Durham University

Durham E-Theses

Twitter Analysis to Predict the Satisfaction of Saudi Telecommunication Companies' Customers

ALMUQREN, LATIFAH

How to cite:

ALMUQREN, LATIFAH (2021) *Twitter Analysis to Predict the Satisfaction of Saudi Telecommunication Companies' Customers*, Durham theses, Durham University. Available at Durham E-Theses Online: http://etheses.dur.ac.uk/13832/

Use policy

The full-text may be used and/or reproduced, and given to third parties in any format or medium, without prior permission or charge, for personal research or study, educational, or not-for-profit purposes provided that:

- a full bibliographic reference is made to the original source
- a link is made to the metadata record in Durham E-Theses
- the full-text is not changed in any way

The full-text must not be sold in any format or medium without the formal permission of the copyright holders.

Please consult the full Durham E-Theses policy for further details.

Academic Support Office, The Palatine Centre, Durham University, Stockton Road, Durham, DH1 3LE
e-mail: e-theses.admin@durham.ac.uk Tel: +44 0191 334 6107
http://etheses.dur.ac.uk

Twitter Analysis to Predict the Satisfaction of Saudi Telecommunication Companies' Customers

Latifah A. Almuqren

000784968

A thesis presented for the degree of

Doctor of Philosophy at Durham University

Supervised by:

Professor Alexandra I. Cristea

Department of Computer Science

Durham University

June 2021

Copyright 2021 by Latifah Almuqren.

"The copyright of this thesis rests with the author. No quotations from it should be published without the author's prior written consent and information derived from it should be acknowledged".

Acknowledgements

First and foremost, all praise and thanks to God for his graces and guidance.

I want to declare my intense gratitude to my PhD supervisors: Prof. Alexandra I. Cristea, for her assistance, in-depth discussions, and excellent advice.

No words can express my love for my parents AbdulRahman and Sarah, thanks for caring, encouragement, for life you gave it to me.

My husband Saud and my kids Saad, AbdulRahman, Mariah, and Sarah are my happiness in life; your presence in my life is the biggest blessing. Thanks for taking care of me in all cases.

My sister's Dr Monerah, Hessah and Al- Joharah thanks a lot for supporting me you are my backbone team.

My brothers Abdul-Allah, Abdul-Aziz, Dr Khalid, and Dr. Abdul-Almohsen thanks for encouragement and assistance.

My nephews, thanks to being in my life.

Thanks to all my friend for help, especially Dr. Nada, and Dr. Nourah.

I would also thank my University, Princess Nourah bint Abdulrahman, to give me the chance to continue my study.

Twitter Analysis to Predict the Satisfaction of Saudi Telecommunication Companies' Customers
Submitted for the degree of Doctor of Philosophy 2021

Abstract

The flexibility in mobile communications allows customers to quickly switch from one service provider to another, making *customer churn* one of the most critical challenges for the data and voice telecommunication service industry. In 2019, the percentage of post-paid telecommunication customers in Saudi Arabia decreased; this represents a great deal of customer dissatisfaction and subsequent corporate fiscal losses.

Many studies correlate customer satisfaction with customer churn. The Telecom companies have depended on historical customer data to measure customer churn. However, historical data does not reveal current customer satisfaction or future likeliness to switch between telecom companies. Current methods of analysing churn rates are inadequate and faced some issues, particularly in the Saudi market.

This research was conducted to realize the relationship between customer satisfaction and customer churn and how to use social media mining to measure customer satisfaction and predict customer churn.

This research conducted a systematic review to address the churn prediction models problems and their relation to Arabic Sentiment Analysis. The findings show that the current churn models lack integrating structural data frameworks with real-time analytics to target customers in real-time. In addition, the findings show that the specific issues in the existing churn prediction models in Saudi Arabia relate to the Arabic language itself, its complexity, and lack of resources.

As a result, I have constructed the *first gold standard corpus of Saudi tweets related to telecom companies*, comprising 20,000 manually annotated tweets. It has been generated as a dialect sentiment lexicon extracted from a larger Twitter dataset collected by me to capture text characteristics in social media. I developed a *new ASA prediction model for telecommunication* that fills the detected gaps in the ASA literature and fits the telecommunication field. The proposed model proved its effectiveness for Arabic sentiment analysis and churn prediction. This is the *first work using Twitter mining to predict potential customer loss (churn) in Saudi telecom companies*, which has not been attempted before. Different fields, such as education, have different features, making applying the proposed model is interesting because it based on text-mining.

Table of Content

List of Tables

List of Figures

List of Abbreviations

CS Customer Satisfaction

CC Customer Churn

SA Sentiment Analysis

ASA Arabic Sentiment Analysis-

Telecom Telecommunication

IAA Inter-Annotator Agreement

JSON Java Script Object Notation

KNN K-Nearest Neighbour

SVM Support Vector Machine

NB Naïve Bayes

ML Machine Learning

MSA Modern Standard Arabic

DA Dialectal Arabic

NLP Natural Language Processing

OCA Opinion Corpus Arabic

OM Opinion Mining

POS Part of Speech

SWN Senti-WordNet

TF-IDF Term Frequency Inverse Document Frequency

WN WordNet

Chapter 1: Introduction

1.1 Introduction, Rationale and Research Problems

Global competition for telecommunication services drives companies to enhance their customers' satisfaction. *Customer satisfaction* defined as customer's response to the expectation level [5].

Customer satisfaction is having been measured using customer interviews and questionnaires, but these methods cannot measure customer satisfaction in real-time [6]. In addition, these methods can be problematic as typically low respond to questionnaires, creating a self-selection bias [7]. Moreover, the time for a turnaround, even for those who respond, creates significant time delays. Many research [8], [9] and [10] has used social media mining to measure customer satisfaction. However, there are currently lacking tools for doing this in Arabic. Measuring customer satisfaction is critical for customer retention [11].

Customer churn is defined within the telecommunication field as customer movement from one telecom company to another [12]. Extensive research indicates that customer satisfaction is positively correlated with customer loyalty and negatively correlated with customer churn [11, 13, 14]. Customer satisfaction and customer churn can be used to optimize industry success: customer churn is reduced when customers are happy. Satisfied customers increase company profits by reducing the costs associated with attracting new customers. It costs five to ten times more to attract a new customer than to retain one, [15], [11], [14], so companies are more concerned with keeping customers than ever before.

Deng et al. [16] assert that keeping customers satisfied is crucial for long term customer relationships. Ranjan et al. [13] has found that positive customer sentiment 'feeling' to be a good predictor for creating new customers. In addition, Li et al. [14] found that a single unsatisfied customer can result in a company loss of 25 additional customers. A churner will influence his social community causing more churning [17].

Avoiding customer dissatisfaction is critical to customer churn. Customer churn prediction requires a thorough customer behaviour analysis [18]. Churn management provides a vital tool for Customer Relationship Management (CRM) in the telecommunication industry [19], [20]. Churn management, keeping the existing customer, is vital [21], [22].

Telecom companies have depended on historical customer data to measure customer churn. However, historical data does not reveal current customer satisfaction or future likeliness to switch between telecom companies.

The literature review presented in Chapter 2 reveals that many studies have been done focusing on developing churner prediction models- based on historical data. These models face delay issues and lack timelines for targeting customers in real-time [23]. Also, these models lack the ability to tap into Arabic language social media for real-time analysis. As a result, the design of a customer churn model based on real-time analytics is needed.

Using real-time methods could help solve problems of delayed data collection - and allow for customer feedback analysis and the creation of effective retention plans. Conversely, delays may cause a drop-in market position, especially for large consumer populations across multiple time zones where daily monitoring data is complex. However, again, the lack of an Arabic language tool limits the usability of this data.

This research addresses the following problems related to customer churn prediction models:

- Time sensitivity: The current churn prediction models have a relatively short life because as they rely on historical customer data, where the data becomes less valuable over time. [20]
- Language-specific issues: The current churn prediction models exclude location and language factors, and that causes a miss important information [24].
- Real-time analytics: A lack of research integrates structural data frameworks with real-time analytics for targeting customers in real-time [23].

Significantly, the reviewed literature indicated that social mining is a powerful tool for predicting customer churn. However, a knowledge gap exists as to how social mining predict customer churn in various industries: how this technique can be used to assess customer behaviour in other industries such as education or marketing.

Social media is a key part of many people's lives today; 85.1% of active internet users are social network platform users [25]. Social media is a communication tool that allows people to share their sentiments, thoughts, opinions and moods [26]. Social media mining can offer rich and diverse data that might be used

to measure customer satisfaction without having to perform surveys [13]. Mining social media data can be considered a real-time technique and involves less time and effort to recap a conclusion [13].

Sentiment analysis or 'opinion mining' refers to the computational processing of opinions, feelings and attitudes towards a particular event or issue [27], [28]. Sentiment analysis of social media platforms can be used by company management to take timely actions to improve customers' experience, avoid customer churn, and create positive customer attitudes about the brands' customers prefer [13]. Sentiment analysis can help organisations support decision-makers in predicting stock markets by identifying the feelings of social network users about financial matters [29].

With seven telecom companies in Saudi Arabia, the three largest are STC, Mobily and Zain [30]. Together, the seven have 41.63 million subscribers who use mobile voice communication services. The prevalence of mobile voice service among the population is 124,6% [31], indicating that many individuals subscribe to more than one mobile service. Over 31% of customers registered complaints in 2018 alone. Saudi Information Technology Commission [31] representing over 13,103M complaints. The percentage of mobile internet subscribers in 2019 in Saudi Arabia is 80% of the total population [32]. In Saudi Arabia, 11 M used Twitter in 2018 [33].

The percentage of post-paid telecommunication customers in Saudi Arabia decreased in 2019 [2]- this represents a great deal of customer dissatisfaction and subsequent corporate fiscal losses. Obviously, these companies need to add more customers while retaining their existing customers. To do this, they urgently need a new method to assess customer satisfaction and predict customer churn. This helps in developing effective customer retention programmes for the company.

Current methods of analysing churn rates are inadequate, particularly so in the Saudi market. The specific issues with current churn models in Saudi Arabia relate to the Arabic language itself. Arabic is a rich morphological language [34], [35], written from right to left and using different forms. There are different forms of the Arabic language: *Classical Arabic (CA)*, as in the book of Islam's Holy Quran, *Modern Standard Arabic (MSA)* used in newspapers, education and formal speaking, *Dialectical Arabic (DA)* used in informal everyday spoken language and found in chat rooms and social media platforms. Arabic is actually a group of dialects. Mubarak and Darwish [36] identified six distinct Arabic dialects: Gulf, Yemeni, Iraqi,

Egyptian, Levantine and Maghrebi. Every dialect has its own grammar and vocab [37], complicating any attempt to build an Arabic lexicon [38], [34], [39].

It has been shown in the literature that the sentiment analysis of the Arabic language is quite challenging [40] for the reasons mentioned earlier. Although ASA is of growing importance, it is still in the early stages of research [41], [42], [43].

Compared Arabic to other languages, Arabic lacks a large corpus [44], [45], [39], [46], [43]. Fewer Saudi dialect corpus and lexicon resources exist than other Arabic dialects, such as the Egyptian dialect, which had a lot of attention; one of the earliest Egyptian corpus is CALLHOME corpus [47]. In addition, Levantine Arabic received much attention, such as the Levantine Arabic Treebank (LATB) [48].

Regarding the Arabic dialects, this is not the case with the Gulf dialect, especially the Saudi dialect. Unfortunately, there are shortcomings to the existing corpora and their availability. This is due partly to the strict procedures for gaining permission to reuse aggregated data, with most existing corpora not offering free access. Saudi lexicon is needed in order to analyse real-time customer positive and negative attitudes toward telecom services.

For that, this research efforts to fill this gap by creating golden standard Saudi corpus AraCust and Saudi lexicon AraSTw for use in data mining - specific to the telecom industry. While some Arabic lexicon resources currently exist, there is no specific Arabic telecom lexicon. Using the lexicon in the data mining tool is the key to tool accuracy. No Saudi dialect lexicon has ever been developed for the telecom industry, even though 467 M Arabic speakers [49]. Future work can include expanding the lexicon to include more regional dialects.

1.2 Initial Hypotheses

The research hypotheses originated from the conflicting findings of the reviewed studies and the gaps identified in the literature in terms of the relationships between customer satisfaction, social media sentiment and customer churn prediction. Each hypothesis addresses one research question (RQ), as listed below.

1. If measurable criteria for customer satisfaction are defined, they could extract services that do not meet the expectations of customers (RQ1 and RQ3).

2. Customer satisfaction with telecom companies in Saudi Arabia can be monitored by analysing microblogging sites (RQ2).

3. The customer churn of telecom companies in Saudi Arabia can be predicted by analysing microblogging sites (RQ4).

1.3 Research Objectives, Questions and Techniques

As a result, I have defined the following umbrella research question:

RQ0: Can Twitter be used to automatically monitor and compute customers' satisfaction with telecom companies in Saudi Arabia and predict customer churn?

This research question is further divided into sub-questions in Table 1.1 below, leading to specific Research Objectives. This research tries to answer RQ0 by answering the sub-questions.

In the same table, I further mention the main research techniques used to reply to that particular research question and target a specific objective.

Research Questions	Research Objectives	Research Techniques	Chapters
What are the traceable, measurable metrics for customers' satisfaction with telecom companies in Saudi Arabia, and how can they be combined for visualisation?	1. To create a framework of measurable, weighted metrics for customers' satisfaction with Saudi telecom companies.	Review the literature, a communications and information technology commission report, and a questionnaire to define the traceable measurable metrics taxonomy.	Chapter four
What types of services for customers of telecom companies in Saudi Arabia are mentioned in tweets, and what is the customer sentiment about these services? **Sub-question:** Are there services not discussed in tweets that could be relevant to the sentiments of customers?	2. To propose recommendations to improve the services of Saudi telecom companies.	Results from step 1 and an annotation process to define the services. Using multi-way sentiment analysis toward the services and visualize the satisfaction and services of the service importance to set the recommendations.	Chapter five and Chapter six
Can we automatically measure and make automatic predictions about customers' satisfaction with telecom companies in Saudi Arabia using Twitter?	3. To identify based on Twitter mining Saudi telecom companies' customers' satisfaction.	Data was taken from Twitter, was processed using a Python Script, manually annotated. Using the developed lexicon to estimate which tweets expressed positive or negative values about which telecom company.	Chapter five
Is it possible to predict the customer churn of telecom companies in Saudi Arabia by analysing customers' tweets?	4. To predict the potential ratio of customer churn.	Feed the churn prediction model with the historical customer data, and the results from step 3.	Chapter seven

Table 1.1: Research questions, objectives and techniques.

1.4 Research Methodology

To accomplish the aim of this study, a deductive approach will be used. Deductive research involves adopting hypotheses and testing them in a causality manner [50] to explore the relationships among the research variables. A review of the literature determined that there are distinct gaps in social media analysis to measure customer satisfaction and predict customer churn. Accordingly, this study will address the relationship among the variables of a Twitter sentiment analysis, customer satisfaction and customer churn. In addition, this study plans to use multiple approaches, such as a supervised approach for an Arabic SA, deep learning, and transformer networks, to develop a model for capturing customer satisfaction and predict customer churn. There is a strong correlation between SA and marketing because customers tend to express their feelings towards products on social media. This provides an opportunity to analyse these feelings or sentiments to measure customer satisfaction.

1.5 Study Aims, Originality and Outcomes

In this study, I used real-time Twitter mining methods in tandem with historical data to predict customer churn. This is a new method not previously used. Subsequent to my publication [51], another researcher [13] mirrored my methods to predict customer base growth. The present study proposes a novel model for the telecom industry that fits the telecom data. The model also takes into consideration language and location factors. In addition, the present study intends to introduce a notion of customer interaction for Saudi telecommunication companies based on the prediction model of the ‘lost customer’ phenomenon (or customer churn).

In addition, this study contributes to the ASA research community by developing three Arabic resources: corpus, lexicon, and Arabic sentiment tool. The corpus and lexicon are in the Saudi dialect. Other sectors will use my Arabic tool to measure customer satisfaction and customer churn, including education and business. Both of these sectors rely on data mining and depend on Arabic text analysis. In the digitized world, especially in the current situation, during Covid-19, using social media to express feelings among the customers or students has dramatically increased [52]- Likewise, increasing the need for a useful SA tool.
This study is unique because:

- It is the first work using Twitter mining to predict potential customer loss (churn) in Saudi telecom companies, which has not been attempted before (Chapter 7), and
- It develops the ASA model by combining both transfer language model and deep learning model (Chapter 5), and
- It creates Saudi resources to solve the lack of Saudi resources issue (Chapter 3).

The outcomes of this study will be:

1. Assessing ASA's current situation, the main approaches contributing to ASA and the challenges that faced ASA using a systematic review.

2. AraCust Corpus: Constructing a first golden standard corpus of Saudi tweets related to telecom companies from this dataset comprising 20,000 manually annotated.

3. AraSTw Lexicon: Generating a Saudi dialect sentiment lexicon extracted from Twitter data sets to capture the texts' characteristics in social media. It comprises 34,755 words. It outperformed another state-of-the-art Arabic lexicon.

4. Contributing to the Arabic Sentiment Analysis (ASA) research community by developing a new model combining deep learning and transfer language models.

5. Identifying and evaluating the main gaps in the current churn prediction models.

6. Propose and evaluate a novel design of a churn prediction model to address the gaps in current churn prediction models by providing a real-time method that suits the telecom data.

7. Providing recommendations for telecom companies based on monitoring real-time customers' satisfaction through Twitter.

1.6 List of Publications

Work conducted in this thesis has contributed in the following ways.

Conference Papers

- Almuqren, L. and Cristea, A.I., 2016, July. Framework for Sentiment Analysis of Arabic Text. In *HT* (pp. 315-317). Core A.
- Almuqren, L. and Cristea, A.I., 2016, July. Twitter Analysis to Predict the Satisfaction of Telecom Company Customers. In *HT (Extended Proceedings)*. Core A.
- Almuqren, L., Alzammam, A., Alotaibi, S., Cristea, A. and Alhumoud, S., 2017, July. A review on corpus annotation for Arabic sentiment analysis. In *International Conference on Social Computing and Social Media* (pp. 215-225). Springer, Cham.
- Almuqren, L. A. R., Qasem, M. M. & Cristea, A. I. (2019). Using Deep Learning Networks to Predict Telecom Company Customer Satisfaction Based on Arabic Tweets. In Information Systems Development: Information Systems Beyond 2020 (ISD2019 Proceedings). Toulon, France: ISEN Yncréa Méditerranée. Core A
- Almuqren, L. & Cristea, A. I. (2021). COVID-19's Impact on the Telecommunications Companies. In WorldCist'21 - 9th World Conference on Information Systems and Technologies. Core C.

Journal Papers

- Almuqren, L., & Cristea, A. (2021). AraCust: a Saudi Telecom Tweets corpus for sentiment analysis. PeerJ Computer Science, 7, e510.

- Almuqren, L., Alrayes, F. S., & Cristea, A. I. (2021). An Empirical Study on Customer Churn Behaviours Prediction Using Arabic Twitter Mining Approach. Future Internet, 13(7), 175.

Papers under review

- Predicting STC Customers' Satisfaction using Twitter Mining for IEEE Transactions on Computational Social Systems.
- Multi-Way Arabic Sentiment Analysis for Information & Management Process Journal.
- AraBERT-GRU Model for Arabic Sentiment Analysis for UMUAI Journal.
- Arabic Text Sentiment Analysis: Reinforcing Human-Performed Surveys with Wider Topic Analysis for IEEE Access Journal.

Participatory

- I have participated in a conference with a poster of this work in Princess Nourah University, King Saud, Saudi Arabia.
- I attended a lot of conferences and seminars about Sentiment Analysis in Saudi Arabia.
- I have participate in the Women's ACM with a poster of this work.

1.7 Research Framework

Figure1.1 shows the research steps that will be followed to achieve the aim of this study.

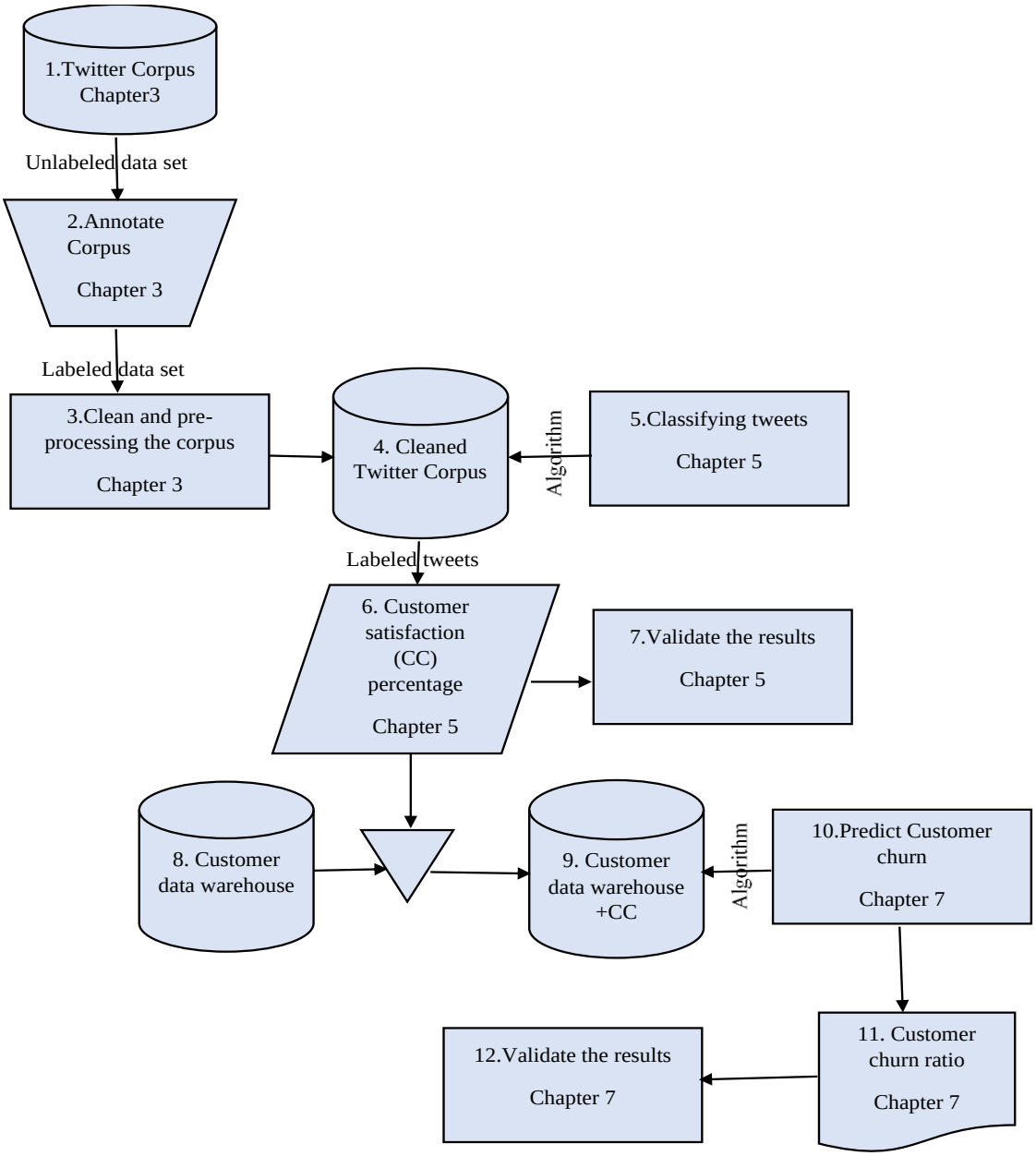

Figure 1.1: Research framework maps.

1.8 Mapping of Research Questions, Objectives and Methodology Across the Thesis Chapters

Figure 1.2 shows a map of the work to be carried out across the chapters of this thesis.

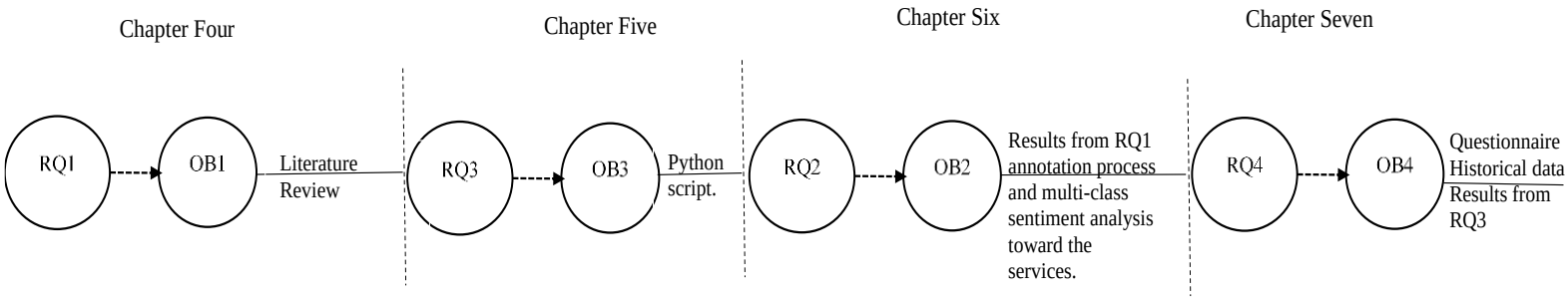

Figure 1.2: Mapping of research questions, objectives, and methodology across the thesis chapters.

The rest of the thesis chapters, Chapter two, a systematic review to assess the current churn models and the current stage of Arabic sentiment analysis and highlight the challenges that faced ASA. While Chapter three explained the construction of the golden annotation corpus and sentiment lexicon.

Chapter 2: Literature Review

2.1 Introduction

The literature review is an essential phase as it enables the researcher to understand the topic and include key concepts. This Chapter investigates the literature and background related to the main areas of this research, which are customer satisfaction, customer churn and Twitter mining on the road, in order to build up knowledge of these key concepts. In addition, this Chapter systematically reviews all the Arabic sentiment analysis models and resources in order to highlight the issues facing the Arabic sentiment analysis approach. This will allow the researcher to build a model commensurate with the nature of the Arabic language and its structure and characteristics. The systematic approach was evaluated using an epistemological engine. Furthermore, this Chapter critically reviews the state-of-the-art churn prediction models to identify gaps in the existing models.

Consequently, this Chapter provides a vital foundation and theoretical basis for the research. The output of this chapter answers RQ1. The scope of this literature review is examining Arabic sentiment analysis models, churn prediction models and their issues.

2.2 Literature Review

2.2.1 Customer Satisfaction and Customer Churn Definitions

Enhancing customer satisfaction (CS) is a popular topic in the marketing literature. Extensive research correlates customer satisfaction with customer churn [11, 13, 14]. Customer satisfaction and customer churn have been identified as two factors that contribute to success in several industries, such as telecommunication [53], medicine [54] and tourism [55].

Customer satisfaction is defined as the outcome of using a service, resulting from the comparisons that the buyer makes with other similar providers in terms of the rewards and costs of the service [56, 57]. Similarly, [58], [59] and [60] define customer satisfaction as an evaluation of the expectations and the actual execution

of the service. Correspondingly, [5] consider customer satisfaction as the customer's response to their satisfaction level.

Meanwhile, customer churn with respect to the telecommunication industry is defined as the turning away of customers from one telecom provider to another [61-63], [18, 21], [64] – e.g., the number of customers switching to another mobile provider.

Both definitions can be linked to one another in the search for an optimum strategy to improve industry success. In other words, customer churn can be stopped by making customers happy. Customers who are satisfied with a company's services make a company more profitable because the cost of attracting new customers is five to ten times greater than the cost of retaining existing customers [15], [11, 14]. Therefore, companies are more concerned with keeping customers than ever before. Ali et al. [11] stated that customer satisfaction is essential for customer retention. Similarly, [16] asserted that keeping a customer satisfied is crucial for a long customer-supplier relationship. Ranjan et al. [13] confirmed that positive customer sentiment can be a good indicator of the possibility of gaining new customers. In addition, [14] claimed that unsatisfied customers can lose the company 25 customers. Hence the importance of customer satisfaction is that it plays a role in avoiding customer churn.

A study states that customer churn prediction requires customer behaviour analysis [18]. Churn management plays a vital role in Customer Relationship Management (CRM) in the telecommunication industry [19, 20]. Churn management has been identified as the processes of keeping existing customers [21], [22].

There are two types of customer churning: voluntary and involuntary. The decision by a customer to move to another telecom company of his own will is called voluntary, while forcing a customer to stop using a telecom company's services for any reason such as death or changing jobs is called involuntary [65]. Usually, scholars are interested in voluntary customer churning because it describes the relationship between a customer and a company. There are two types of customer scheme, post-paid or pre-paid. Post-paid customers receive a monthly bill for the company's services while pre-paid customers are charged in advance for the company's services.

I define customer churn in our study as a post-paid customer who voluntary leaves the company and stops using their services within our time window. In contrast, a non-churner in our study is a post-paid customer who remains with the company within our time window.

2.2.2 Social Media Mining

Big data is the term used to refer to extensive data sets. In this study, big data refers to the enormous data generated from social media platforms and users' activities on these platforms. Social media data can be tweets, reviews, posts, etc. [66]. Social media encompasses various platforms that allow people to share and exchange information, producing an abundance of valuable data. Some social media platforms are Twitter, Facebook, Yelp, Linked In, YouTube. The social media data on these platforms graduate from structured data to unstructured data such as news or microblogging. Mining these data would provide a wealth of knowledge and useful information on many levels. Using different data mining techniques to analyse social media data is nevertheless a dynamic domain of research. There are various social media mining techniques, such as sentiment analysis, opinion mining, or social media analysis. Social media mining can offer rich and diverse data that could be used to measure customer satisfaction without having to perform a survey [67]. Mining social media data is a real-time technique, and a conclusion can be drawn with less time and effort than with some other methods [13].

Several studies have highlighted the benefits of social media analysis, especially SA, for organisations [68, 69], [70], [71], [72], [73], [74], [75], [13, 76]. According to [68], social media analysis can help organisational decision-makers predict the stock market by identifying financial and social network users' feelings. Other researchers have stated that mining social media data is important for marketers and customers for several reasons. For example, it produces an abundance of useful data, which provides a wealth of information about customers for the company [77]; it helps to develop a recommendation system to maintain existing customers or gain new ones; it helps organisations to change customers' decisions about their services as they can access customers' opinions; it is also effective for building confidence among customers and stakeholders [67]; it helps to prevent customer churn [78].

Regarding the reasons why social media is preferable for users, [13] explained that users might be more comfortable expressing their opinions on social media platforms than traditional methods. In addition, users tend to express their true feelings towards a brand and its services through social media platforms rather than filling in a questionnaire [13].

2.2.2.1 Why Twitter?

Twitter is a popular, widely used messaging service categorised as a microblogging website [26]. It was launched in July 2006. Messages on Twitter are embedded into so-called 'tweets', which are individual, unstructured text messages with a limit of 140 characters, recently raised to 280 characters.

Twitter has been selected for this research because millions of users post their opinions, sentiments and moods daily. It is an easy platform to use on many devices. Twitter offers available access to tweets for developers, using Application Programming Interface (API) and Open Authentication (OAuth) for API [79]. The developer can retrieve the last seven days' tweets, while prior than seven days seeing as a historical tweet. The collected data include the tweet text, the tweet author id, the tweet date, the tweet location, etc.

Across the world, Saudi Arabia ranks seventh in terms of the number of personal accounts on social media [49]. In 2020, Twitter users in Saudi Arabia reached 12 million [80]. Unsurprisingly, Twitter is one of the most visited sites in Saudi Arabia, and the number of Twitter users continues to rise rapidly. Al-Jenaibi [81] explained why Saudi people prefer Twitter over other social media websites, including the fact that Twitter allows Saudi people to freely express their views about certain prohibited subjects, thus changing the nature of this once closed, conservative community.

2.2.3 Providing Customer Satisfaction with social media

This review aims to define how to measure customer satisfaction using social media mining, specifically Twitter mining. Firstly, I reviewed several case studies on the applications and methods used in relation to customer satisfaction. Four classification methods were identified. These methods, listed in Table 2.1, which also provides a general description of each method, include advantage, disadvantage, the purpose of application and example of application in case studies.

No.	Tools	Advantages	Disadvantages	Applications	Example of Case Study
1	Face-to-face survey	Captures verbal and non-verbal queries. Ability to hold and keep respondents' focus longer. Effective for responding to open-ended questions. Queries can be answered. **Supports quantitative survey design.**	Expensive (e.g., travel). Time consuming.	Complicated or lengthy subjects. Focused on key customers. Customers are geographically grouped.	[60]
2	**In-app or postal survey**	Can be developed and administered by the researcher. Low cost. Suitable for high numbers of respondents. **Flexibility for the respondent to complete the survey at a suitable time.**	Poor response rates. Poor response to open-ended questions. Tendency to misunderstand (the questions). **Attracts either unsatisfied or very satisfied customers.**	Strong relationship with a company (e.g., surveying employee attitude), subject (e.g., homebuyer survey). Focused on respondents that are obligated to complete it.	[82] [83]
3	**Telephone survey**	Low cost. Ability to control the interviewer standards and the number of samples.	High cost (if outsourced). Time consuming. Difficult to reach respondents who lack phone access.	Widely used in business-to-business (b2b) customer surveys.	[84]

		Simple to collect rating answers using scales.	**Tendency to misunderstand (no visual representation).**		
4	**Data Mining**	Suitable for large numbers of respondents.	High cost (using a sophisticated tool).	Widely used for predicting the level of customer satisfaction.	[85] [86]

Table 2.1: Findings Related to the Methods Used to Review Customer Satisfaction.

The review results show that the most popular means of gathering and measuring customer satisfaction is through surveys [67]. Most studies examined data mining techniques to measure customer satisfaction and predict customer churn in the telecommunication industry. However, few studies have measured customer satisfaction, particularly in the telecommunication industry, using social media mining. Table 2.2 shows the various studies reviewed in relation to customer satisfaction and social media mining and the findings reported in those studies.

Variable	Reference(s)	Summary of Findings
Sentiment	[69]	- Measured public transport rider satisfaction towards transit system services using the riders' tweets. - This helped to improve their service quality and safety monitoring. - The findings showed that sentiment analysis (SA) can successfully detect rider sentiments in real time towards a transport organisation.
	[71]	- The researchers used SA to propose a tool for evaluating customer satisfaction in a job search company. - The results showed that over 42% of the company's clients had positive impressions

		about the provided services and 34% did not express any feelings in their remarks.
	[73]	- Sought to predict the 2014 European elections based on Twitter use and opinion polls using a lexicon-based classifier for SA. - The researchers achieved better results than several baselines, including polls, prediction websites and replication of old works.
	[72]	- Mining random tweets on Twitter to determine consumer's sentiments towards certain brands through SA. - The findings proved that there is positive consumer sentiment towards famous brands.
	[87]	- Measured customer satisfaction for two online transportation service providers in Indonesia.SA using support vector machines (SVM), NB and Decision tree (DT). - Their data set includes 9,191 tweets. - They found that the customers preferred to express bad sentiments on the companies' Twitter accounts, instead of positive; SVM and DT had the highest performance. - They did not use features for pre-processing and classifying the data, which could have given better classifier results.
	[88]	- This paper used SA and customer satisfaction to measure brand reputation. - They used Naïve Bayes, SVM and DT classifier methods. - Their data set included 10,000 tweets. - Their results proved that the best algorithm is the Support Vector Machine algorithm.

Mood	[75]	- Examined the influence of the public mood on the closing value of the Dow Jones Industrial Average (DJIA). - Analysed the texts of daily tweets using mood tracking tools: Opinion Finder and Google-Profile of Mood States (GPOMS). - The outcomes indicated that not all changes in the public mood matched the DJIA value shifts but that some of the public moods, categorised as 'calm', can predict the DJIA values.
Opinion, Attitude and Sentiment	[70]	- Analysed consumers' behaviour with regard to food products, using their micro blogging messages (i.e., tweets) to monitor and analyse consumer opinion, attitude and sentiments expressed in shared posts and comments. - The results showed that the success of branding required sentiments to be monitored for a long period of time, because these sentiments do not change quickly.

Table 2.2: Summary of the Literature that Links Customer Satisfaction, Social Media Mining and Twitter Features.

From the analysis of studies that analysed customer satisfaction through Twitter mining, it appears that a few studies used 'mood' as a Twitter feature when measuring CS, e.g. [75], who examined the influence of public mood on the closing value of the Dow Jones Industrial Average (DJIA). In addition, a few studies used 'opinion' or 'attitude' as a Twitter feature to measure CS. For example, [70] analysed consumers' behaviour regarding food products by using their microblogging messages (i.e., tweets) to monitor and analyse consumers' opinions, attitudes and sentiments expressed in shared posts and comments. However, most studies used the sentiment as a Twitter feature to measure CS, such as [69] and [72].

Figure 2.1 shows our customer satisfaction model, which contains all the variables that resulted from the literature analysis linking Twitter features, customer satisfaction and customer churn.

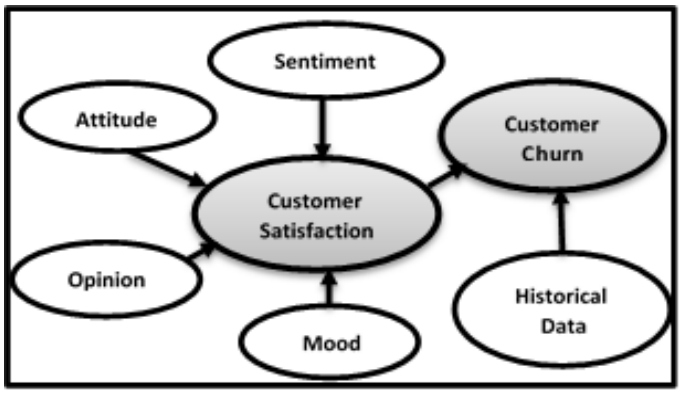

Figure 2.1: Customer satisfaction model.

Subsequently, studies that linked customer satisfaction with social media mining in the telecommunication industry were reviewed to find the gaps in research that this study could fill (see Table 2-3). Studies that reviewed techniques other than data or social mining (e.g., surveys) were excluded because this research aims to understand social mining techniques. Hence, reviewing other techniques would not be beneficial for this aim. In addition, studies that reviewed social mining applications for other purposes, i.e., other than for assessing CS, were excluded. As this study focuses on applying of social mining for assessing CS, selecting other application fields would not contribute towards its aim.

Reference	Aim	Technology	Data Set	Findings	Gap Identified
[8]	Measured customer satisfaction towards telecommunication companies in Saudi Arabia using different algorithms.	Sentiment Analysis (SA) using KNN, NB and ANN.	1,331 tweets	KNN was superior to the other algorithms with 75.6% for F-measure.	The data set included only English tweets, although the majority of customer tweets about Saudi telecommunication companies are in Arabic. This limited the capture of customers' real sentiments.
[9]	Analysed Jordanian telecommunication companies' customer comments on Facebook.	SA using KNN, SVM, NB, and DT	14,332 customer posts on Facebook	SVM classifier outperformed the other three classifiers with 95% accuracy.	They classified the comments into positive, negative, other, or question, considered the negative and positive comments and discarded the 'other' and 'question' classifications.

[89]	Combined the SA ViseKriterijumsa Optimizacija I Kompromisno Resenje (VIKOR) [90], which is a ranking and optimization approach to developing a framework for predicting customer satisfaction with mobile services	Text mining using a developed dictionary. The result was evaluated with VIKOR.	Customer reviews on the web.		First, not validating the results by comparing them with actual results for customer satisfaction. Second, using a basic method for SA.
[10]	Proposed an approach for measuring customer satisfaction with mobile companies	Naïve Bayes (NB) Classifier	8,000 Indonesia tweets from Twitter	The approach obtained a F1-score of 93.5% and accuracy of 99.09%.	This research depends on one technique. It neglected to use other more advanced techniques.

Table 2.3: Gap analysis.

Table 2.3 illustrates that although several studies are using social media to measure customer satisfaction, few studies measure customer satisfaction through social media mining in the telecommunication industry. Most studies measured customer satisfaction in the telecommunication field using English tweets. Consequently, there is a need to mine Arabic tweets in the telecommunication industry.

Therefore, our study plans to use Sentiment Analysis for Arabic tweets to measure CS and use the CS results and historical customer data to predict CC. It will use advanced techniques, including deep learning and transformer network models. In addition, it will validate the customer satisfaction results from Twitter mining with actual results using statistical methods.

2.2.4 Customer Churn Prediction and Social media mining

This section aims to critically evaluate the literature and discuss the themes and gaps discovered in the churn prediction models that would allow the arguments for this PhD research to be appropriately framed. The

review included a full-text assessment of the studies, and Table 2.4 presents the data sets and algorithms used and the results of the reviewed studies.

Due to the lack of studies using social media mining to develop a churning prediction model, this section will discuss the studies that adopted different state-of-the-art techniques for developing churn prediction models based on historical data or other parameters. By analysing the reviewed studies, I was able to identify the different techniques used to develop the existing churn prediction models and their frequency –Table 2.5. The following paragraphs analyse some of the research listed in Table 2.4 in more detail.

Olle and Cai [91] investigated hybrid models that build on data mining techniques to explain churn behaviours. Their data set was a telecom data set from an Asian mobile operator. They applied a logistic regression used in parallel with the voted perceptron for classification purposes and then combined this with clustering for churn prediction packaging in Waikato Environment for Knowledge Analysis (WEKA). The results showed that the new hybrid model was more accurate than single methods with a 0.721 ROC value.

In addition, [92] proposed a hybrid approach for customer churn prediction based on neural networks. Their data set was from an American telecommunication company. They combined a self-organised map (SOM) and an artificial neural network (ANN). Their proposed approach started with data reduction for the unrepresentative training set using ANN. Then, the output was fed into SOM to develop the churn prediction model. The results showed that the hybrid approach enhanced the prediction accuracy more than using a single neural network. This work addressed a limitation, which is that data reduction causes loss in the training set.

Many researchers have compared different classification techniques for the customer churn problem. Ali et al. [11] predicted customer churn behaviour by using various data mining techniques. Their data set was an online customer data set available at Kaggle (https://www.kaggle.com/). They used different classifiers implemented in WEKA, i.e. support vector machine (SVM), bagging, stacking, C50/J48, PART, naïve Bayes, Baysen Net and Adaboost. They summed up their findings to conclude that bagging and the Sequential minimal optimization (SMO) algorithm outperform the others with an accuracy of 99.8% using 14 attributes. Hassouna et al. [20] empirically compared two techniques for customer churn: decision tree (DT) and logistic regression models with 15,519 and 19,919 customers, respectively, from a UK mobile telecommunication

provider. They stressed the need for more advanced methods of churn modelling. Additionally, [93] proposed a model for churn prediction for telecommunication companies using logistic regression and DT in R. Their data set was the historical records extracted from the telecom industry. They concluded that data mining techniques could be a promising solution for customer churn management.

Some researchers, such as [94], argue that a DT is the best classification model. Based on their result from a churning prediction experiment, the DT model surpasses the neural network model in predicting churn using a PAKDD – 2006 data mining competition data set [95]. The same result was obtained by [96]. They applied DT, neural network and regression techniques to develop a churn prediction model. They stated that the DT outperformed some of the existing data mining techniques.

Another study by [97] argued that random forest achieved better results than other prediction classifiers based on a comparison between different classification models. However, to overcome the random forest model issue, they planned to use deep learning models. In connection with this, [98] recommended applying random forest to develop a customer churn prediction model. However, some researchers have argued that random forest is not appropriate for customer churn prediction [18] because it is complicated in to understand [55].

On the other hand, [99] claim that an SVM is the best model for churning prediction. They used data sets gathered from the machine learning UCI database, University of California. The data were related to home telecommunication. They concluded that SVM has a better performance in terms of accuracy than other classifiers such as Logistic Regression and Naïve Bayes. The same result was obtained by [100]. They used different data mining techniques to predict customer churn: SVM, Neural Network, DT and NB. The data set included employee and customer information for a year and a half. They concluded that the SVM is the best classifier to develop an accurate churn prediction model.

Some studies attribute the superior performance of SVM to its ability to handle the random curve [101]. In addition, [102] stated that the disadvantages of using SVM might be overcome by predicting the boundaries between true positive and true negative. However, [103] confirmed that SVM has the main disadvantage: it produces a black box, which causes an illusion.

To overcome the disadvantages of SVM, some researchers have proposed using a neural network model to predict customer churning. Clark et al. [104] used an ANN in SPSS on 2,427 customer records from the UCI Repository, University of California. Their proposed approach predicts customer churn with 92% accuracy. Similarly, [18] analysed the meaning of churn management in the mobile telecommunication industry, and they designed a new prediction model to predict churn. They used two data sets: one was original data, and the other was statistical data. They found that the neural network was superior as a scoring model by the overall instance of lift.

Moreover, [105] proposed a novel model for churn prediction in the field of insurance using deep and shallow models such as long short-term memory (LSTM), convolutional neural network (CNN), random forest and AdaBoost. They concluded that the combination of deep and shallow models enhanced performance more than a deep model and shallow model independently.

To enhance the classification performance, researchers have proposed combining one technique with another technique. Ahmad et al. [106] predicted which customers were most likely to churn using data provided by the Syriatel telecom company. They used DT, random forest, gradient boosted tree, and extreme gradient boosting (XGBOOST). The best results were obtained by applying the XGBOOST algorithm with 93.3% area under the curve (AUC) value.

In addition, [107] compared different classification methods for the customer churn problem with the boosting versions. The churn data set utilised was from the UCI Machine Learning Repository, California University[1]. They applied DT, ANN, SVM, boosting, NB and regression analysis. SVM-POLY with AdaBoost achieved the best performance with 97% accuracy.

Furthermore, [108] used rotation forest and Rotboost. The Rotboost method is utilised in the rotation forest and AdaBoost combining. The authors concluded that the Rotboost outperformed rotation forests in terms of accuracy. Comparably, [109] developed a churn prediction model using 50,000 pieces of customer information from telecom data publicly available for Orange large & Cell2Cell. They compared different ensembles such as RotBoost (RB), Random Forest, Rotation Forest and Decorate (DEC) with minimum

[1] https://archive.ics.uci.edu/ml/index.php

redundancy and maximum relevance (mRMR) packaged in WEKA and MATLAB. They found that the approach using the mRMR method combined with RotBoost achieved the best implementation in terms of accuracy with 0.761% AUC.

In addition, [110] assessed data preparation (transformation of the categorical and continuous data) in the prediction of model performance. Their data set was obtained from 30,104 customers from a large European telecommunication company. They applied logistic regression and found that data preparation enhanced the prediction model's performance by 14.5%, as measured by AUC. Some of the subjects related to developing a churn prediction model, uncertainty sample.

Amin et al. [111] considered the uncertainty of the samples in the churn prediction model performance. They found a positive relationship between the size of the sample and the lower distance test set (LDT) sample performance. The LDT has a better performance than the upper distance test set (UDT) samples when the uncertainty sample is raised. Another aspect examined in the literature is the impact of social network analysis on churning model performance.

Dasgupta et al. [112] used 60 GB of data from the largest telecom company, including voice call details, SMS details, etc. They applied the J48 Decision Tree and other classifiers packaged in WEKA. They concluded that social network analysis increases the accuracy of customer churn predictions. Similarly, [24] proposed using customers' social network information and their call log details to predict user churn using the Pokec [2]social network data. They generated synthetic call log details of the social network users (25,000 users of the Pokec data set). They used influence maximisation, calculated by considering a user's topic of interest from the users' social network data and call duration (CLD), together with message length from the user call log data. Future analysis should factor in both location and language to avoid geographical and cultural sampling errors.

In the literature, some researchers study growth, and although churn is the opposite of growth, they are quite related. Ranjan et al. [13] developed a prediction model for the growth rate of new subscribers for telecom subscribers by using sentiment score. A total of 153,651 distinct tweets for the Twitter handles of five popular

[2] http://snap.stanford.edu/data/soc-pokec.html

telecom brands in India were analysed with semantic analysis. The authors proved that SA could manage the high growth rate of new subscribers added to the brand in the study period.

There is only one study that followed our study [113] found a relationship between the sentiment of Twitter feeds related to Telcom's broadband internet service and the customer churn rate. They applied LSTM for SA. Their results showed that churn prediction could be improved by monitoring the negative sentiment by around 1.47% Mean Average Percentage Error (MAPE). However, their study did not use social media mining to develop the churn prediction model. Related research mainly uses company-provided data for churn prediction. Whilst this is a useful source, this is not always available.

After critically reviewing previous studies, I have concluded that this current study is a pioneering work in using Twitter mining to develop the churn prediction model [51].

A significant finding from the literature review that investigated churn prediction models is that social mining is a powerful tool for predicting customer churn. However, the fact that only one study was found in this review that used social media mining shows a knowledge gap in how social mining can predict customer churn in various industries. The following problems related to existing customer churn prediction models were found in the literature:

- The current churn prediction models have a relatively short life as they rely on customers' historical data. The data become less valuable over time for making [20], which may not provide telecom companies with the best churn prediction experience.

- There is a lack of research that integrates a structural data framework with real-time analytics to target customers in real time [23].

- The current churn prediction models exclude location and language factors and that causes geographical and cultural sampling errors [24].

Therefore, this study will use real-time Twitter mining methods and a data warehouse to develop a churn prediction model to prevent customers from turning to other companies, thus enhancing competitiveness. The model will also take into consideration language, time and location factors. The present study intends to

introduce a notion of customer interaction for Saudi telecommunication companies based on the prediction model of the ‘lost customer’ phenomenon (or customer churn).

Ref	Data set	Algorithms	Results
[106]	Spark environment (https://spark.apache.org/) by working on a large data set created by transforming big raw data provided by the Syriatel telecom company.	DT, random forest, gradient boosted machine tree and extreme gradient boosting (XGBOOST).	The best results were obtained by applying the XGBOOST algorithm with 93.3% AUC value.
[23]	Structured data and unstructured data. The unstructured data included: 1) Call centre interaction: details of customer complaints and feedback. 2) Data records captured, such as data regarding purchase, download of apps, etc.	RFM technique, which identifies the customers who will churn by examining how recent customers have made purchases (Recency), how often they made purchases (Frequency) and how much they spent on their purchases (Monetary).	They recommended the integration of the structural data framework with real-time analytics to target customers in real time on the basis of location, time, etc.
[107]	Churn data set from the UCI Machine Learning Repository, California University. (https://archive.ics.uci.edu/ml/index.php).	DT, ANN, SVM, boosting, NB and Regression analysis.	SVM-POLY with AdaBoost achieved the best performance with 97% accuracy.
[114]	Data obtained from 30,104 customers from a large European telecommunication company.	Logistic regression.	They found that data preparation enhanced the performance of the prediction model by14.5%, as measured by the AUC.
[20]	Details of 15,519 and 19,919 customers, respectively, from a UK mobile telecommunication provider.	DT and logistic regression models	They stressed the need for more advanced methods of churn modelling.
[93]	Available historical records extracted from the telecom industry.	Logistic regression and DTs in R.	The data mining techniques could be a promising solution for customer churn management.

[91]	Telecom data set from an Asian mobile operator.	Model was built using WEKA. A logistic regression was used in parallel with the voted perceptron, which is ANN sole node for classification purposes and then combined with clustering for churn prediction.	The results showed that the new hybrid model is more accurate than single methods with 0.721 ROC value.
[94]	Their data set was from a PAKDD – 2006 datamining competitio [95].	Used DT and neural network for churn prediction.	They observed that the DT model surpassed the neural network model in predicting churn.
[64]	Two telecom industry data sets were considered. Type-1 contained 3,333 records, and Type-2 contained 20,468 records. Both provided telecom customer details but used different attribute sets.	Axiomatic fuzzy set theory and parallel density-based spatial clustering of application with noise – a data clustering algorithm on the Hadoop MapReduce framework.	The proposed model was more efficient than the existing system in terms of time and performance. In the future, new methodologies for churn analysis should be explored by integrating different data mining techniques and machine learning algorithms to achieve better and more efficient results.
[115]	Telecom company's billing data set.	Rule-based classification.	The result obtained is not promising because their data set was incomplete.
[11]	Online customer data set available at Kaggle.	Used different classifiers implemented in WEKA, i.e., SVM, bagging, stacking, C50/J48, PART, naïve Bayes, Baysen Net and Adaboost.	Concluded that bagging and the SMO algorithm outperform other techniques with an accuracy of 99.8% using 14 attributes.

[14]	Telecom customer churn in UCI (Machine Learning Repository) and Orange Telecom.	Cluster stratified sampling logistic regression model.	Showed that their prediction method performs satisfactorily and can be effective in forecasting telecom customer churn.
[18]	There were two data sets: one was original data and the other was statistical data.	SAS Enterprise Miner.	Found that the neural network was superior as a scoring model by overall instance of lift.
[116]	The data set was taken from the Indian telecommunication service industry.	Counter propagation neural networks, classification, regression trees (CART), J48 and fuzzy ART MAP.	To predict churning in telecommunication, the authors suggested the use of fuzzy ART MAP and CART instead of other techniques.
[117]	Historical customer data.	Open-source software framework Apache Hadoop, along with Map Reduce sub-framework and the help of NB algorithm.	Found that the Apache PIG has some disadvantages.
[13]	A total of 153,651 distinct tweets for the Twitter handles of five popular telecom brands in India.	Sentiment analysis.	Proved that SA can manage the high growth rate of new subscribers who were added to the brand in the study period.
[113]	Tweets related to Telkom's broadband Internet service and customer churn rate data history from the company's data warehouse.	Applied SA using recurrent neural network LSTM.	Results indicated that the accuracy of churn rate predictions that occur three months correlated with negative mood.
[24]	Used the Pokec social network (http://snap.stanford.edu/data/soc-pokec.html) data and generated synthetic call log details of the social network users (25,000 users of the Pokec data set).	Used influence maximisation, calculated by taking into account a user's topic of interest from the users'	Future analysis should factor in both location and language to avoid geographical and cultural sampling errors.

		social network data and call duration (CLD), together with message length from the user call log data.	
[118]	Customer data.	Data mining; spreading activation, threshold-based and decision tree-based clustering; K-Ties heuristic; J48 classifier in the WEKA tool.	Proved that social relationships play an influential role in affecting churn in an operator's network.
[119]	Telecom data set containing 3,333 pieces of customer data.	The authors made a comparison between various rules and algorithms such as Covering, Exhaustive, Genetic and LEM2 rule-generation Algorithms. This study used rough set theory (RST) to predict customer churn.	Their approach of using rough set theory with Genetic Algorithm achieved high performance – 98.1% accuracy. Their data set has some drawbacks: 1. Unbalanced data set, which would affect classifier performance. 2. Did not consider influence of customer profile on churning predictions and that would affect the decision makers' decisions about keeping a customer.
[102]	5,000 pieces of customer data from telecom provider.	K-means collective with Naïve Bayes.	Their proposed approach achieved high accuracy. Using other methods like SVM, DT and NB may overcome the boundaries between true positive and true negative.

[120]	5,000 customers' data provided by wireless network telecommunication company.	Random Forest, DT, C4.5 and AntMiner+.	Their proposed approach achieved better performance in terms of specificity than DT, C4.5 and AntMiner+.
[121]	Bank data set.	SVM, NBTree and SVM AdaBoost	Their approach achieved high result.
[100]	Data set of employee and customer information for a year and a half.	SVM, Neural Network, DT and NB	Their conclusion was to use the SVM to develop an accurate churn prediction model.
[122]	600,000 customer records from South Asian telecom company.	Comparison between different methods such as Linear regression, SVM and DT with fuzzy classifiers such as VQNN, FuzzyNN, FuzzyRoughNN and OWANN	Their conclusion was in favour of using a Fuzzy classifier to predict customer churning.
[123]	Customer data including real phone call records from China Mobile Communications Corporations (CMCC).	Different classifications based on PB, e.g., Particle Swam Intelligence PBCCP, BP and PSOBP algorithms	The PBCCP algorithm performed better than other algorithms to predict customer churn.
[124]	100,000 pieces of customer data provided by a South Asian telecom operator.	Used brute force to find features that could be used as an prediction input for supervised learning algorithms.	Their approach achieved high accuracy with 90%.
[97]	The data set was gathered from the web repository http://www.ics.uci.edu~mlearn/MLRepository.html.	Random Forest, Naïve Bayes, SVM, C4.5, LIBSVM, ANN, and Probability and Gaussian Weighted Integration.	Random forest performed better than other classifiers. They plan to use deep learning networks in future.

[105]	Data from New China Life Insurance Company LTD.	Deep and Shallow Models such as LSTM, CNN, Random Forest and AdaBoost.	They concluded that combining deep and shallow models enhances performance more than deep model and shallow model on their own. In the future they will use different shallow and deep models.
[125]	6,000 pieces of user data from an E-Commerce platform.	Logistic regression with EBURM model	The results show the accurate performance of the model.
[99]	The data sets were gathered from the machine learning UCI database, University of California. The data is about home telecommunication.	SVM	SVM has a better performance in terms of accuracy than other classifiers such as Logistic Regression and Naïve Bayes.
[126]	100,000 pieces of customer data from Malaysian telecom companies.	Data mining by evolutionary learning (DMEL)	The result proved that DMEL effectively predicted customer churn in the telecom industry.
[62]	160,000 pieces of customer data from Taiwan telecom company.	Neural Network back propagation (BPN) and DT.	BPN achieved a higher performance than DT.
[112]	60 GB of data of the largest telecom company that include detail about voice call details, SMS details, etc.	J48 Decision Tree, and other classifiers packaged in WEKA.	Using social network analysis, the authors accurately predicted customer churn.
[127]	65,000 pieces of customer information, Duke University.	SVM-RFE in MATLAB	Their results showed that SVM-RFE predicted customer churn acceptably.
[128]	Different data sets from bank, supermarket, telecom company, TV, and newspaper. 100,205 pieces of customer data in the Telecom Data Set.	Logistic Regression and Random Forests packaged in WEKA	Their results showed that using the under-sampling technique enhanced the performance of the prediction model.

[129]	895 pieces of call record data from a telecom company.	Data mining using Back Propagation Neural Network algorithm in MATLAB	Predicted customers at risk of possibly churning.
[104]	2,427 customer records from UCI Repository, University of California, Irvine.	Data mining using ANN in SPSS	Their proposed approach predicts customer churn with 92% accuracy.
[130]	5,000 customer records from UCI Repository, University of California, Irvine.	C4.5, AntMiner+ and Ripper packaged in WEKA	The results showed that C4.5 and Ripper achieved highest accuracy.
[131]	5,000 pieces of customer data from UCI Repository, University of California, Irvine.	Adaptive Neuro-fuzzy Inference system (ANFIS), C4.5, and Ripper packaged in MATLAB	Neuro-Fuzzy performance outperformed C4.5 and Ripper
[132]	5,000 pieces of customer data from mobile service provider.	SVM, DT, and neural network packaged in WEKA.	The Neural Network and SVM achieved the same accuracy with 83.7%.
[109]	50,000 pieces of customer information in telecom data publicly available for Orange large & Cell2Cell	Random Forest, Rotation Forest, RotBoost and Decorate ensembles, concurrently with minimum redundancy and maximum relevance (mRMR), Fisher's ratio and F-score.	MRMR proved that it has more suitable features than Fisher's ratio and f-score. MRMR with RotBoost accuracy performed best with 0.761% AUC.
[133]	3,333 pieces of customer data from UCI Repository, University of California, Irvine	SVM packaged in IBM SPSS	SVM achieved 88.56% for accuracy.

[134]	89,412 pieces of customer information including personal information and call data record	Logistic regression and multilayer perceptron neural networks packaged in MATLAB	Using SPA method as the propagation process to generate more input variables enhances the performance of the traditional machine learning model, which depends on the historical data of a customer in the company data base.
	Their data set was from an American telecommunication company.	They combined a self-organized map (SOM) and an artificial neural network (ANN).	The results showed that the hybrid approach enhanced the prediction accuracy more than using a single neural network. This study addressed a limitation, which is that data reduction causes loss in the training set.

Table 2.4: Synthesis of the Included Studies related customer churn and social media mining.

References	Technique
[96], [135], [106], [136], [94], [18], [20], [107], [93], [137], [138], [139], [132], [140], [138], [141], [85]	DT
[20], [107], [93], [91], [140], [139], [85], [135], [142], [114], [96], [138], [14], [143]	Logistic regression
[113], [107], [96], [91], [94], [116], [138], [132], [140], [139], [85], [133], [144]	Neural network
[107], [11], [132], [133], [99], [145], [146]	SVM
[11], [116]	J48
[116], [11]	CART
[117], [107], [11], [138], [140], [133], [138]	NB
[116]	Fuzzy Classification
[115]	Rule-based classification.
[132], [139]	k-means algorithm
[106], [147]	Random Forest

Table 2.5: Most common techniques used for customer churn prediction models.

As elaborated above, different techniques and approaches are used for customer churn prediction. However, I cannot conclude which technique is the best to address the customer churn prediction problem. For the research community, this is still an open question.

2.2.5 Saudi Telecom Companies

With the emergence of new technologies, the telecom field has changed accordingly. This is the case with the telecom market in the KSA, which expanded in 2003 by attracting new investors. As a result, the Saudi telecom market became a viable market [148]. The current Saudi telecom market is dominated by three telecom companies: the Saudi Telecom Company (STC), the Etihad Etisalat Company (Mobily), and Zain KSA. In addition, the market also hosts other internet service providers and mobile virtual network operators.

During 2019, STC performed well, and it became the leading provider of digital services in Saudi Arabia: STC revenues increased by 4.63%, equivalent to SR 54,368 million [149]. Mobily was launched in 2004. It covers a wide area in Saudi Arabia and is one of the biggest wireless networks in terms of coverage in the country [150]. Additionally, its data centre system is one of the largest worldwide [150]. The third telecom company in Saudi Arabia is Zain. It had a strong enrolment in the Saudi market and by 26 August 2008, just four months after its launch, it had more than 2,000,000 customers [151].

2.2.6 Sentiment Analysis

This section reviews significant research on Arabic Sentiment Analysis, but it starts with English language Sentiment Analysis studies. Research on SA in English texts began in 2002 [152], [153] using different learning methods and focusing on reviews as data sets. With the emergence of social media, studies on Twitter SA began in 2009 by exploiting of supervised machine learning classifiers.

Go et al. [154] conducted one of the earliest Twitter SA studies. They used emoticons to collect tweets, classified as 177 negative tweets and 182 positive tweets. They applied three machine classifiers: NB, SVM and maximum entropy. The accuracy ranged between 80% and 83%. Subsequently, [155] conducted a similar study but with a larger corpus (30,000 tweets) and with the addition of a neutral label to the classification labels. They searched for emoticons in tweets and then classified the tweets based on them. They used three machine learning classifiers – NB, SVM and conditional random fields – and their best result was achieved using the NB classifier. These studies relate to this research study on using Twitter as a data set with NB and SVM as classifiers.

Similarly, [156] collected their data by utilising 15 emoticons and 50 hashtags to label tweets. They performed a k-nearest neighbours (k-NN) classification, where the F-score was 86. Paltoglou and Thelwall

[157] and [158] proposed the SentiStrength approach, which depends on lexicon-based methods. It deals with informal language use on Twitter, such as negation, capitalisation and emoticons, by applying linguistic rules. It works by ranking the strength of tweets from 1 (not positive/negative) to 5 (extremely positive/negative). They achieved an F-score of 86.5 for the classification of the data set obtained from Twitter.

Two years after the development of the SentiStrength approach, [159] and [160] produced SentiCircles, a platform for SA that considers the contextual and conceptual semantics of words. The platform's novelty lay in its consideration of a term within its context. The approach used in SentiCircles outperforms that used in SentiStrength and the average F-measure for SentiCircles was 65.98.

SemEval is an international shared-task workshop on semantic evaluation. It holds an annual competition to encourage participants from around the world to evaluate semantic analysis systems as well as to release a sentiment lexicon and annotated data sets. Its proceedings synchronise with one of the Association for Computational Linguistics conferences. Some of the SA research studies on Twitter tasks conducted from 2013 to 2020 are reviewed below.

Since SemEval 2013, SA research has focused on English, [161] created two classifiers to detect sentiments on Twitter using two tasks: message-level and term-level tasks. The team won first place out of 44 teams during the SemEval 2013 Task 2 competition. They used three lexicons: the Multi-Perspective Question Answering (MPQA) lexicon [162], the Bing Liu lexicon [163] and the National Research Council (NRC) emotion lexicon [164] [165]. They automatically generated two lexicons from Twitter, one with sentiment-word hashtags and the other with emoticons. Their F-scores were 69.02 for the message-level task and 88.93 for the term-level task. Their observations indicated that the sentiment-lexicon features were the best.

Similarly, SemEval 2014 replicated the work performed in SemEval 2013, [166] presented a state-of-the-art SVM classifier for a Twitter SA, which was placed first in the SemEval 2014 competition for Twitter SA tasks. The highest F-scores achieved were 70.45 for the message-level task and 89.50 for the term-level task.

SemEval 2015 Task 10 was used for determining Twitter SA [167]. This task focused only on single English words and negated two-word expressions. There were 41 teams across five subtasks. In the SemEval 2015 Twitter SA task competition, [168] were placed first on the phrase level and second on the message level.

Their system was based on neural networks, and they achieved scores of 84.49 for the phrase-level subtask and 64.59 for the message-level subtask.

SemEval 2016 Task 7 [169] and SemEval 2017 Task 4 [170] have focused on English and Arabic Twitter SAs. Refaee and Rieser [171] were the winning team for the Arabic task. the winning team for the Arabic task. They observed that Arabic Twitter data set results were lower than those for a similar English Twitter data set in 2015. Moreover, the results for single words were higher than those for phrases, especially for the Arabic Twitter data set. They produced a publicly available SA tool for Arabic tweets, which collects tweets from Twitter under certain queries and then classifies them according to three sentiment labels: positive, negative and neutral.

Regarding SemEval 2018 Task 1 [172], the task includes five subtasks. The third and fourth subtasks are for Valence Regression (V-reg) and Valence Classification (V-oc). The data sets for these two subtasks included 2,600 English tweets and 900 Arabic tweets. Seventy-two teams participated in for the two subtasks for the English data set and 26 teams for the Arabic data set. The winning team for both tasks using the English data set was [173], while the winning team for both tasks using the Arabic data sets was EiTAKA [174].

SemEval 2019 Task 3 [175] concerned classifying the sentiment of a text into four emotion classes, while SemEval 2020 Task 9 [176] concerned the SA of Code-Mixed Tweets. The released corpora were for the Hindi-English and Spanish English Languages.

With the advent of big data and the proliferation of social media (e.g., Facebook, Twitter), SA has observed an increase in academic research over the past decade. SA or 'opinion mining' refers to the computational processing of opinions, feelings and attitudes towards a particular event or issue [27], [28]. To identify subjective opinions in sources, SA applies natural language processing (NLP) and textual analytics techniques [177] [178]. SA helps to reveal the polarity of texts by identifying whether a fragment of the text indicates a positive, negative or neutral impression [179, 180].

Subjective Analysis can classify the text into subjective or objective, where subjective text has opinions and sentiment, and objective text has facts [162]. Sentiment analysis can classify the text based on many ways, many-way classification, i.e., binary classification (positive or negative), or three-way classification

(positive, neutral or negative) [181, 182]. It can be employed through two approaches, flat classification or hierarchal classification. In flat classification, the classifier classifies the text in many-way classification on one level. In hierarchical classification, many classification layers used, usually in the first layer, then the text is classified as objective or subjective, and the subjective text is classified according to its polarity.

SA has been investigated at three levels: document level, sentence level, and entity and aspect level [183]. Document-level considers the whole text as one unit that holds opinions, such as product reviews, while sentence level deals with each sentence as one unit that holds sentiment. Usually, sentence-level classification is applied to short texts in social media such as Twitter [184], [185]. On the aspect level, SA is carried out on an entity.

Sentiment analysis has been, and is still, a thriving research area. However, the task of Arabic sentiment analysis is less represented by the body of research [44], [45], [39], [46]. This section offers an in-depth analysis of existing ASA studies of textual content and, it identifies their common themes, domains of application, methods, approaches, technologies, and algorithms used. A total of 133 ASA papers published in the English language between 2002 and 2020 were identified in four academic databases and one other source. The papers were screened and analysed, with the results identifying 133 papers related to Arabic text SA. Their contents were carefully analysed, and our study presents the different approaches used to conduct this analysis.

Social media sites have become popular in recent years, and since then, the SA approach has grown to become prominent for capturing public opinion. Doing so can improve the effectiveness and efficiency of decision-making by using textual analytics to make better-informed decisions [179], [186]. SA has been adopted in a wide range of fields, including marketing and e-commerce, customer relationship management, market intelligence, strategic planning, political polls, employment, sociology, health care, education and scientific research, and humanitarian assistance and disaster relief [28], [187]. SA plays a vital role in obtaining realistic information related to public opinion. For example, SA helps to determine customers' preferences and evaluate their satisfaction with products on e-commerce sites like Amazon, thus improving quality and standards based on the actual needs of customers [75].

SA has been applied from different perspectives, either for general [177], [188], [189] or specific challenges [28, 190], [191] techniques [178], [179], [187] and languages [191], [192], [193]. In the context of language, the majority of the research pertains to English rather than Arabic SA [194], [41]. Both languages differ in their expressive power of sentiments, which makes the detection of sentiment polarity considerably more complex [188]. This issue is particularly challenging given that natural languages are unstructured, making the interpretation of sentiment a tiresome task [195]. There are important studies that have handled this problem in the English language [28], [189], [196], but it remains largely unexplored concerning Arabic [188].

Arabic differs from English in several key aspects. Arabic is a rich morphological language [34], [35], written from right to left, using different forms, thus presenting researchers with specific challenges. Arabic has many forms, which are Classical Arabic, as in the book of Islam's Holy Quran, Modern Standard Arabic used in newspapers, education and formal speaking, and Dialectical Arabic, which is the informal everyday spoken language, found in chat rooms and social media platforms. The Arabic language consists of 28 Arabic alphabet letters. Additionally, there are 10 letters with a second form [46], Table 2.6. To represent the meaning of an Arabic word, diacritics are used, which are small signs over or under letters positioned to reflect the vocals. The absence of these Arabic diacritics in the DA text makes text interpretation more complicated. Moreover, DA forms differ from one Arab country to another, making understanding a specific DA difficult for the people not speaking that DA. Mubarak and Darwish [36] have defined six Arabic dialects: Gulf, Yemeni, Iraqi, Egyptian, Levantine and Maghrebi. This makes the SA process more complex, for instance when attempting to build an Arabic lexicon [34], [38], [197], [198]. In addition, the mix of Modern Standard Arabic and DA employed by Internet users [199], [40] presents challenges which have resulted in limited research on ASA [200], [201].

28Arabic alphabet letters	أ,ب,ت,ث,ج,ح,خ,د,ذ,ر,ز,س,ش,ص,ض,ط,ظ,ع,غ,ف,ق,ك,ل,م,ن,ه,و,ي
10 letters with a second form	ء (أ,إ,ء,ى,ئ,ؤ) هـ (ه,ة) ا(ى,آ)
Basic diacritics	ً,ٌ,ٍ,َ,ُ,ِ,ّ,ْ

Table 2.6: Arabic Letters.

Some studies have identified the need to create a comprehensive resource for the various Arabic dialects that exist, combining their morphological analysis and tokenisation into one process. Doing so may resolve the Arabic tokenisation issue and also improve the processes when conducting SA and opinion mining [28], [34] [40]. Although ASA is of growing importance, it is still in the early stages of research [41], [202], [43].

Early ASA research addressed SA of newswires [203], [204], whereas the most recent studies focus more on ASA of social media [43],[205], [206], [170]. Although many survey studies extensively address SA in the English language [207], [208], [209], [210], [196], ASA survey research is still relatively modest [202]. Some ASA research addresses specific issues, such as creating an Arabic lexicon [211, 212], while others focus only on specific SA techniques [213], [214], [215], [180]. However, these studies provide narrow insights into ASA; they do not comprehensively address ASA in general [202]. Thus, ASA remains largely unexplored [170].

In this systematic review, the contributions are categorized as: providing a comprehensive review of ASA studies published in the current literature; capturing a holistic view of the most significant approaches, tools and resources used in ASA research; and assessing the most significant challenges identified in the reviewed studies and proposing suggestions to overcome the challenges.

Identifying the review questions is the first step in a systematic review. The research questions of this study are as follows:

RQ1. What is the current stage of research related to ASA?

RQ2. What are the most effective approaches, tools and resources used in ASA?

RQ3. Which are the most significant challenges identified in the reviewed studies?

RQ4. What are the suggestions for overcoming the challenges?

2.2.6.1 Literature searches

To identify relevant studies, a systematic review of the literature (as recommended by [216]) was conducted in four different academic databases (SAGE, IEEE, Springer, WILEY) and on Google Scholar up to June 2020. I applied the following keywords in our search strategies within the database searches: 'Arabic semantic analysis', 'Arabic subjective analysis', 'Arabic emotion detection', 'Arabic text categorization', 'Arabic opinion mining', 'Arabic lexicon', 'Arabic corpora', 'Arabic sentiment analysis', 'Arabic sentiment classification', and 'Arabic Opinion Mining'. Some terms were excluded, such as 'Arabic indexing', 'information retrieval' and 'code-switching'.

2.2.6.2 Study selection

I applied the following manual, careful search, and review strategy:

1. I reviewed the studies at the title and abstract level, after eliminating the duplicates.
2. The remaining articles were evaluated in detail at full-text level and were included if the reviewer identified them as relevant. The appraisal was carried out using the following inclusion criteria: selected studies:
   (a) reported the application of text ASA, and
   (b) were written in the English language. Studies were excluded if they reviewed something other than text ASA, such as speech, voice or images.

2.2.6.3 Data extraction and analysis

Data extraction was performed as follows during the review of the included studies and ASA methods. The extracted data are synthesised and presented in Subsection 2.2.6.4. The information gathered from these syntheses was used to find the common themes in this review.

2.2.6.4 Searches and sifting results

702 potential candidate studies were identified via the search strategies, with 687 studies remaining following removal of duplicates. Step 1 of the search and review strategy was conducted through the screening of titles, abstracts and methodologies of the included studies. A total of 554 studies were thus excluded, as these did not meet inclusion criteria. The second level of the review consisted of a detailed assessment of the remaining 133 studies. No studies were excluded at this stage, as all met the inclusion criteria set for this review. The result from the two stages of sifting is presented in a Transparent Reporting of Systematic Reviews and Meta-

Analysis (PRISMA) diagram [3] (the reporting components most used for systematic reviews) in Figure 2.2. In order to enhance readability and reduce diagram complexity, guidelines [3] were referred to when designing the diagram.

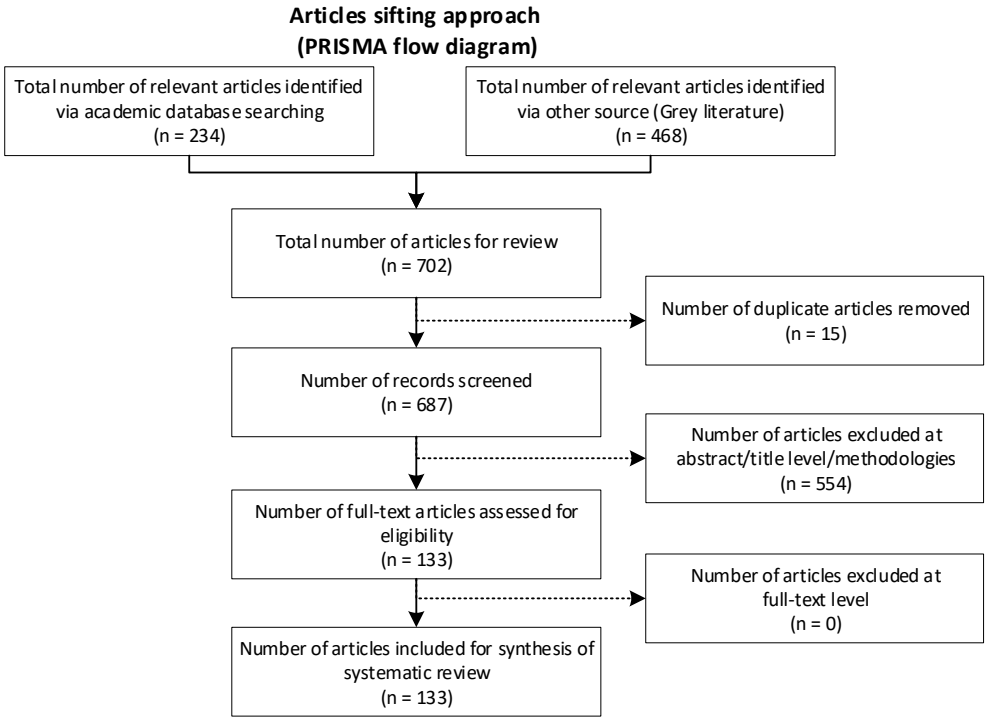

Figure 2.2: PRISMA diagram [3] of the ASA Literature Filtering Process.

The most frequent terms related to the topics examined in our first stage (Step 1) of screening were, unsurprisingly, 'Arabic', 'sentiment', 'mining', and 'opinion' (Figure 2.3). Meanwhile, more informative terms appear in the second stage (Step 2), as these showcase the languages correlated with Arabic, vis-a-vi English and Urdu. This analysis further indicates the domains involving ASA research, i.e., finance and news, and the topics of concern in these papers, which were linguistics, corpora and lexicons (Figure 2.3A, 2.3B) and other topics not further discussed in this survey, such as feature engineering.

A **B**

Figure 2.3: Word cloud depicting the most frequent words appearing in step one (A) of the ASA screening.

2.2.6.5 Overview of the included studies

The outcomes of our review of selected studies are presented in this section to highlight the similarities and differences between state-of-the-art ASA studies. Our findings indicate that ASA is applied to a range of various domains, including health services [217], [218], [219], telecommunication services [51], [9], [139], customers' satisfaction with e-products [220] , government services [221], security [222], volunteer work [223], politics [224], [225] and finance [226].

Findings also indicate that common approaches used to implement ASA (2002 to 2020) include machine learning approach (Figure 2.4). It is noted that the hybrid approach (mixing machine learning with the unsupervised learning) first emerged in 2010 and was frequently used until the year 2019. From 2005 to 2020, there was a prevalence of machine learning approaches, specifically supervised learning. Between the years 2017 and 2019, the application of machine learning rose drastically, more than other approaches. The use of deep learning has increased rapidly in the research community since 2017. It is evident here that transfer learning is still new to the ASA field, with only one study found in 2019 and one study found in 2020.

The following sub-sections provide a detailed analysis of the different approaches identified for ASA.

2.2.6.4.1 Supervised learning approaches

There are several learning algorithms applied based upon a supervised learning approach [227] and [228],

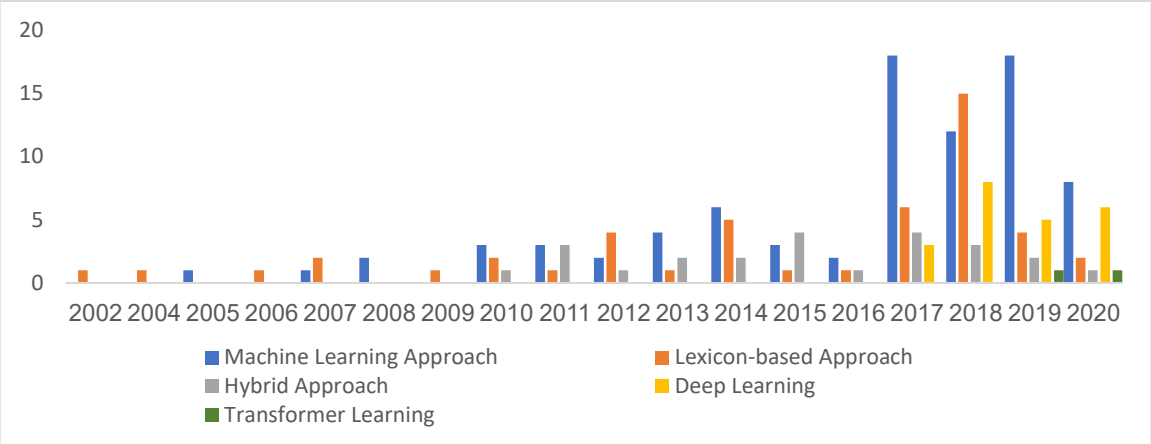

Figure 2.4: The most common approaches found in ASA literature.

which are: naïve bayes (NB), support vector machines (SVMs), decision trees (DTs), logistic linear regression, random forest, neural networks and the k-nearest neighbours (K-NN) algorithms. These algorithms were employed as the base classifiers for ASA [194].

A number of experimental studies [194], [43], [218], [229], [211] a have used different machine learning algorithms for standard Arabic datasets and dialectal Arabic. For example,_ [229] evaluated the application of NB and DTs using a multi-dataset in MSA and dialectal Arabic. This dataset consisted of 658 comments from Facebook written in English, 2648 reviews from Aly and Atiya [230] written in MSA, and 409 reviews used by [231]. Sentiment analysis was performed using RapidMiner for a two-way classification (positive, or negative). Evaluation was conducted based on two parameters (accuracy and runtime) with results demonstrating some significance. The two classifiers performed poorly on dialectical Arabic, with 50.76% accuracy for DT and 54.43% for NB. Regarding MSA, the performance was raised to 97.16% with DT and 89.52% with NB. In addition, the performance of both classifiers was enhanced on the English corpus, with 84.25% for NB and 83.87% for DT. They concluded that NB performance is higher on the English corpus, when comparing with dialectical Arabic.

Gamal et al. [43] applied different machine learning algorithms NB, AdaBoost, SVM, Ridge Regression (RR), and Maximum Entropy (ME). Their dataset consisted of 151,000 tweets written in the MSA and Egyptian dialect, balanced between positive and negative tweets. Their main finding indicated that RR with Term Frequency -Inverse Document frequency (TF-IDF) as the feature extraction method and 10-fold cross-validation achieved the best result, with 99.90% accuracy, precision, recall and F-measure.

Following this work, [232] compared different algorithms, which are NB, SVM, BNB, Multinomial NB (MNB), Stochastic Gradient Decent (SGD), Logistic Regression (LR), Maximum Entropy (ME), RR, Passive Aggressive (PA), and Adaptive Boosting (Ada-Boost), with different n-gram features using 10-fold cross validation on their dataset [43]. Results favoured the unigram feature set, with PA achieving 99.96% for precision, recall, accuracy and F-measure.

In the work of [233], a comparison of linguistic and statistical features was compared between SVM, KNN and ME. Linguistic features included stemming and part-of-speech (POS) tagging, whilst statistical features included TF-IDF. Their dataset consisted of 10,006 tweets labelled with (positive, negative, neutral and objective); where 'objective' means a tweet without opinion and 'neutral' is a tweet with positive and negative opinion in the same tweet. They concluded that SVM outperformed the other classifiers by obtaining 75.21% for precision, 72.15% F-score and 69.33% recall. This implied the suitability of using the suggested features with SVM.

Farha and Magdy [234] compared between SVM and NB on 6,921 reviews and comments collected from Yahoo and Maktoob social networks using 10-fold cross validation and the TF-IDF scheme. They classified comments and reviews into four categories (social, technology, science and arts) then labelled these comments and reviews using three sentiment labels (positive, negative or neutral). Much of the dataset was written in MSA and different Arabic dialects (Egyptian, Khaliji, Levantine, and Arabizi - which is a combination between MSA and English). They concluded that SVM was superior to NB for their unbalanced dataset, obtaining 64.1% for accuracy and recall and 63.8% for precision.

SVM emerged superior to other algorithms on different datasets, such as that of [235], who proposed the largest offensive words Arabic dataset extracted from Twitter, based on different Arabic dialects. Their dataset contained 10000 tweets labelled with four labels (clean, hate speech, vulgar or offensive). The dataset

was evaluated using the different algorithms; DT, RF, Gaussian NB, Perceptron, AdaBoost, Gradient Boosting, Logistic Regression and SVM with different pre-trained embedding. They achieved the best F1 79.7%, 88.6 recall with SVM using the Mazajak embedding [234].

In addition, [217] proposed a health services dataset written in Arabic, comprising 2026 tweets classified as positive or negative. This area of research applied different Deep and Convolutional Neural Networks and Machine Learning algorithms, such as Logistic Regression, SVM and NB. In addition, they applied the uni-gram and bi-gram as a feature-selection technique and TF-IDF as a weight scheme, with the best accuracy of 91% being achieved by an SVM with Linear Support Vector Classification (LSVC).

Bahassine et al. [236] assessed the Improved chi-square feature selection (ImpCHI), by using SVM and DT on 5070 documents classified into six classes (Business, Entertainment, Middle East, SciTech, Sport and World). SVM with ImpCHI outperformed ImpCHI with DT, obtaining 84.93%, 85.17% and 85.29% for average F-measure, recall and precision. They concluded that when the feature number was between 40 to 900 features, the ImpCHI feature selection outperformed the other feature selections, which were Information gain, Mutual information, and Chi-square.

The same feature algorithm was used by [237] who proposed an approach using a Chi-Square algorithm for feature selection and KNN for classification. They used a Twitter dataset for Arabic Sentiment Analysis [238]. It included a perfectly balanced dataset of 2000 tweets, classified as 1000 positive tweets and 1000 negative tweets. They obtained 65.00% using Chi-Square as feature selection and KNN for classification when K=3.

Another study proved the effectiveness of using KNN with ASA [239]. This work proposed an improved K-NN Arabic text classifier using word-level n-grams (unigrams and bigrams) in document indexing and compared this to document indexing based on a single term. They applied their experiment on an Arabic corpus constructed by [240] from online websites and newspapers, with the corpus being placed into the Computer, Economic, Education or Engineering categories. Their approach obtained 87%, 64% and 74% for average precision, recall and F-measure. The study demonstrated that the average accuracy from using n-grams was 74%, while the accuracy from single-term indexing was 67%, thus indicating that the use of n-grams to represent each document provides a higher level of performance compared to using a single term. In comparison, an

alternative study proved that KNN has a poor performance because of their supposition that a tweet of the same meaning would lead to the same classification [233].

Thus far, I have identified both SVM and NB as being competitively effective at supervised sentiment classification in the context of Arabic, [241], [235], and they are widely accepted [242]. Some studies have proved the superiority of using NB with Arabic text classifiers, considering it a well-performing algorithm for data mining [243], [244], with a demonstrated accuracy of 82% and 86.5% respectively for a macro-averaged precision, and 84.5% for macro-averaged F-score using the NB classifier. The dataset was 815 comments written in colloquial Arabic, sourced from two online Saudi newspapers. They manually classified the data using four labels (strongly positive, positive, negative, and strongly negative). This finding is consistent with that of [155], who concluded that NB provides a higher degree of accuracy than SVM.

On other hand, there exist many studies proving the high accuracy of the SVM [245], [245], [242], [246]. This is especially true for sentiment analysis where SVM was considered not only the best classifier [247] [152] for supervised learning, but also most efficient [248]. Aldahawi [249] demonstrated that for text classification the best results were obtained using an SVM, as this does not require parameter tuning. The architecture of SVM - inserting a hyperplane to separate between classified data is explained as being behind its effective performance [248]. In addition, [250] claimed that an SVM was superior to NB regarding accuracy, as there is no reliance upon probabilities and is suitable for high-dimension text. This success has even been reflected in graphical languages such as Chinese [6]. This presents the possibility of success in other graphical languages, such as Japanese and Arabic. In addition, some studies have stated the reason for the superior SVM performance to be the ability of SVM to handle many classes, and the vectorisation architecture of SVM, which represents the text via a good quality representation [251] and [233].

As a result, SVMs have been abundantly applied to movie reviews, while NB has been used within web discourse sites [233, 249]. Furthermore, some studies affirmed that the performance differences between NB and SVM algorithms are based on textual characteristics_ . Finally, some studies even claimed that SVM is hard to interpret [236].

It has been noticed that the supervised learning approach (particularly machine learning) is the most popular approach for ASA [44] due to the high accuracy that it provides using supervised learning [44] and [252]. However, there are some challenges which come together with this approach:

- It requires labelled training data, which is time-consuming and costly [253] and [254];
- Due to the need of labelled training data, which requires humans for the annotation, this makes the availability of high-quality datasets slight [253];
- It is domain-dependent because the model performance that was trained on a specific dataset will decrease when trained on a different dataset with a different domain [253] and [44];
- It requires a lot of features to differentiate between sentiments [233].

Some studies applied other machine learning techniques (i.e., the unsupervised approach) to identify groups, as is further described in the next section.

2.2.6.4.2 Unsupervised learning approaches

Although the supervised approach has been proven to be superior to the lexicon-based approach [183] and [255], it requires for data to be labelled, which is hard to construct. Many studies have attempted to apply a lexicon-based approach with the aim of building an Arabic version [199], [201], [42], [205], [211], [256], [257], [258], [259], [260], [261], [262].

The use of lexicon-based approaches differed in the literature for ASA. El-Beltagy et al. [263] combined the lexicon-based approach and rule-based approach to propose an Arabic Aspect-based Sentiment Analysis. This dataset consisted of 2071 Arabic reviews from government apps. The approach achieved an accuracy and F-measure of 96.57% and 92.50% respectively. In addition, [262] improved the unsupervised approach based on Arabic sentiment analysis through the use of valence shifter rules. They applied available lexicons, such as the lexicon proposed by [264], and proposed by [263], AraSenti-PMI by [45], and Arabic Senti-Lexicon by [211], etc. The research concluded that the proposed rule enhanced the classification performance by 5%. Moreover, [44] proposed a weighted lexicon-based algorithm (WLBA) of SA for Saudi dialect. The WLBA concept is to learn from the corpus and not depend upon the lexicon to calculate the weight. The algorithm subtracts the associations between sentiment-bearing and non-sentiment-bearing words, and then based on the association it calculates the weights for the word. The researchers applied WLBA to their Saudi dataset for Sentiment Analysis consisting of 4700 tweets. They compared between their proposed approach

and two different lexicon-based approaches, the double-polarity approach [265] and the simple algorithm [254]. The simple algorithm method relies on counting in each sentence the positive words and negative ones, whilst double polarity depends on the frequency of sentiment words in the sentence. The researchers concluded that WLBA performed better than the double-polarity approach, however worse than the simple method. They counted some features to enhance the performance, such as supplication (Do'aa), to capture the linguistic complications of the Arabic language. This provided a performance increased to 85.4% for the average accuracy. Results demonstrated that consideration of linguistic features in ASA is important, and not widely covered within the literature. In addition, the Saudi dataset contained a large amount of Do'aa, therefore presenting its importance of being included within a corpus. Moreover, they proved that there is a strong relation between the sentiment-bearing words and the non-sentiment-bearing words in the Saudi dialect corpus. The same result was obtained by [45] the importance of considering linguistic features, such as negation, for ASA. This research compared a lexicon-based method, a supervised method and a hybrid method.

The lexicon-based method relied on counting the positive and negative words. Their proposed approach achieved an accuracy of 91.75%. The same was performed by [257], who compared corpus-based and lexicon-based approaches for ASA. They constructed a lexicon for ASA from a seed of 300 words. Then, they added synonyms to expand the lexicon. After that, they summed all the weights for the word polarity, including the negation to the weights. They concluded that the lexicon-based method performed inadequately when the lexicon is small.

As you can see, the lexicon-based approach depends on the creation of a lexicon of good quality [44]. The major advantage of the lexicon-based approach is domain-independence, when constructing a comprehensive lexicon. However, it is hard to construct a comprehensive lexicon [182].

For building a lexicon there are two approaches in the literature [183]: manually [266] or automatically [159]. The automatic approach includes corpus-based and translation-based approaches [254], [183].

Many researchers applied a manual approach, since the manual approach provided a more accurate lexicon [44], [254], [258]. Abdul-Mageed & Diab [267] constructed manually the Sifaat, which is an Arabic lexicon with 3325 adjectives. They subsequently extended it to the Multidialectal Arabic Sentiment Lexicon (SANA). In

addition, they proved that the lexicon manually constructed was more accurate than one automatically built; however, researchers are limited in the size of lexicon they may construct [254].

A dictionary-based approach depends on using a dictionary to find the synonyms and antonyms of seeds of positive and negative words, until no word is found anymore [45]. One of the drawbacks of using the popular lexicon for translation is this approach is not accurate, due to errors in translation, or cultural variations [268], [269], [45]. Table 2.7 shows some sources and dictionaries that were used in previous studies to construct Arabic lexicons. The corpus-based approach depends upon the corpus to generate the polarity words, then uses different approaches to find the synonyms and antonyms of these words to generate the lexicon [254]. Some scholars have utilised specific algorithms to construct an automatic lexicon, such as the pointwise mutual information (PMI) statistical method [264]. Following their steps, [212] offered two sentiment lexicons, AraSenTi -Trans and AraSenTi-PMI, built from the Twitter dataset AraSenTi-Tweet [39]. They used two automatic approaches for generating the lexicon and used a simple lexicon-based approach to evaluate the two lexicons. To generate the first lexicon AraSenTi-Trans, they applied the MADAMIRA tool [270] for pre-processing the dataset. Then, they employed two sentiment lexicons: the Liu lexicon [163] and the MPQA lexicon [162].

The second lexicon was generated using PMI [271], which calculates the association between two terms, in terms of the sentiment analysis, i.e., the frequency of a word in a positive text compared to the frequency of the same word in negative text. AraSenti-Trans includes 131,342 words and AraSenti-PMI includes 93,961 words classified as negative and positive words. They applied a simple lexicon-based approach for evaluating the lexicons on three datasets RR [272], Arasenti-tweet [39], and ASTD [273]. The results showed that the AraSenti-PMI lexicon outperformed the other lexicon. The best F-avg of 88.92% was obtained by the AraSenti-PMI lexicon on the AraSenti-Tweet dataset. Regarding the AraSenti-Trans lexicon, the best F-avg of 59.8% was achieved on ASTD [273]. Compared to the manual building of a lexicon, the automatic approach requires a considerable reduction in the effort required and ensures that significantly larger lexicons may be produced [254].

Many studies have attempted to apply a lexicon-based approach with the aim of building an Arabic version [199], [201], [42], [205], [6], [211], [257], [258], [259]. A large portion of these studies used manual construction, though lexicon-based approaches, which provided highly accurate sentiment classification. Some

researchers have used available resources including Arabic WordNet [274] to build MSA lexicons [275], [276]. [277] applied a semi-supervised approach with Arabic WordNet [274] to build a MSA lexicon and achieved 96% classification accuracy. Other researchers have focused on the building of an Arabic dialect lexicon using different approaches (i.e., manual or automatic) to generate lexicons [278], [257], [265], [212], [43], [229], [256]. Although lexicon components have been used successfully for Arabic, a large Saudi dialect lexicon has not yet been fully applied to ASA [44].

I further critically reviewed the construction of Saudi lexicons, since they are within the scope of this research. [253] constructed a Saudi dialect sentiment lexicon (SauDiSenti) that consisted of 4431 words and phrases written in MSA and Saudi dialect. It is available online[3]. It is manually constructed from a Saudi dialect twitter corpus (SDTC) [279]. They yielded two annotators to extract the positive and negative terms from the corpus. They extracted 1079 positive terms and 3351 negative terms. For evaluation of the lexicon, they compared it to one of the biggest Arabic lexicons, AraSenTi [212]. The result showed that SauDiSenti outperformed AraSenTi when accounting for the neutral tweets, together with the positive and negative tweets, with 0.437% for the average F-measure. In addition, the AraSenTi outperformed the SauDiSenti, when considering positive and negative tweets, with 0.760 for average F-measure. Assiri et al. [44] provided a Saudi dialect lexicon. The lexicon includes 14,000 terms. The lexicon was constructed through three steps: first, the lexicon was expanded using seed words and a learning algorithm [265]. Secondly, the lexicon was built by [280]. In the third step, they added new words manually. Moreover, Adayel and Azmi [281] built a Saudi dialect lexicon including 1500 words (500 positive words and 1000 negative words) as a part of the hybrid approach of ASA. They employed SentiWordNet [282] to translate some words to Arabic and assign the sentiment to them. Furthermore, [283] provides a domain-dependent Saudi Stock Market lexicon (SSML). SSML contains 3,861 terms and their sentiment polarities (positive and negative) with two levels of strengths. They constructed the lexicon manually from Twitter and Saudi shares forum[4].

[3] http://corpus.kacst.edu.sa/more_info.jsp
[4] www.saudishares.net/vb/.

Label	Source for Arabic lexicon	Creator of the source	ASA works using the sources
S1	SentiWordNet (SWN)	[282]	[258]
S2	General Inquirer (GI)	[284]	[258], [267]
S3	Twitter	N/A	[265]
S4	MPQA Lexicon	[162]	[198]
S5	Liu Lexicon	[163]	[198]
S6	SentiStrength	[285]	[286], [254]
S7	Penn Arabic Treebank	[287]	[199], [246], [267]
S8	Arabic WordNet	[274]	[277]
S9	SWN3- SentiWordNet	[288]	[267]
S10	Affect Control Theory Lexicon	[289]	[258]

Table 2.7: Sources used to construct Arabic lexicons.

Lexicon	Lexicon Size	Construction Approach	Source	Reference
Arabic Senti-lexicon	3,880 terms	A term translation process was revised manually	S4	[290]
NileULex	5,953 Egyptian dialectical words and phrases	Manually	[265] and [291]	[292]
Saudi Dialect Sentiment Lexicon (SauDiSenti)	4,431 words and phrases	Manually	Saudi dialect Twitter corpus (SDTC)[279]	[253]
Large-scale Standard Arabic Sentiment Lexicon (SLSA)	35,000 lemmas	Machine learning models, with limited use of heuristics	A morphological analyser for Standard ArabicAraMorph [293] and S1	[294]
Large-scale Standard Arabic Sentiment Lexicon (ArSenL)	157,969 words	Combination of Arabic WordNet and English-based dictionary.	S8, the Standard Arabic Morphological Analyzer (SAMA) [295], English Senti Wordnet (ESWN) [282] and English WordNet (EWN) [296]	[280]
Sifaat	3,325 adjectives	Manual	S7	[267]
Arabic lexicon	1.8 million phrases	Arabic similarity graph and Manual	Business reviews from web	[259]
Polarity lexicon	3,982 adjectives	Manual	S7	[258]
Egyptian dialect lexicon	4,392 terms	Manual	S3	[265]
Lexicon that transfers the Jordanian dialect to MSA	300 words	Manual	Social websites, Internet and chat logs	[6]
Lexicon that transfers Arabizi to MSA	N/A			
Lexicon that transfers emoticons to MSA	N/A			

Ara-SenTi-Trans Ara-SenTi-PMI	2.2 million tweets	Automatic	S3	[212]
Arabic lexicon	16,800 lexical items	Integration between manual and automatic	S6	[256]
Arabic lexicon	N/A	Manual	S7	[246]
Arabic subjectivity word	2,600 human-classified comments	Integration between manual and automatic	S6 and online dictionary	[286]
Arabic sentiment lexicon	7,500 words	Semi-supervised learning	S8	[277]
SANA a dialect Arabic sentiment lexicon	224,564 entries	Automatic	S1, S2, and S10	[258]
Dialect/slang subjectivity lexicon	2,000 subjective terms	Automatic	S3	[297]
Idioms/proverbs lexicon for the Egyptian dialect	32,785 idioms/proverbs	Manual	Arabic websites	[260]
Arabic version of SentiStrength	N/A	Automatic	S1	[298]

Table 2.8: Comparison between Arabic Lexicons.

The lexicon-based approach presents some disadvantages that have been summarised as follows: For aspect-level sentiment analysis, it causes a minimum recall [211]. The lack of training in using the lexicon-based approach is not as effective as the supervised approach for sentiment analysis [266]. Due to that, the lexicon-based approach depends on the database used; this causes a lack of extensibility [258].

Due to the variety of Arabic dialects, each dialect needs a special lexicon, because of the uniqueness of its lexical information. Therefore, the lexicon-based approach is dialect-dependent, and domain-dependent with the claim that for sentiment analysis a domain-dependent lexicon outperforms general lexicons [299], [300].

However, there are some advantages from using the lexicon-based approach: No need for model training, which makes it simple [211] and [253]. Providing background information via a lexicon with the machine learning training could be optimal [211]. A lexicon-based approach provides the understanding of the impact of the theoretical framework [301].

2.2.6.5.3 Hybrid approaches

The hybrid approach is a combination between supervised learning and unsupervised learning [45], [211] and [286]. Several studies have found a hybrid approach to be the most suitable technique for SA [199], [211], [286], [302], [45], [303], [304].

Hybrid approaches can contribute to solving the shortcomings of both supervised learning and lexicon-based approaches [305]. The combination of the high accuracy from supervised learning approaches and the legibility (clear to understand and read) from unsupervised approaches makes hybrid approaches perform the best with SA [211]. Furthermore, some studies proved that the hybrid approach outperforms the supervised approach in accuracy [211].

The hybrid approach is common within the ASA research; for example, [302] applied a lexicon-based approach using SentiWordNet for ASA, alongside with a machine learning classifier. They used SentiWordNet as a feature for SVM. They concluded that using SentiWordNet as a feature for the SVM algorithm improved the term counting method by 7 times and raised the accuracy from 65.85% to 69.35%. The same lexicon SentiWordNet was used by [281] to label the tweets; then they used supervised learning, SVM, with n-grams, to classify the text. The results validated the effectiveness of the hybrid approach with 84% and 84.01% for F-measure. Additionally, the accuracy was raised using the hybrid approach over the individual lexicon-based or machine learning approaches.

In another work, [199] proposed SAMAR (a sentence-level ASA for Arabic social media genres). They utilised a polarity lexicon (PL) manually composed of 3982 adjectives labelled with (positive, negative, or neutral) on the DARDASHA and TAGHREED datasets to investigate the task of sentence-level construction with MSA and Arabic dialects. TAGREED includes 3015 MSA and dialectal Arabic tweets, while DARDASHA (DAR) includes 2798 Egyptian dialect Arabic chats from the Maktoob website http://chat.mymaktoob.com. They applied PL as a binary feature, to check chat or tweet whether they have a positive adjective or negative one that existed in the PL. The work concluded that the accuracy of the sentiment analysis was raised after applying the polarity lexicon. The best F-score obtained was 95.52%.

Some studies proposed the hybrid approach for Arabic sentiment analysis by using an Arabic lexicon to find the sentiment score of the words in a sentence, for example: [45] used rule-based knowledge to be included in a statistical method as a feature. They utilised the AraSenti lexicon [212] as a tweet-score feature for the SVM

and NB classifiers. This feature was applied using the AraSenti lexicon. They confirmed the superiority of a hybrid approach with two and three-way classifications. The same hybrid approach was used by [198] for binary ASA, three-way ASA and four-way ASA. The best model performance was 69.9% F-score for binary classification, 61.63% F-score for three-way classification, and 55.07% F-score four-way classification.

Another study used a hybrid approach [250], with a manually built lexicon to define the sentiment scores. They applied the hybrid approach on an Egyptian tweets' dataset. They validated their approach on 4800 tweets annotated as positive, negative, or neutral. The results showed that integrating lexical-based features into machine learning enhanced ASA. In addition, [211] proposed an Arabic senti-lexicon including 3880 terms classified as positive or negative. They utilised the Arabic senti-lexicon to extract the features for machine learning algorithms NB, k-NN, SVMs, logistic linear regression and neural networks. The results demonstrated that feature vectors extracted from Arabic sentiment lexicon enhanced the classifier performance, with the best macro-F-score of 97.8% favouring logistic linear regression. Similarly, [263] integrated features derived from the NileULex sentiment lexicon [292] into machine learning algorithms. The datasets used were obtained from social media written in MSA and different Arabic dialectics, such as Saudi, Egyptian, and Levantine. It has been used with Complement Naïve Bayes (CNB) [306]. The results showed that used the lexical-based features raised the accuracy of the model.

An interesting concept considered for the hybrid approach was introduced by [286], who presented the hybrid approach as a combined approach, which applied different methods sequentially, to classify the sentiment of a text. It applied two methods to classify Arabic documents, i.e., a lexicon-based one, and a machine learning method using the maximum entropy followed by a K-NN algorithm on the 8793 Arabic statements found in 1143 posts. The research constructed a lexicon by translating the wordlist from the SentiStrength software [158] from English to Arabic. As Figure 2.5 shows, the accuracy was raised from 50% to 80% using a combined approach. This hybrid approach was applied later in [286], to examine students' opinion changing in two consecutive semesters.

Alhumoud et al. [305] proved the outperformance of a hybrid approach over supervised and unsupervised approaches. The research applied a lexicon-based approach to label the dataset of 3000 Saudi dialect tweets. Then, they trained the SVM classifier on the labelled dataset. The hybrid learning results were 96% for precision, 97% for recall and 90.3% for average accuracy. In addition, [307] used the hybrid approach on the same dataset

[305] labelled with the same sentiment lexicon using two machine learning approaches, SVM and K-NN. The results demonstrated the advantage of the hybrid approach over the supervised approach with 90.5% average accuracy using K-NN and 90% average accuracy using SVM.

A similar improvement was demonstrated in another study by [308], who depended on the hybrid approach to switch the sentiment-bearing words with their consistent label in the text. The results showed that the hybrid approach surpassed the corpus-based approach, and the best accuracy (96.34%) was obtained by utilising random forest.

In recent studies, [233] presented the hybrid method as combining linguistic features and statistical features for Arabic sentiment analysis. POS and stemming were considered as linguistic features, while (TF) and (IDF) were considered as statistical features. They applied SVM, K-NN and ME. The results proved the effectiveness of the hybrid method, additionally, the superiority of SVM over other algorithms, with 72.15% as F-score.

Alternatively, [277] applied semi-supervised learning to evaluate the Arabic lexicon (Arabic SSL). They incorporated Arabic SSL into NB and SVM. They applied their experiment on the OCA corpus [309] and a book review corpus manually collected and annotated. They applied the lexicons to calculate the scores and feed it as features for ASA classifiers. Results demonstrate the superiority of NB over SVM with 97% accuracy. In addition, they concluded that the classification accuracy did not improve using the semi-supervised learning, due to ignoring other factors rather than the sentiment score, such as the order of words within a text.

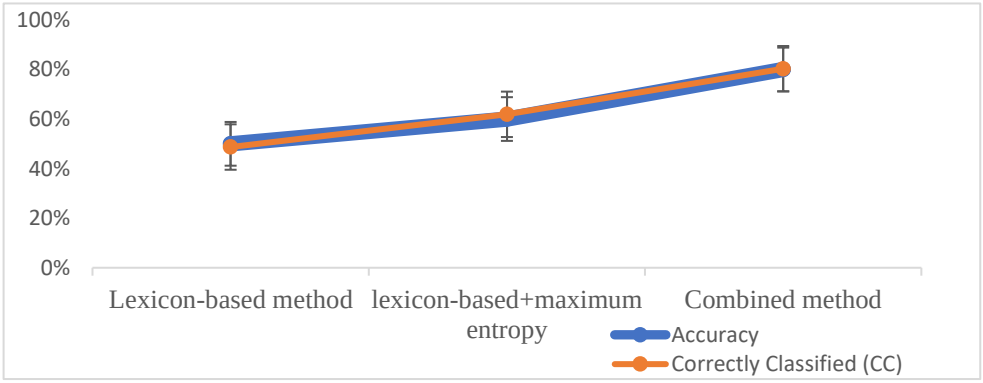

Figure 2.5: Comparison between the performances of the three methods.

2.2.6.5.4 Deep learning and transfer learning

Artificial neural networks (ANNs) are used to mimic human neurons in the brain. They hold many pieces of information, as do artificial neurons when processing a lot of information [310]. A neural network processes a large amount of information stored in the artificial neurons and ordered in layers. ANNs can be feedforward or recurrent/recursive neural networks.

The architecture of feedforward ANNs, the simplest ones, consists of three layers: the input layer L1, the hidden layer L2 and the output layer L3 (Figure 2.6). Other more complicated ANNs depend on this architecture. The input layer has many input vectors Xi and an intercept value +1. The hidden layer and output layer have neurons, each of which has an activation function – these are the computation components of ANNs. Every vector in the input layer connects with a neuron in the hidden layer using weight; it is a controlling value between two neurons, controlling the learning process by changing it. The learning process starts in the neuron by reading the output from the preceding layer, processing the data, and producing an output that is sent to succeeding neurons in the next layer. The output layer uses a SoftMax function for the last classification. There are various activation functions such as the sigmoid function, the rectified linear function (ReLU) and the hyperbolic tangent function (tanh). Lately, the ReLU function has become more widely used because it is easy to calculate. It has a quick uniting in training, and it improves the performance of ANNs [311]. It is necessary to mention backpropagation [312]; training the ANNs causes a loss of SoftMax function called cross-entropy loss. To minimize it, backpropagation should be used to make a gradient descent. One example of a feedforward ANN is a CNN [313].

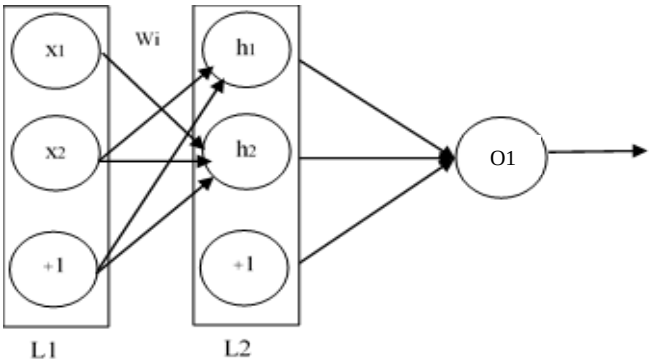

Figure 2.6: Feed forward neural network.

In the late 1990s, interest in the use of ANNs waned because the research community focused on simple ANNs with a maximum of two layers. This was due to the difficulty and high cost of training a neural network with more layers, i.e., a "deep" neural network. However, in recent years with the accessibility of computing power, especially graphics processing units (GPUs), the use of deep learning models became state of the art [314] in many tasks such as speech recognition [315], computer vision [316] and NLP [317], [318], and, in recent years, for SA tasks [313]. Deep learning models used feature extraction in the sequential multiple layers, starting with the lower layer for simple features and progressing to higher layers for more complex features. The lowest layer is the word count vector while the highest layer is the binary classification of the learning process.

Today, deep learning becomes a widespread approach in the NLP community [319]. The deep learning mechanism depends upon multiple hidden layers to represent the data, especially with large datasets. Examples of deep learning networks include Convolutional neural networks [320], and Recurrent neural networks (RNNs) [312]. CNN is a feed-forward network mostly applied in computer vision [321]. While, RNNs are applied with sequential data [321].

Using deep learning algorithms means that the changes in the input data are constant because the algorithms use an abstract interpretation. The advantage of deep learning models is that they use an uncomplicated model to achieve complicated functions, as they extract the deep learning models as a nonlinear feature, which is used as input into a linear model [29]. Furthermore, they use a huge volume of data (Big Data) effectively, dealing with the variety of data formats by using abstract data. This reduces the demand for feature extraction and the deep learning model can learn complex features by itself, whereas other simple machine leaning algorithms such as SVM and DT cannot extract complicated features [313].

A Recurrent Neural Network (RNN) [322] is another type of ANN. It is employed for processing sequential information because it has a memory that can process a long sequence of inputs, unlike feedforward ANNs. The memory allows RNN to do the same process for each component in the sequence, so the output depends

on all the prior calculations. RNN remembers the information that has been processed before (see Figure 2.7).

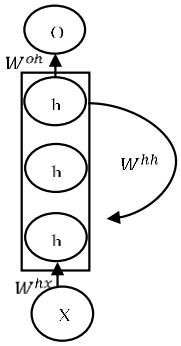

Figure 2.7: Unfolded recurrent neural networks.

Every term in a sentence being processed in RNN is considered to be one time-step (layer). For example, RNN processes a sentence with four words through four layers. One of the advantages of using RNN is that there are fewer parameters than with a feedforward ANN because it uses the same parameters in each layer with different inputs. On the other hand, the most important drawback of RNNs is the vanishing gradient problem [323], which is caused because the RNNs practically cannot handle really long sequences of information [313].

To overcome the shortcomings of RNNs, special types of RNNs have been developed: the long short-term memory network (LSTM) [324], the bidirectional RNN [325] and deep bidirectional RNN. Bidirectional RNN considers the output element depends on the previous element and next element to predict. Deep bidirectional RNNs apply the same idea as bidirectional RNNs except that for every time-step they use a multiple layer, which needs more training data. LSTM [324] is usually utilised in sequential data [313]. LSTM is more complicated than simple RNNs because instead of one layer for each word in a sentence it has four layers interacting with each other. In addition, there are two states: hidden and cell. At time-step/layer (t), LSTM decides which information it will forget based on a forget gate (ft), which is a sigmoid function σ. The function takes the previous hidden layer output, ht−1 and the present input, xt. The output will be [0 or 1], where 0 means “forget”, and 1 means “keep”.

Two steps are carried out by LSTM to update the cell state Ct−1. In the first step, the input gate (layer), which is a sigmoid function, decides the values to update. The second step produces a vector for a new value Ct

using the tanh function/layer. The new value Ct will append to the cell state. The output of LSTM depends on the cell state; first, LSTM decides which information to pass to the output gate based on a sigmoid layer. Second, the cell state passes to the tanh function and multiplies with the sigmoid gate output.

It is worth mentioning that LSTM can overcome the vanishing gradient through the forget gate, which lets the memory update and removes. The gated recurrent unit (GRU) [326], [327] is the same as the LSTM architecture except it combines the two gates "forget" and "input" into one "update" gate. It is less complicated than the LSTM model and widely used [313].

In the area of sentiment analysis, many scholars proved the deep learning models efficiency [328], [329] [330], [331]. Recently, a number of studies have investigated the use of deep learning models for ASA [217] [218], [332], [333], [205], [334], [335], [336], [337] ,[338], [339], [340, 341], [342], [343], [344], [345], [346], [347], [348].

CNNs proved good results in many NLP researches [219], due to the structural attributes of CNNs. In the literature, [349] combined CNN with word-embeddings to classify tweets, with results demonstrating the success of this approach. A number of researches applied this method, such as [350], [218], [340], [341] and [345]. CNNs capability in choosing excellent features was lauded [218]. In addition, CNN was shown to decrease the number of weights within a model and accordingly decrease complexity [321].

The traditional RNN was seen to struggle during processing of long sequential data [351]. The proposing solution was using the Long Short Memory (LSTM) and Gated Recurrent Unit (GRU) [352, 353] because of the capabilities of the LSTM and GRU in processing long sequential data [218] and in their abilities of inclusiveness in learning – i.e., including the previous output [354]. Thus, the most common RNN models used are Long Short Term Memory LSTM and Gated Recurrent Unit (GRU) [336], [333], [332], [29], [338] [337], [334], [217, 340], [341], [345].

One of oldest works using deep learning models for ASA [355], applied a CNN for aspect-based sentiment analysis for multilingual analysis, as a part of SemEval-2016 Task. The work obtained an accuracy of 82.72% for ASA. Alayba et al. [217] presented a health dataset written in Arabic, which included 2026 tweets classified as positive and negative labels. Different Deep and Convolutional Neural Networks DNNs and CNNs and

Machine Learning algorithms were used, such as Logistic Regression, SVM and NB. Within this research, the best accuracy was obtained by an SVM with 91%, closely followed by 90% achieved by a CNN. The same dataset was used later by [218]. They examined the integration of a CNN and LSTM approach to ASA. The study aimed to improve the ASA accuracy using their dataset of Arabic health service, with results proving that an integrated approach improved the sentiment classification with 94% for accuracy.

Alwehaibi and Roy [333] applied Long Short-Term Memory Recurrent Neural Network (LSTM-RNN) to ASA using three different Arabic word2Vec: AvaVec, ArabicNews and AraFT. The greatest accuracy was achieved by AraFT (93.5%), followed by ArabicNews (91%) and AraVec (88%). The results also demonstrated that a pre-trained word-embedding approach enhanced the performance of the model.

The same method was applied by [356], who employed GRU and CNN on tweets written in MSA and Arabic dialect. In addition, they used word embedding. Mohammad et al. [172] focused on the third task in Semeval-2018, which is about defining the intensity of the sentment. Their data came from Twitter in three different languages including Arabic. They used word embedding as a feature. The best results were achieved by CNN, LSTM and Bi-LSTM.

Al-Smadi et al. [338] carried out the application of an LSTM for an aspect-based SA using Arabic reviews of hotels, outperforming the state-of-the-art method. The same steps were followed later by [357], who applied LSTM on aspect-sentiment analysis with two different settings. The first model is a bidirectional LSTM with a character level and conditional random field (Bi-LSTMCRF). The second one is an aspect-based LSTM. Their dataset contained hotel reviews written in Arabic. They employed two embedding features: character- and word-level. Their results outperformed prior research to theirs.

Recently, pre-trained language models have achieved good results with different NLP target tasks, due to the ability of these models to learn with few parameters [358], unlike the previous approaches that depended on features [359]. The advantage of these pre-trained models is that they can be trained on a large general domain data set to acquire the language characteristics, and then fine-tuned on a small data set. They therefore do not depend on large data sets.

The whole idea of transfer learning is to transfer the parameters [360]. The simplest transfer model, word embedding (WE), is word2vec [361]. WE deal with each word as vector, and it is a valuable

technique to obtain a numeric feature from words. It maps neighbouring and analogous words in the high-dimensional space. The most popular WEs are Word2Vec [361], Global Vector for word representation (GloVe) [362] and FastText [363]. Word2Vec is a neural network that comprises one input and hidden and output layers. It concentrates on the distance of words, and it represents sentiment analogy between words. Training these WEs entails a massive corpus, and it takes a long time. However, there are few publicly available pre-trained WEs for Arabic NLP. There are two Arabic pre-trained WEs in [364]: AraFT [365] and Arabic_news [366]. In addition, there is AraVec [367], which depends on different models of Word2Vec: continuous bag of words (CBOW) and skip-gram models. CBOW model is a bi-gram model that predicts one word that is most likely to be the following word [361], while the skip-gram model [368] is the reverse of CBOW – it finds the possibility window of words for each word.

AraVec and Arabic_news used Word2Vec with CBOW to train on an Arabic Wikipedia corpus for AraVec and various Arabic corpora for the Arabic_news model. AraFT used FastText with skip-gram. The literature proves that CBOW performs better than skip-gram [361]. Usually, WE is the first layer of a transfer learning model [360].

The most important transfer learning model is OpenAIGPT [369], which is a language model that has achieved the state of the art in textual entailment and question answering [370]. The ULMFiT pre-trained language model [359], which is composed of three "AWDLSTM" layers [371], is very accurate on different NLP tasks. ULMFiT performed very well with different NLP tasks. The newest language model is Pre-training of Deep Bidirectional Transformers for Language Understanding (BERT) [372]. It uses a transformer network [373]. BERT outperformed the other pre-trained language models, due to its ability to manipulate context from both directions. Another pre-trained language model is RoBERTa [374], which is an enhanced version of the BERT model [372].

In the parameter-transfer method applied to the NLP field, Word2Vec [29], which is a simple transfer technology only for the first layer of the model, has many applications. It has a great impact in practice and can be used in many advanced technologies.

The use of transfer language models is still new for ASA studies. Only a few studies that have been published so far [375], [358], [376]. Al-Twairesh & Al-Negheimish [375] used the BERT model [372] on an Arabic tweet data set. Their generic and sentiment-specific word-embedding model outperformed the BERT model. They explained that this was because the BERT model was trained on Wikipedia, which is written in MSA, whereas dialects are used on Twitter. Other research studies have used Arabic word embedding. Zahran et al. [377] using a Word2Vec model on the AraVec data set [333], [356], [172], while [314] used LSTM and CNN with doc2vec to enhance the performance of SA for a financial site (i.e., Stock Twits). The results found that a deep learning approach helped to improve the accuracy of the financial SA.

HULMonA [358] is the first Arabic universal language model. It is based on ULMFiT [359]. It was pre-trained on a large Arabic corpus and fine-tuned to many tasks. It consists of three stages: 1) training an AWD-LSTM model [371] on an Arabic Wikipedia corpus; 2) fine-tuning the model on a destination corpus; and 3) including a classification layer for text classification. The results showed that hULMonA achieved state of the art in ASA.

The most recent Arabic universal language model is AraBert [376]. It is a BERT-based model trained on different Arabic data sets. It used the BERT basic configuration [372] except it added a special pre-training phase prior to the experiment specific to the Arabic language. It tried to find the solution for the lexical sparsity in Arabic [350] by using "ال" "Al" before the word (a prefix without meaning) and by using a Fast and Accurate Arabic Segmenter (Farasa) [378] to segment the word.

2.2.6.5.5 Corpora

Compared to other languages, Arabic lacks a large corpus [44], [45], [39], [46], [43]. A number of scholars depended on the translation from one language to another to construct their corpus, for example for the Opinion Corpus for Arabic (OCA). It is one of the oldest corpora for ASA by [379], comprising more than 500 Arabic movie reviews. The reviews were translated using automatic machine translation, and the results compared to both Arabic and English versions. Subsequently, most research efforts have focused on enhancing classification accuracy with the OCA dataset [380]. In addition, the MADAR corpus was proposed by [381]. It included 12,000 sentences from Basic Traveling Expression Corpus (BTEC) [382] translated to French, MSA, and 25 Arabic dialects. This corpus for Dialect Identification and Machine Translation is

available online[5]. One of the earliest Arabic datasets created as MSA Resource was the Penn Arabic Treebank (PATB) [287]. It consisted of 350,000 words of newswire text. It had 12 parts. This dataset has been a main resource for some state-of-the-art systems and tools such as MADA [383], and it's successor MADAMIRA [270], YAMAMA [384], and the tool by [385]. It is available for a fee[6].

Regarding the Arabic dialects, the Egyptian dialect had a lot of attention; one of the earliest Egyptian corpora is the CALLHOME corpus [47]. In addition, Levantine Arabic was well-studied, leading to the Levantine Arabic Treebank (LATB) [48]. It includes 27,000 words in Jordanian Arabic. Some efforts were made for Tunisian [34], [386], and Algerian [387]. Regarding the Gulf Arabic corpus, there is the Gumar corpus [388]. It consisted of 1,200 documents written in Gulf Arabic dialects from different forum novels. It is available online[7]. Using the Gumar corpus, a Morphological Corpus of Emirati dialect has been created. Khalifa et al. [389] consisted of 200,000 Emirati Arabic dialect words and is freely available[8]. More details about the Arabic corpora are summarized in Table 2.9. However, there are shortcomings to the existing corpora and their availability. This is due in part to the strict procedures for gaining permission to reuse aggregated data, with most existing corpora not offering free access. Furthermore, it is clear from Table 2.10 that the most frequently applied source for Saudi corpora is Twitter. Unfortunately, all Saudi corpora that were found in the literature are not available. In addition, some of them did not mention details about the annotation, which may cause a limitation for using these corpora. Finally, Figure 2.8 illustrates the percentage of different Arabic corpus types. Interestingly, since 2017, I found that dialectal Arabic has been used in more corpora than MSA.

[5] http://nlp.qatar.cmu.edu/madar/

[6] https://catalog.ldc.upenn.edu/LDC2005T20

[7] https://nyuad.nyu.edu/en/research/centers-labs-and-projects/computational-approaches-to-modeling-language-lab/resources.html

[8] https://nyuad.nyu.edu/en/research/centers-labs-and-projects/computational-approaches-to-modeling-language-lab/resources.html

Corpus Name	Ref.	Source	Size	Type	Online Availability
Twitter Benchmark Data Set for Arabic Sentiment Analysis	[43]	Twitter	151,000 sentences classified as positive or negative	MSA/ Egyptian dialect	Not Available
SUAR (Saudi corpus for NLP Applications and Resources)	[278]	Different social media sources such as Twitter, YouTube, Instagram and WhatsApp	104,079 words	Saudi dialect	Not Available
Health data set	[217]	Twitter	2,026 tweets classified as positive or negative	Arabic dialect	Not Available
DARDASHA	[199]	Chat Maktoob (Egyptian website)	2,798	Arabic dialect	Not Available
TAGREED		Twitter	3,015	MSA/Dialect	
TAHRIR		Wikipedia Talk pages	3,008	MSA	
MONTADA		Forums	3,097	MSA/Dialect	
Hotel Reviews (HTL)	[291]	TripAdvisor.com	15,572	MSA/Dialect	Not Available
Restaurant Reviews (RES)		Restaurant Reviews (RES) from Qaym.com	10,970	MSA/Dialect	
Movie Reviews (MOV)		Movie Reviews (MOV) from Elcinemas.com	1,524	MSA/Dialect	
Product Reviews (PROD)		Product Reviews (PROD) from Souq.com	4,272	MSA/Dialect	
MIKA	[390]	Twitter and different forum websites for TV shows, product and hotel reservation.	4,000 topics classified as positive, negative or neutral	MSA and Egyptian dialect	Not Available

Arabic Sentiment Tweets Dataset (ASTD)	[273]	Twitter	10,000 Egyptian dialect tweets	Egyptian dialect	Freely available at https://github.com/mahmoudnabil/ASTD
Arabic Twitter Corpus	[391]	Twitter	8,868 tweets classified as positive, negative, neutral or mixed	Arabic dialect	Available via the ELRA repository.
Large Arabic Book Review Corpus (LABR)	[230]	Book reviews from GoodReads.com	63,257 book reviews	MSA/Dialect	Freely available at www.mohamedaly.info/datasets
Al-Hayat Corpus	[392]	Al-Hayat newspaper articles	42,591	MSA	Available for a fee http://catalogue.elra.info/en-us/repository/browse/ELRA-W0030/
An-Nahar Corpus	[393]	Newspaper text		MSA	Available for a fee https://catalog.elra.info/en-us/repository/browse/ELRA-W0027/
AWATIF (a multi-genre corpus of Modern Standard Arabic)	[394]	Wikipedia Talk Pages (WTP), The Web forum (WF) and Part 1 V 3.0 (ATB1V3) of the Penn Arabic Treebank (PATB)	2855 sentences from PATB, 5,342 sentences from WTP and 2,532 sentences from WF	MSA/Dialect	Not Available
The Arabic Opinion Holder Corpus	[204]	News articles	1 MB news documents	MSA	Available at http://altec-center.org/
Arabic Lexicon for Business Reviews	[259]	Reviews	2,000 URLs	MSA	Not Available
Tunisian Arabic Railway Interaction Corpus (TARIC)	[34]	Dialogues in the Tunisian Railway Transport Network	4,662	Tunisian dialect	Not Available

Table 2.9: Comparison between different Arabic corpora.

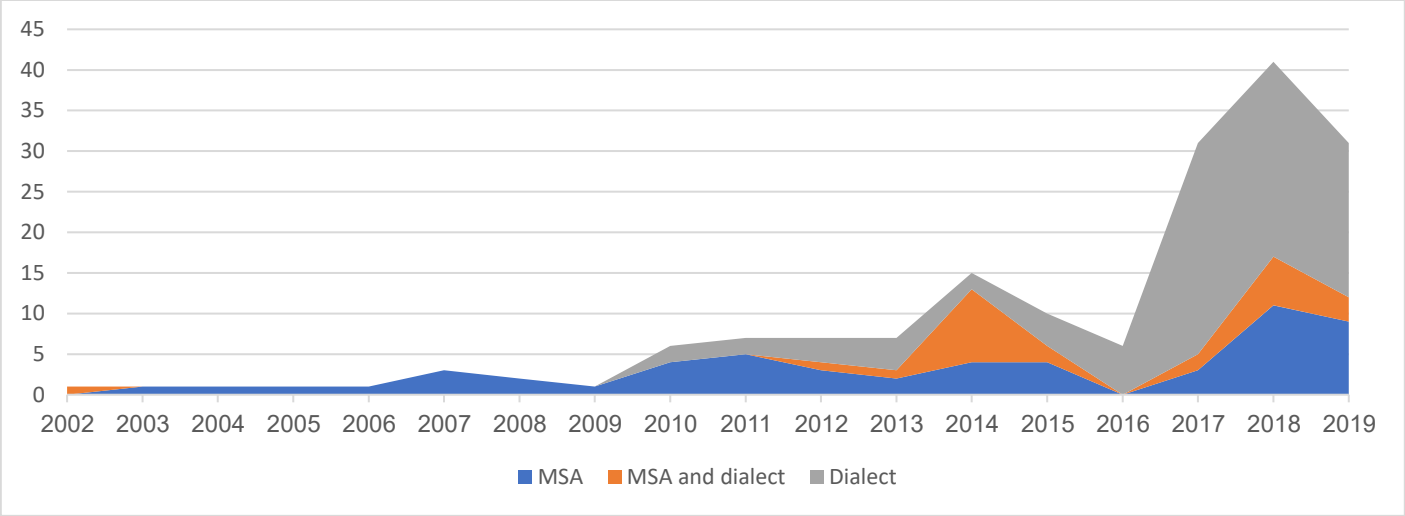

Figure 2.8: Percentage of Arabic corpora over time based on the type of corpora.

Corpus Name	Ref.	Source	Size	Classification	Online Availability
AraSenti-Tweet Corpus of Arabic SA	[39]	Twitter	17,573 tweets	Positive, negative, neutral, or mixed labels.	Not Available
Saudi Dialects Twitter Corpus (SDTC)	[279]	Twitter	5,400 tweets	Positive, negative, neutral, objective, spam, or not sure.	Not Available
Sentiment corpus for Saudi dialect	[395]	Twitter	4000 tweets	Positive or negative.	Not Available
Corpus for SA	[396]	Twitter	4700 tweets		Not Available
Saudi public opinion	[244]	Two Saudi newspapers	815comments	Strongly positive, positive, negative, or strongly negative	Available upon request
Saudi corpus	[397]	Twitter	5,500 tweets	Positive, negative, or neutral	Not Available
Saudi corpus	[398]	Twitter	1,331 tweets	Positive, negative, or neutral	Not Available

Table 2.10: Comparison between different Saudi dialect corpora for ASA.

2.2.6.5.6 Systems and Tools

Many systems and tools that support Arabic are for Morphological Analysis (MA) [399], [383], [270]. The oldest and pioneering system in this field is the Buckwalter Arabic Morphological Analyzer (BAMA) [399]. It depended on an Arabic dictionary. This dictionary included prefixes, stems and suffixes. Pasha et al. [270] proposed MADAMIRA based on two systems: MADA [383] and AMIRA [400]. A lot of Arabic works were based on BAMA, for example, SAMA 3.1 [401] and MADA + TOKAN [383].

It is worth emphasising other important works, such as AMIRA [400], an Arabic online tool for POS-tagging, tokenization, and lemmatization. Another Arabic system is Khoja's Stemmer [402]. It eliminates the prefixes and suffixes from the word and takes out the root.

Named Entity Recognition (NER) tools are considered important for extracting semantic features of the text [403]. However, works applying (NER) to Arabic are few [404]. One of the very recent works on NER is by AL-Jumaili and Tayyeh [233] who proposed a real-time named entity recognition system using news from Internet. The F-score for person, location, organization, noun, and verb was 72.61%, 68.69%, 55.25%, 77.62%, and 65.96%, respectively.

The review of the ASA literature confirmed the effectiveness of techniques (e.g., data mining) for analysing abundant data (i.e., Arabic text) and for projecting patterns for further discussion and analysis (e.g., forecasting). Our review revealed that, although there is an increasing interest in the use of ASA tools, unfortunately, there is no clear recommendation on a reliable enough tool to perform this analysis within a real-world context. In addition, the tools that are widely used in the SA field don't support ASA, such as Tableau [405], and Power BI[9]. For this, tech giants like the Saudi Telecom Company are translating Arabic tweets into English and then using Tableau type software to perform SA.

One of the tools that are used for ASA, [406] compared between a created Opinions Polarity Identification (AOPI) tool and two free online SA tools supporting ASA, which are SentiStrength [298] and Social-Mention[10]. They applied them on a corpus including 3,015 opinions written in MSA and different Arabic dialects. The results proved the efficiency of AOPI over the other tools.

[9] Data Visualization | Microsoft Power BI
[10] *Real Time Search - Social Mention*

The reviewed studies have covered a number of techniques enabling opinion-oriented information-seeking systems. These highlight the intellectual richness and breadth of the research area. In addition, numerous studies have also proposed many data-mining systems for MSA. Some systems were designed for dialectal Arabic, while others were designed for both MSA and dialects. Table 2.11 illustrates and compares the features of these systems within the ASA literature.

The findings of our review indicate that the majority of existing systems for Arabic text used SVM, KNN and NB classifiers, which have proven to be effective with Arabic. However, future research in ASA is expected to adopt other techniques, such as deep learning and neural networks, which has showed already early promise.

	System	Pre-processing	Algorithms	Data Source	Evaluation
ASA tools for Modern Standard Arabic (MSA)	Standard Arabic sentiment analyzer (SentiArabic) [407]	Yes	Lexicon-based combined with a decision tree	SentiTest contained online news and a PATB sentiment annotated by [203].	F-score of 76.5% on a blind test set.
	[408]	Yes	Unsupervised technique	Restaurant reviews	60.5% accuracy
	Aara [244]	Yes	NB classifier	Newspaper comments (Alriyadh and Aljazirah)	F-score was 84.5%. The accuracy of the system is 82%
ASA tools for Dialectal Arabic (DA).	[409]	Yes	DT, SVM, and NB	Users' comments in Facebook	73.4% accuracy
	[410]	Yes	DT, SVM, and NB	28,300 reviews from YouTube www.youtube.com	94.5% accuracy
	Mazajak/[234]	Yes	CNN followed by an LST	SemEval ([411] ASTD [273], ArSAS [412]	92.0% accuracy

	[410]	Yes	DT, SVM, and Naïve Bayes	28300 reviews from YouTube www.youtube.com	94.5% accuracy
ASA tools for both MSA and DA Arabic	Colloquial non-standard Arabic-Modern Standard Arabic Sentiment Analysis (CNSA-MSA-SAT)/ [413]	Yes	IBK (KNN) Classifier	Arabic reviews and comments from online social website	The accuracy was 90%
	SAMAR/ [199]	Yes	SVM light	Chat websites, social media, web forum and Wikipedia talk pages	The highest accuracy for sentiment classification was for web forum at 71.82%

Table 2.11: Data mining tools used in ASA.

2.3 Summary

This chapter reviewed all the studies related to the variables of customer satisfaction, customer churn and Twitter. As a result, this review has identified the current churn prediction model issues, the customer churn variables and the measurable metrics for customer satisfaction. In addition, this chapter reviewed research on ASA to provide a holistic view of the approaches, tools and resources used in this field. This review aims to identify the different approaches used for ASA: machine learning, lexicon-based and hybrid approaches. This chapter offers insight into the issues and challenges associated with ASA research, and it provides suggestions for ways to move the field forward. For example, even now, there are many gaps and deficiencies in the studies on ASA. Specifically, Arabic tweets, corpora and data sets for SA are currently only moderately sized.

Moreover, Arabic lexicons with high coverage contain only MSA words, and those with Arabic dialects are quite small. New corpora need to be created. Additionally, there is a need to develop ASA tools that can be used in industry and academia for Arabic text SA.

Chapter 3: Sentiment Resources for Saudi Dialect

3.1 Introduction

With the growing use of social media sites worldwide over the last ten years, Sentiment Analysis has recently become a prominent and useful technique for capturing public opinion in many different disciplines. However, sentiment polarity detection is a challenging task due to the limitations of sentiment resources in different languages. Whilst a substantial body of research exists for English [44], [45] it remains a largely unexplored research area for the Arabic language [44], [45], [39], [46]. This is due chiefly to the complexity of Arabic [46], [45], [278]. Hence, Dialectal Arabic (DA) analysis, targeted here, is complicated, requiring a native speaker. Moreover, DA datasets and lexicons, especially freely available Gold Standard Corpora (GSC), Saudi dataset are lacking [44], most resources being for Egyptian and Levantine [45], and current effort has concentrated on and the Gulf Dialect [388] and the Palestinian Dialect [414].

Nevertheless, there is still a need for DA for Arabic corpora [415]; for Saudi Dialectical Arabic, this need is stringent [45]. Therefore, I attempt to alleviate this matter by focusing on Arabic Sentiment Analysis and provide solutions to one of the challenges that face Arabic SA by creating a *Saudi GSC and Saudi Sentiment lexicon and AraSTw lexicon*. These resources are based on data extracted from Twitter. This chapter presents how I have constructed, cleaned, pre-processed, and annotated the 20,0000 Gold Standard Corpus (GSC) AraCust, the first Telecom GSC for ASA for Dialectal Arabic (DA). AraCust contains Saudi dialect tweets, processed from a self-collected Arabic tweets dataset and annotated for sentiment analysis, i.e., manually labelled (k=0.60). In addition, I have illustrated AraCust's power, by performing an exploratory data analysis to analyse the features that were sourced from the nature of the AraCust corpus, to assist with choosing the suitable ASA methods for it. The AraCust corpus released[11] for the research community.

3.2 Data Collection

[11] AraCust: a Saudi Telecom Tweets corpus for sentiment analysis [PeerJ]

To build the dataset, I used Python to interact with Twitter's search application programming interface (API)[12] to fetch Arabic tweets based on certain search keys. The Python language and its libraries are one of the most flexible and popular approaches used in data analytics, especially for machine learning. To ensure pertinence to our target application, I started with hashtags related to the three largest Saudi telecom companies: the Saudi Telecom Company (STC), the Etihad Etisalat Company (Mobily), and Zain KSA, which dominate the market. As a result, I extracted the relevant top hashtags, as follows: #STC, #Mobily, #Zain, #موبايلي, #الاتصالات_السعوديه and #زين_السعودية which were used for the search. These initial seed terms were extracted based on the following Python function:

tags = API.trends. place ()

From the tweepy library. Additionally, I used the Twitter accounts of these companies as search keywords. As the aim of this collection was to allow for a longitudinal, continuous study of telecom customers' sentiments, I gathered data continuously from January to June 2017, mainly because this period includes customers' reactions to the Saudi Communications and Information Technology Commission's new index, which refers to complaints submitted to the authorities [416]. While seemingly a short period, it in fact generated the largest Arabic Telecom Twitter dataset for ASA. I was aware that I needed to account for the dataset subsequently reducing in size after spam and retweets were eliminated. The initial result obtained comprised 3.5 million tweets. After filtering and cleaning (based on location and time-zone and stratified random sampling), the dataset was reduced to 795,500 Saudi tweets, which comprise the large AraCust dataset.

For our own further experimentations, in order to reduce computational costs and time in constructing our working AraCust corpus, I chose a sub-sample of Saudi tweets randomly from the dataset to prevent bias [417]. The principal notion behind the size reduction of the corpus was that the annotation process is manual, time-consuming, and costly. Specifically, to avoid bias in the sample, I applied the following steps: identify the population, specify the sample frame, and choose the right sample technique. As stated, the population in this study is STC, Mobily and Zain customer tweets. The sample frame is a Saudi tweet that describes the

[12] Twitter Inc. Twitter Inc. obtained Gnip, and it became the official tweets provider.

tweet author's point of view regarding one of these companies. The probability sample technique is Simple Random Sample (SRS), applied stratified over the three sets (STC, Mobily, and Zain). The advantage of SRS is that all of the population has the same chance of being selected [418]. In addition, scholars have proven the efficiency of the random sampling technique for social media, because items that are repeated multiple times in the data set are likely to appear frequently in the sample as well [419], [420].

The sample size decision was based on a pattern-extraction experiment using Network Overview, Discovery, and Exploration Node XL [421]. Node XL is an add-in tool for Microsoft Excel used in social media analysis and visualization. Up to 2000 Arabic tweets were retrieved using the previously mentioned hashtags. Based on the findings of another study that 110 tweets per day are enough to capture customer sentiment [44], I needed 20,000 tweets over 6 months. In addition, I found that the services provided by Saudi telecommunication companies most frequently mentioned in the customers' tweets were: Internet speed, signal coverage, after-sales service, call centres, and fibre communication.

The size of our AraCust corpus of 20,000 Saudi tweets (Table 3.1) is in line with that of previous studies, which showed that datasets over 20,000 tweets are sufficient to produce state-of-the-art systems for Twitter Sentiment Analysis (SA) [166], [161].

As the companies I targeted were from Saudi Arabia, I further filtered the tweets based on user location and time zone to identify Saudi tweets. Saudi Arabia ranks seventh in the world in the number of personal accounts on social media [49]. I found that many tweets do not have a location field set in the profile of the users who posted them. To resolve this issue, I used a list of city names, landmark names, city nicknames, etc., for Saudi Arabia, as additional labels for the user location of tweets, following [36]. Also following Mubarak and Darwish, I used a list from the GeoNames website,[13] a geographical database that includes 8 million place names for each country, which includes 25,253 place names for Saudi Arabia.

Finally, in the context of our data collection process from Twitter, it is worth mentioning that ethical concerns of using social media data have stirred an ongoing controversy in research communities in terms of confidentiality and privacy. The availability of social media data is thought to potentially expose social media

[13] https://www.geonames.org/

users to risks. Although social media data is prominently public still, the emergence of profiling by business owners for business purposes has led to criticism and apprehension. Regarding our own study, on Twitter, users' phone numbers and addresses are not made public, to provide some level of privacy. Additionally, in our current research, I further deleted any phone numbers or names that were included in the tweets themselves, for additional privacy. Finally, I collected only the tweet texts, time, and location, without collecting any other user-related information from them.

Company	Twitter Handle and hashtags	# of Unique Tweets
STC	@STC_KSA, @STCcare, @STCLive	7,590
Mobily	@Mobily, @Mobily1100, @MobilyBusiness	6460
Zain	@ZainKSA, @ZainHelpSA	5950
Total		20,000

Table 3.1: Companies and the total number of unique tweets from each in AraCust.

3.3 Corpus Cleaning and Pre-Processing

To avoid noise in the corpus, cleaning was performed on the dataset. One way of cleaning is removing spam, thus any tweet with a Uniform Resource Locator (URL) was excluded, as in [45] and [217], because most tweets in the dataset with a URL were news or spam. In addition, I excluded repetitive information, such as retweets, as recommended by [422] and [217]. Moreover, non-Arabic tweets were excluded from the data set by filtering for Arabic language (lang: AR), because translation damages the classifier efficiency. Pre-processing was completed on the corpus using a Python script to remove unnecessary features in the tweets that might lower accuracy from the tweet corpus before applying classifiers, such as user mentions (@user), numbers, characters (such as + = ~ $) and stop words (such as ",", ".", ";"), as suggested by [391] and [45]. The tweet corpus was processed using the Natural Language Toolkit (NLTK) library in Python for normalization and tokenization. Although emoticons could arguably express sentiment, they were deleted, because prior research reported a classifier misunderstanding between the parentheses in the quote and in the emoticon [45]. In addition, importantly, as I dealt with Arabic tweets, [391] showed that retaining emoticons

in classification decreased the performance of the classifier; they stated that this was due to the way Arabic sentences are written from right-to-left, which is reversed in emoticons.

Next, the words in the tweets were tokenized, which means that sentences were segmented into words for easier analysis, as in [45] and [423]. Finally, the tweets were normalized. For Arab text, normalization entails the unification of certain types of Arabic letters of different shapes, as in [39], i.e.:

- Replacing the Arabic letters "أ", "إ", and "آ" with bare *alif* "ا".
- Replacing the letter "ى", "ئ" and "ىء" with bare *ya* "ي".
- Replacing the final "ة" with "ه".
- If a word starts with "ء", replacing it with "ا".
- Replacing "ؤ" with "و".

As stemming algorithms do not perform well with Dialectal Arabic (DA) words [158], they were not applied. The data collection, filtering, cleaning, and pre-processing steps are illustrated in Figure 3.1. The subset before and after the pre-processing is illustrated in Table 3.2.

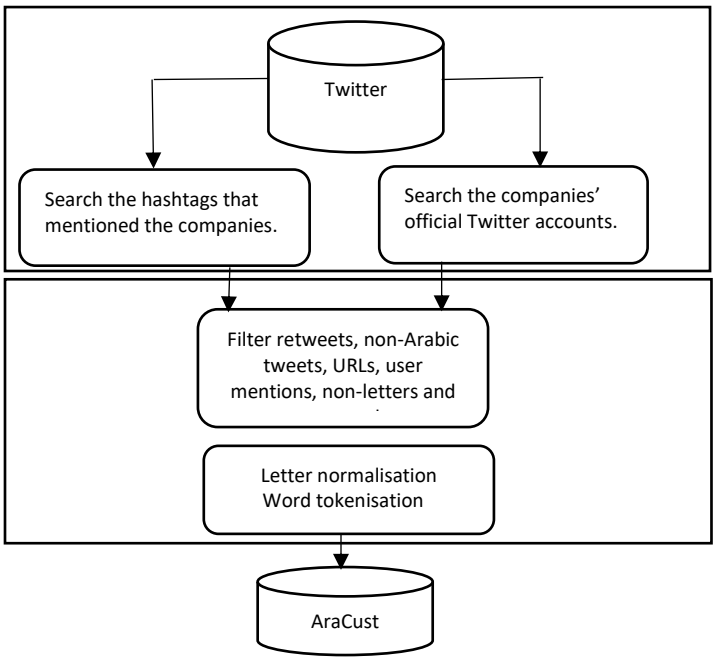

Figure 3.1: AraCust corpus collection, filtering, and pre-processing.

Tweet in Arabic	Label	Company	Tweet in English	Tweet After pre-processing
@So2019So @STCcare غيري الشركه	Negative	STC	Change the Company	غيري شركه
@alrakoo @mmshibani @GOclub @Mobily اشكرك 😊	Positive	Mobily	Thank you	اشكرك

Table 3.2: Subset of the corpus before and after pre-processing.

3.4 Exploratory Data Analysis

Before doing the sentiment analysis task, it is important to analyse the corpus. This includes the data types that I will deal with in the classification and prediction experiments, as well as the features that originate from the nature of the corpus, which may affect the model's performance. Our data analysis involved many features set analyses, from character-based to dictionary-based, and syntactic features [424]. This exploratory data analysis was accomplished using character-based, sentence-based, and word-based features, to allow for processing at a variety of levels. The exploratory data analysis was completed using the NLTK library via a Python script.

From the exploratory data analysis, I observed first that there were more negative tweets than positive tweets for all three companies (see Table 3.3 and Figure 3.2). I interpret this result as being due to all Arab countries having suffered difficult economic circumstances in the past few years; this result is in line with the findings by [425] and [268].

Company	Negative	Positive	Total
STC	5,065	2,525	7590
Mobily	4530	1930	6460
Zain	3972	1978	5,950
Total	13,567	6433	20000

Table 3.3: Companies and the total number of positive and negative tweets.

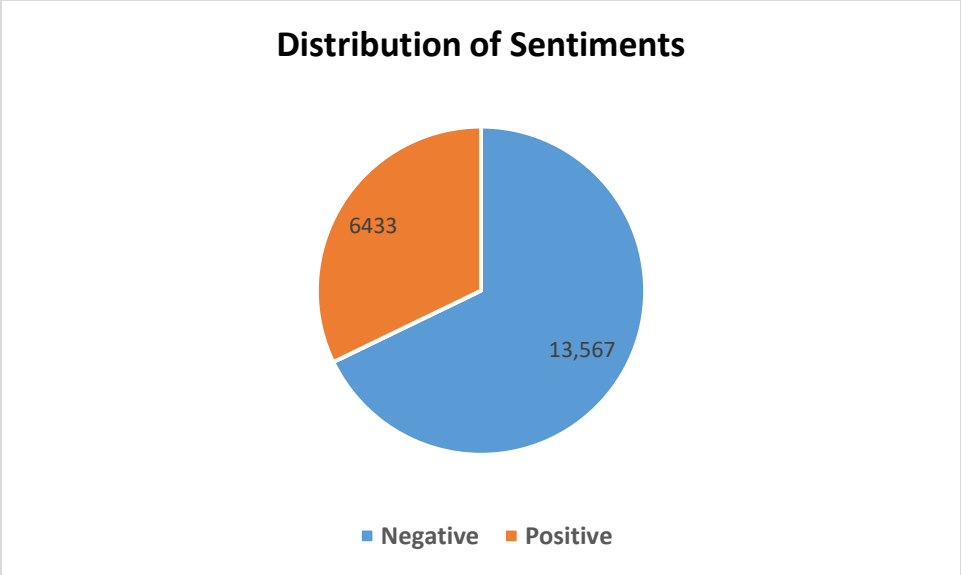

Figure 3.2: Distribution of Negative and Positive Sentiment.

Next, I analysed the differences between the tweet length distribution across the sentiment to determine whether there was some potential correlation there and because prior research used the tweet-length feature as input to a machine learning classifier in SA research [185], [198] (Figure 3.3). I observed that tweets tend to be longer when customers express a negative sentiment. In addition, interestingly, I found that STC customers had longer tweets overall than other companies' customers (Figure 3.4). These results guided us to use the All-Tweet Length feature in the classification task to estimate the impact of tweet length on the classifier's performance.

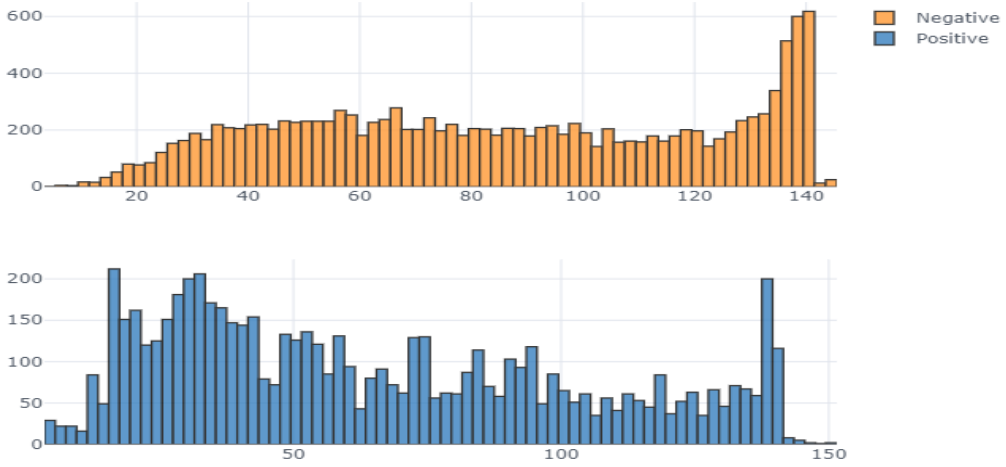

Figure 3.3: Tweet length distribution across sentiment.

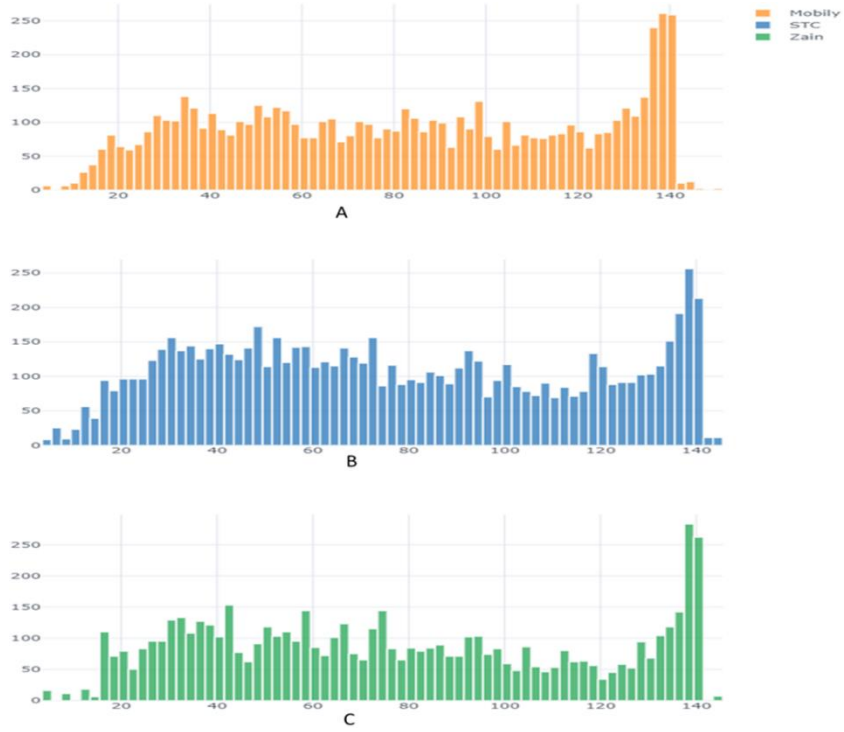

Figure 3.4: Tweet length distribution across companies.

The ten most frequent words in the corpus and their number of appearances in the corpus are given in Table 3.4. It appears from the table that there is a repeated use of the word “God,” but just from this information I do not know whether it was repeated in a negative or positive way. In addition, there was just one positive

expression among these frequent words: “thank you” (which is one word in Arabic, Table 3.4). The highest frequency was, naturally, for the word “Internet,” which potentially indicates the importance of this service; but likewise, I cannot tell at this stage if the reason for having “internet” among the most frequent words is positive or negative. To better understand the way these words are used, I first studied the context of usage by using the “most frequent” bigram to provide a more comprehensive view of the data.

Word in Arabic	Frequency	Word in English
نت	1770	Internet
الله	1760	God
سلام	1363	Hello
والله	1179	Swear God
خاص	1315	Private
حسبي	637	Pray
عملاء	599	Customers
شكرا	560	Thank you
مشكلة	549	Problem
شريحة	515	Sim card

Table 3.4: Most Frequent Words in the AraCust corpus.

The most frequent bigram on the corpus, as shown in Figure 3.5, is “pray” (note that this is expressed as two words in Arabic); this is mainly used in a negative way, as explained below. Greetings are next in frequency, followed by “data sim card,” which I thought may to be due to a frequent problem source. I observed that internet service is described as slow, so most of the tweets that mentioned the internet are complaints, as shown below. Additionally, “customer service” is one of the most frequent bigrams in the corpus.

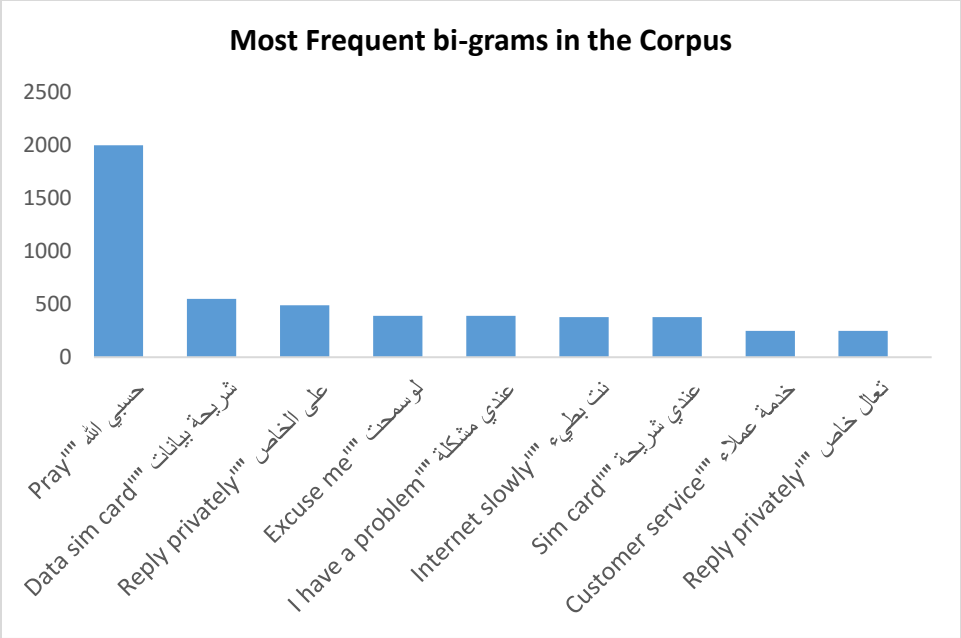

Figure 3.5: Most Frequent Bigrams in the AraCust corpus.

Next, I calculated the positive and negative rate for each word in the most frequent word chart to determine whether the word was used with a positive or negative sentiment. I calculated the positive rate pr(t) and negative rate nr(t) for the most frequent words (term t) in the corpus as follows (Table 3.5):

$$pr(t) = \frac{term\_freq\_df[t,'positive']}{term\_freq\_df[t]} \quad (3.1)$$

$$nr(t) = \frac{term\_freq\_df[t,'negative']}{term\_freq\_df[t]} \quad (3.2)$$

Where term_freq_df [t, val]; val ; is $\in\{positive,negative\}ethicalc$ the frequency of the word *t* as a word with valence (sentiment) *val* in the corpus:

$$term\_freq\_df[t,val] = \sum_{tw\in C}^{n} \text{bool1}(\text{tw},\text{t},\text{val}) \quad (3.3)$$

Where *tw* is a tweet in corpus C; and bool1() is a Boolean function:

$$bool1(tw,t,val) = \begin{cases} 1, & valence(tw,t) = val \\ 0, & rest \end{cases} \quad (3.4)$$

With valence(tw,t) a function returning the sentiment of a word *t* in a tweet *tw* and term_freq_df [t] is the total frequency of the word *t* as both a positive and negative word in the corpus:

$$term\_freq\_df[t] = \sum_{tw\in C}^{n} \text{bool2}(\text{tw},\text{t}) \quad (3.5)$$

Where bool2() is a Boolean function:

$$bool2(tw, t) = \begin{cases} 1, & t \in tw \\ 0, & t \notin tw \end{cases} \qquad (3.6)$$

I found that "internet" is used as a negative word more than a positive word, as mentioned before. In addition, possibly surprisingly, the word "God" is used in negative tweets more than in positive ones. The words "Hello", "Swear to God", "private", "sim card" and "thank you" are used as positive words more than as negative words. This solved my initial misconception about 'sim card' being a problem. Moreover, I found the word 'customers' used as a negative word more than a positive word.

These results led us to use the *Has Prayer feature* in the classification task; this feature allows us to evaluate whether the existence of a prayer in a tweet increases the classifier's performance.

Term in Arabic	Term in English	Negative	Positive	Total	Pos_rate	Neg_rate
نت	Internet	975	795	1770	0.44	0.55
الله	God	977	783	1760	0.44	0.55
سلام	Hello	765	895	1363	0.65	0.56
والله	Swear God	567	704	1179	0.59	0.48
خاص	Private	656	659	1315	0.50	0.49
حسبي	Pray	425	212	637	0.33	0.66
عملاء	Customers	413	186	599	0.31	0.68
شريحه	Sim card	271	289	560	0.51	0.48
مشكله	Problem	279	270	549	0.49	0.50
شكرا	Thank you	235	280	515	0.54	0.45

Table 3.5: Most Frequent Words in the AraCust corpus and their sentiment probability.

The feature set analysis is illustrated in Tables 3.6, 3.7and 3.8. The character-based features (Table 3.6) reflect the existence of symbols, such as a minus sign, punctuation marks such as a comma, and numbers. The ratio was measured between the number of characters in a tweet and the number of characters overall.

Word-based features (Table 3.7) include word standard deviation, which was calculated using the standard deviation of word length, word range (the difference between the longest and shortest word), characters per words calculated by the mean number of characters for each word, and vocabulary richness, which is the count of various words.

Sentence-based features include the mean number of words for each sentence, the standard deviation of sentence length, and range, the latter expressing the difference between the longest and shortest sentence (Table 3.8).

Character-based Feature	Ratio
Punctuation marks	8.0
Numbers	6.03
Symbol	1.0

Table 3.6: Character-based Features.

Word-based Feature	Ratio
Word standard deviation	6.51
Word range	30
Chars per word	5.22
Vocabulary richness	1.0
Stop words	0.0
Proper nouns	0.11

Table 3.7: Word-based Features.

Sentence-based Feature	Ratio
Words per sentence	16.23
Sentence standard deviation	7.17
Range	30

Table 3.8: Sentence-based Features.

3.5 Annotation

Before the SA, I needed to train the classifier and create a readable version for the machine using corpus annotation. Annotation is the process of assigning interpretative information to a document collection for mining use [65]. Hinze et al., [426] defined the annotation as using a predefine classes to mark the text, sentence, or words. Salmeh et al. [268] defined annotation as providing the opinions and sentiments towards the target. There are different levels of corpus annotation. For example, sentiment annotation and syntactic

annotation is the process of parsing every sentence in the corpus and labelling it with its structure, grammar and part-of-speech (POS) – that is, labelling every word in the corpus with a corresponding appropriate POS label.

Several approaches used to annotate the corpus, including the manual approach, which depends on human labour, and the automatic approach, which uses an annotation tool.

Gold Standard Corpora (GSC) are an important requirement for the development of machine learning classifiers for natural language processing with efficiency; however, they are costly and time consuming and this is the reason for the rare existence of GSC, especially in Arabic [427].

The process of construction of the GSC is based on manual annotation by different experts who review the data individually, then the inter-annotator agreement is computed to confirm the quality [427].

For sentiment annotation, several studies used three-way classification labels (positive, negative and neutral) to express the sentiment orientation [428], [391], [171], [45]. The output from the classification is based on the labels used in the annotation.

In this research, I classified the corpora using five-way classification and binary classification for two different experiments. I classified the text using a five-way sentiment classification (Strongly Positive, Positive, Neutral, Negative, Strongly Negative), which is consistent with the SemEval 2017 Task 4 for Arabic tweets [170]. In addition, I followed some of the studies that used SA to predict customer satisfaction or consumer sentiment, which rated the strength of the sentiment [69], [72]. This is compatible with user behaviours in rating sentiments towards a product or restaurant in the business world [429]. Each sentiment label is considered as a degree of customer satisfaction towards specific telecom services. The annotator should assign one label to represent the strongest emotion expressed per tweet, as noted in many studies [45], [171], [391], [170].

Mixed class refers to cases in which the text expresses both positive and negative sentiments simultaneously, making it difficult to decide which is the strongest sentiment expressed. In some SA studies, the mixed class is ignored, based on the assumption that it is uncommon to find texts conveying more than one sentiment [252]. Meanwhile, other studies considered a text with mixed sentiments as neutral sentiment [247].

Therefore, to be as comprehensive as possible, I decided to follow the latter and considered the mixed class to be a neutral label in the five-way classification. While in binary classification, the mixed class considered as Indeterminate.

In addition to the five-way classification mention above, I also, additionally, used a binary classification (Negative vs Positive) in this research, to predict the customer satisfaction toward the telecom company, following many studies that used binary sentiment classification with Arabic text [430], [171], [45], [199]. Several prior studies proved that binary classification is more accurate than other classifications [171], [45]. Each sentiment label is considered as a degree of customer satisfaction: satisfied and unsatisfied.

Sarcasm is defined as a form of speech in which a person says something positive while he/she really means something negative or vice versa [181]. Sarcasm is notoriously hard to detect; in English, there are only a few studies on sarcasm detection using supervised and semi-supervised learning approaches [181]. In ASA, no study was found that takes on sarcasm detection. Therefore, I asked the annotators, optionally, to also label tweets with sarcasm - according to the sentiment they conveyed. This allowed us to be able to use sarcasm as a feature for machine learning classifier, following [171]. I thus opened the way for the first sarcasm-detection Arabic NLP work.

The corpora were divided into three corpora, based on the telecom company as the keyword (STC, Mobily, Zain). The manual approach was adopted. To ensure the high quality of the manual annotation process, the annotation process needs clear guidelines to maintain consistency between annotators [45].

As recommended by [217], [45], three annotators were hired in this research to annotate the corpus. The annotators called A1, A2, and A3 were all Computer Science graduates, native speakers of the Saudi dialect and had experience in the annotation process. The reason for choosing three annotators instead of the usual, and simpler, two, was to increase the quality of the resulting corpus by alleviating conflicts that could arise from discrepancies between only two annotators. Hence, if two annotators disagreed with respect to one tweet classification, I took a vote between all three annotators. In addition, [431] stated that more than two annotators are more preferably.

To encourage a thorough examination of the tweets and high-quality results, the annotators were paid. Moreover, to ensure fair pay, in order to determine the annotators' wages, I conducted a pilot study to calculate the average time they needed to annotate the tweets, as recommended by [45]. I provide the annotators with 110 tweets [44] and the annotation guideline, and then calculated the average time that they needed for annotation. They took 33 minutes, 20 minutes, and 35 minutes to annotate 110 tweets. Thus, the average time that they needed was 30 minutes to annotate 110 tweets. I then paid them to annotate the 20,000 tweets over the course of 2.5 months, two hours per day for five workdays per week.

Before I began the annotation process, the annotators were provided with annotation guidelines in both Arabic and English in one-hour session; some of the annotation guidelines are shown in Table 3.9. I stored the annotations in an Excel file. The annotation guidelines were also included in the Excel file in case the annotator needed to read it (Figure 3.6) and the full guideline in Appendix A.

As suggested by [431], I build an easy interface in the Excel file which has the tweets, an automatic list box of labels to avoid typing errors, , the sentiment-bearing words, and the telecom services mentioned in the tweet, if found (Figure 3.7).

To build a gold standard Arabic corpus, three rotations were used to annotate the corpus. As mentioned before, I divided the corpora into three corpora based on the Telecom companies, STC, Mobily and Zain. They started the first rotation by annotating the STC corpus, then the Mobily corpus, followed by the Zain corpus. After the first rotation, I reviewed the annotators' choices and discussed with them before the new rotation started. After the second rotation, I calculated the similarity percentage between A1, and A2, A2 and A3 and A1 and A3 for three corpora. At the third rotation, I asked the annotators to revise the labels for the corpus that have low similarity percentage. After the three rotations, the author revised the three annotation labels done by the annotators and compared their choices, using vote to make decisions. I found that 83% of the tweets were labelled with the same label by the A1 and A3, 75% of the tweets were classified with the same labels by A2 and A3, while 74% of the tweets were classified by A1 and A2 with the same labels.

The aim of this study is to predict customer satisfaction with telecommunication company and telecommunication services by analysing customer tweets on Twitter according to the Table shown below.	**هذه الدراسة تهدف الى قياس رضاء المستخدمين اتجاه الخدمات المقدمة من شركات الاتصال عن طريق تحليل أراء العملاء في تويتر وتصنيفها حسب الجدول الموضح بالاسفل.**
1. **Standpoint**: The Sentiment should be considered from the tweet author's point of view, not the annotator point of view.	1. المنظور: اختيار نوع الرأي ايجابي او سلبي يجب أن يكون كما أراد كاتب التغريدة التعبير عنه لا كما يراه الواسم. أي من منظور الكاتب وليس من منظور القاريء.
2. **Background**: The choosing of the sentiment label should be made according to the tweet content, not the annotator's background.	2. المحتوى: اختيار نوع الرأي يجب أن يكون كما يظهر في محتوى التغريدة وليس حسب معلومات سابقة للقاريء.
3. **Neutral**: A tweet that has mixed negative and positive sentiments and within which both polarity sentiments have the same strength, or, if the tweet does not include sentiment.	3. كلاهما: الرجاء اختيار (محايد) عندما تكون التغريده تحتوي مشاعر مختلطه ايجابيه وسلبيه وكلا المشاعر لها نفس القوه في التغريدة او كانت التغريدة بلا رأي.
4. If the service is unclear, please leave it empty.	4. عند عدم وضوح الخدمة التي يصفها المغرد تترك فارغه.
5. If the sentiment-bearing word is unclear, please leave it empty.	5. عند عدم وجود كلمة مؤثره ولكن التغريدة تدل دلالة إيجابيه أو سلبية تترك الكلمة المؤثره فارغه.
6. The **polarity** of a sentiment-bearing word is either positive or negative.	6: تصنيف الكلمة المؤثرة أما يكون أيجابي أو سلبي.

Table 3.9: Annotation Guidelines.

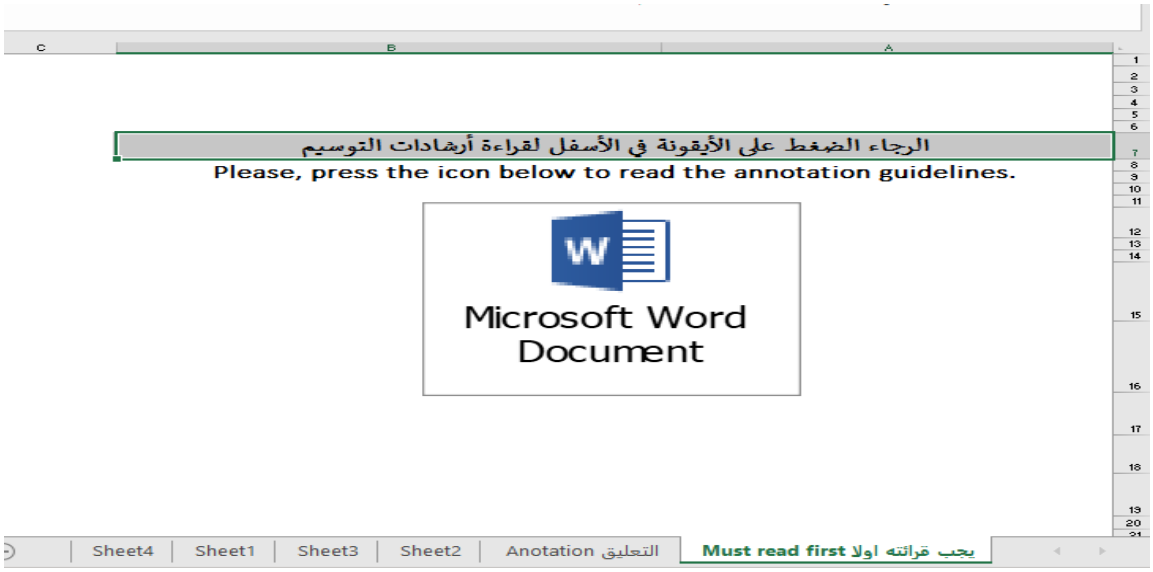

Figure 3.6: The included annotation guidelines in the Excel file.

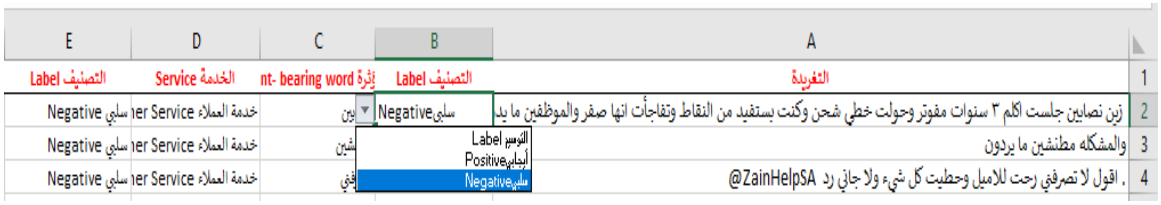

Figure 3.7: The annotation file.

3.6 Annotation Challenges

The annotators faced some challenges in the annotation process, similar to those experienced in prior research [432], such as:

- *Quoting and supplications*: It is difficult to define the sentiment of a tweet author whose tweet includes a quote or supplication, and to determine whether the author agrees with the sentiment of the quoted author. The annotators chose the sentiment that was expressed in the quote or in the

supplication. Then, I checked the sentiment that they allocated. I did not ignore or remove the tweets with quotes or supplications, because the quotes/supplications were a form of expression of author sentiment.

- *Sarcasm*: It is extremely hard to detect sarcasm in a tweet, because the explicit sentiment is different from the implicit sentiment. Nevertheless, as people are better at this than machines, annotation of tweets with this label is invaluable due to the difficulty of the sarcasm detection task [433]. For that, I asked them to label a tweet accordingly if they could detect sarcasm in it.

- *Defining the telecom services on the tweet:* The annotators indicated that not all of the tweets mentioned telecom services. This may be associated with the nature of the tweet, which is short. For this reason, I asked annotators to define the telecom services if they found them in the tweet.

- *Absence of diacritics*: this makes the pronunciation of a word difficult, because without diacritical marks, some words have two possible meanings. For these, I asked the annotators to interpret the word in the context of its sentence.

3.7 Inter-annotator Agreement

To identify the reliability of the annotation scheme, the inter-annotator agreement (IAA) was used. I used the similarity index as an early indicator of the annotators' agreement. Fleiss' Kappa [434] was used to measure the consistency for the 5-way classification (Strongly Positive, Positive, Neutral, Negative, Strongly Negative) and for the binary classification (Positive, Negative), because there were more than two annotators [434], [435].

The kappa **k Fleiss** [435] is defined as:

$$k = \frac{\overline{P} - \overline{Pe}}{1 - \overline{Pe}} \qquad (3.1)$$

Where $\overline{Pe}$ expresses the normalization of the agreement that is attainable randomly and $\overline{P}$ gives the normalized probability of agreement achieved by chance. If the annotators are in complete agreement, then k = 1. If there is no agreement among the annotators, then k < 0. The value I obtained was of 0.50 for 5-way

classification and 0.60 for binary classification for the three annotators, which is a moderate level based on the level of acceptance [1], Figure 3.8. In addition, I checked for agreement two-by-two between A1, A2 and A1, A3, and A2, A3 and I took the average A, Table 3.10.

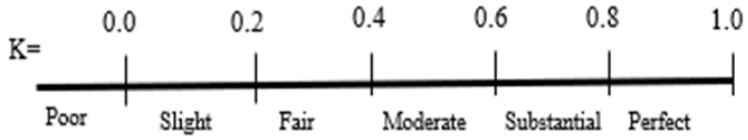

Figure 3.8: The acceptance level of k [1].

Annotators	k
A1& A2	0.7
A2 & A3	0.74
A1 & A3	0.87
Avg A	0.77

Table 3.10: Two-by-two agreement for binary classification between the three annotators.

3.8 Evaluation the corpus

To evaluate AraCust corpus, I applied a simple experiment using a supervised classifier to offer benchmark outcomes for forthcoming works. In addition, I applied the same supervised classifier on a publicly available Arabic dataset created from Twitter, ASTD [273], to compare the results of AraCuat and ASTD; the details of these datasets are provided in Table 3.11. I used an SVM, which has been used in Arabic sentiment analysis in recent research with high accuracy [235], [217], [236]. I used a binary classification (Positive, Negative) and eliminated tweets with different classification labels from the ASTD data set. I used a linear kernel with an SVM classifier, as some studies have stated that this is the best kernel for text classification [161], [39], [171]. The AraCust and ASTD corpora were split into a training set and test set; additionally, 10-fold cross-validation was performed for both to obtain the best error estimate [436]. The findings are in the test set, Table 3.12. For oversampling due to the dataset being biased towards negative tweets, I used the popular Synthetic Minority Over-Sampling Technique (SMOTE).

I analysed the features term presence, term frequency (TF) (the frequency of each term within the document) and term frequency–inverse document frequency (TF–IDF) (the frequency of each word based on all records' frequencies). I found that term presence is the best feature to use with binary classification, in line with what was found by [198], which is that term presence is best feature for binary classification due to a lack of term repetition within a short text, such as a tweet. In addition, [437] stated that a term presence model can provide information such as term frequency for short texts. Pang [247] noted that using term presence leads to better performance than using term frequency. The results in Table 3.12 show that our dataset AraCust outperforms the ASTD result. Further research may also investigate using deep learning algorithms on our newly created GSC AraCust dataset.

Data Set	Positive tweets	Negative tweets	Total
Aracust	6433	13,567	20000
ASTD	797	1,682	2,479

Table 3.11: Datasets used in the evaluation.

Data Set	Positive			Negative			Total	
Aracust	Precision	Recall	F1	Precision	Recall	F1	F1-avg	Accuracy
	93.0	76.0	83.6	91.0	98.0	94.4	89.0	91.0
ASTD	79.0	65.0	71.3	76.0	96.0	84.4	77.9	85.0

Table 3.12: Evaluation results of the SVM on the datasets.

3.8 Building the AraSTw Lexicon

Liu [183] mentioned two techniques to build a lexicon: *automatic* and *manual* techniques. Automatic techniques include two approaches: *dictionary-based* approach and *corpus-based* approach [183]. The manual approach is time- and labour-consuming but is more accurate than the automatic approach [45]. The dictionary-based approach depends on using a dictionary to find the synonyms and antonyms of seeds of positive and negative words, recursively, until no word is found anymore [45]. Oppositely, the corpus-based approach depends on the corpus to generate the polarity words, then using different approaches to find the synonyms and antonyms of these words to generate the lexicon [254].

I used the corpus-based approach to build my Arabic Sentiment Lexicon (AraSTw) from (AraCust) corpus, the golden annotated corpus created from Arabic tweets, due to the fact that all the data came from the AraCust corpus [254]. In the first phase, annotators who annotated the AraCust corpus were asked to extract the sentiment-bearing word from each tweet and classify the word as a positive or negative word. Then, I checked the words manually and gave a +1 score to positive words and -1 score to negative ones. In the second phase, I used two publicly available Arabic sentiment lexicons: AraSenTi [212], consisting of Saudi dialectical and MSA words and phrases and SauDiSenti [253], comprising also MSA and Saudi dialectical words and phrases. The third phase was filtering words and phrases that were positive or negative in the two lexicons, to keep only the Saudi dialectical words and phrases and eliminate the MSA ones. The proposed AraSTw lexicon statistics is shown in Table 3.13. Finally, I used the AraSTw lexicon with the classifier as a feature that will be further explained in Chapter 5 and used to predict the customer satisfaction.

Label	**Number of words**
Negative	28358
Positive	6397
Total	34,755

Table 3.13: AraSTw -lexicon statistics.

3.9 Evaluating the AraSTw Lexicon

To evaluate the AraSTw lexicon, I performed a simple lexicon-based approach, where I implemented an automatic count of negative and positive words, to define the sentiment of a tweet. I used one internal data set, which is the AraCust corpus, and one publicly available Arabic dataset that was created from Twitter, ASTD [273]; the details of these datasets are listed in Table 3-14. I have used a binary classification (positive, negative) and eliminated the neutral tweets from the data sets. Moreover, I compared the performance of AraSenTi [212] and my lexicon AraSTw on the same external dataset, Table 3.15 and 3.16. AraSenTi [212] created from Twitter, has Saudi dialectical words.

Data Set	Positive	Negative	Total
AraCust	6433	13,567	20000
ASTD	797	1,682	2,479

Table 3.14: Datasets used in the evaluation of the AraSTw and AraSenTi lexicons.

Data Set	Positive			Negative			Total	
AraCust	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**	**F1 avg**	**Accuracy**
	93.0	76.0	83.6	91.0	98.0	94.4	89.0	91.0
ASTD	79.0	65.0	71.3	76.0	96.0	84.4	77.9	85.0

Table 3.15: Evaluation results of the AraSTw lexicon on the datasets.

Data Set	Positive			Negative			Total	
AraCust	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**	**F1 avg**	**Accuracy**
	0.90	0.60	0.72	90.0	0.98	0.94	0.83	90.0
ASTD	42.83	84.69	56.89	86.49	46.43	60.42	58.65	58.73

Table 3.16: Evaluation results of the AraSenTi lexicon on the data sets.

As shown from Tables 3.15 and 3.16 that AraSTw lexicon outperformed the results of AraSenTi lexicon [212] on the same data sets. AraSTw lexicon outperformed the AraSenti lexicon by 44.7% accuracy on ASTD. In addition, AraSTw outperformed the AraSenTi lexicon by 1.11% accuracy on AraCust.

In order to find out whether there is a statistical significance between ‘the two lexicons AraSTw and AraSenTi’ on ‘ASTD corpus’, I conducted a correlation test analysis. I calculated p-value. It has been found that there is a statistical significance at the level of significance $\alpha = 0.05$, where p-value = 0.0285. Therefore, the null hypothesis (H0) was rejected, and the alternative hypothesis (H1) was accepted:

- H0: There is not statistically significant between the two lexicons.

- H1: There is a statistically significant between the two lexicons in favour to AraSTw lexicon.

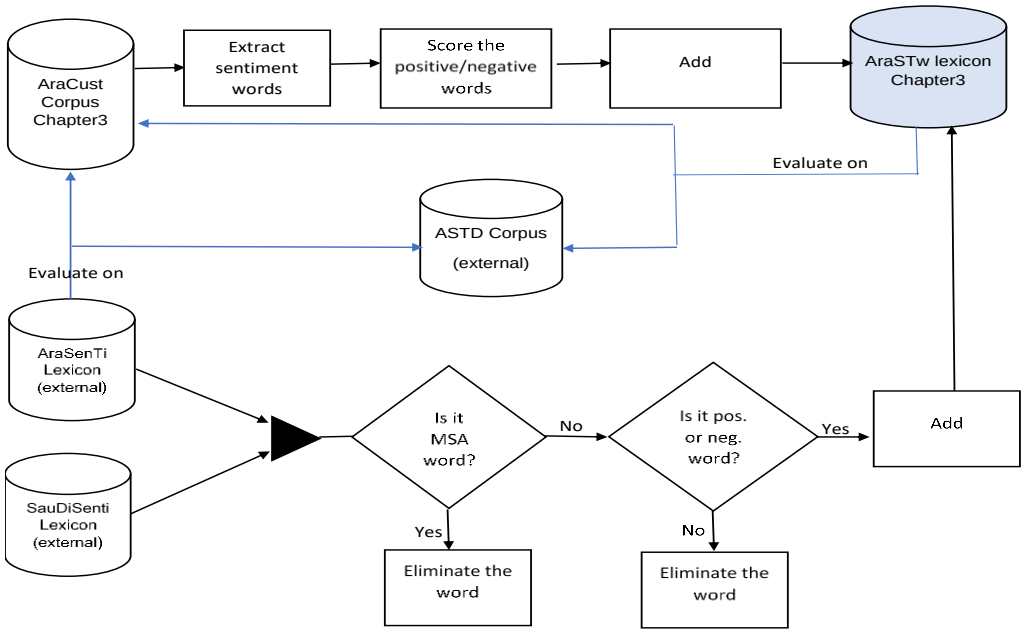

Figure 3.9: AraSTw lexicon creation and evaluation.

3.10 Summary

This paper attempts to fill the gap found in the literature by proposing the largest GSC of Saudi tweets corpus created for ASA. It is freely available to the research community. This paper describes in detail the creation and pre-processing of AraCust. In addition, this paper has explained the annotation steps that adopting for annotating the AraCust. It has described some features that were sourced from the nature of the corpus. The corpus consists of 20,000 Saudi tweets. A baseline experiment was applied on AraCust to offer benchmark results for forthcoming works. Additionally, A baseline experiment was applied on ASTD to comparing the results with AraCust. The results show that AraCust superior to ASTD. In addition, this chapter explained the details about the creation, annotation, and evaluation of the AraSTw lexicon.

Chapter 4: Metrics to Measure Customer Satisfaction and Churn

4.1 Introduction

This chapter provides the answer to *RQ1* concerning the traceable, measurable criteria for customers' satisfaction with telecom companies in Saudi Arabia and how to combine them for visualisation. Furthermore, the chapter achieves *RO1* by creating a framework of measurable weighted criteria for customers' satisfaction with Saudi telecom companies and determining a means for companies to visualise this framework through real-time graphs. Furthermore, it contributes to achieving *RO2*- namely, proposing the recommendations to improve the services of Saudi telecom companies.

I use a questionnaire to present an evaluation of metrics for measuring customer satisfaction and the possible variables that can differentiate between the behaviour of churners and non-churners. This assessment was based on a thorough review of metrics and variables used in previous works, which led to a preliminary exclusion of variables due to both the difficulty of obtaining them from telecom companies and privacy. Following this analysis, I created a taxonomy of metrics and then evaluated it based on questionnaires with telecom customers to test the metrics and the relationship between the collected variables and churning behaviour from a customer's point of view. Finally, I conducted an informal interview with a Saudi telecom expert (a telecom business consultant) to show her the collected variables and question her about other variable suggestions from the company's point of view.

The metrics are used in Chapter 6 to visualise the customers' satisfaction toward the services provided by the selected telecom companies and extract the recommendation. The collected characteristics and churn behaviours are used in Chapter 7, to predict the customer churn percentage.

4.2 Related Research

The business criteria used to develop the questionnaire were based on the telecom performance indicators specified by the Saudi Communications and Information Technology Commission [438], data annotation process and other related research (Table 4.1). Specifically, [439] found that price perceptions positively affect overall customer satisfaction, whereas [440] examined the behaviour of mobile telecommunication

customers in Hong Kong and found that transmission quality and network coverage are the most important factors driving customer satisfaction. Athanassopoulos and Iliakopoulos [441] considered positive recommendations to be the most useful loyalty indicator in European telecommunication companies and defined measures affecting customer satisfaction, such as the quality of voice transmission, access at peak time, speed of service and correct operation.

Hung et al. [62] applied data mining techniques to predict customer churn in a Taiwanese wireless telecommunication company – namely, *k*-means segmentation with the decision tree method followed by a neural network approach. The company provided the researchers with data on 160,000 customers. Subsequently, they conducted interviews with telecom experts, such as telecom business consultants, marketing analysts and mobile sales personnel, to determine the symptoms preceding customer churn and found that the age of customers, the length of tenure and the number of overdue payments all affected the churn probability. Moreover, customers who did not often make phone calls to others within the same operator's mobile network were found to be more likely to churn. The authors concluded that the performance of the neural network approach is better than that of the decision tree model.

Furthermore, [112] analysed call records (basic factors such as call duration and churner properties) using data mining – spreading activation and threshold-based and decision tree-based clustering. The authors proved that social relationships play an influential role in affecting the churn within an operator's network. Similar findings were obtained by Nguyen [442] using the data mining technique and machine learning algorithms (i.e. C4.5 decision tree, alternating decision tree, Naïve Bayes and logistic regression) to predict customer churn in mobile operators. Specifically, they used data belonging to five different categories: demographics, billing data, refill history, calling patterns and network features. Nguyen [442] found that all classifiers achieved above 60% overall accuracy [443].

Ref	Study's Aim(s)	Customer Churn Variables
[23]	Providing a framework to telecom companies for identifying potential churn customers.	Name, age, sex, income, number of minutes used, total duration of calls, number of messages, total duration of international calls and internet data usage (structure data).
[444]	Identifying a relational learner, determining the rank of the relational classifiers and investigating the collective inference methods to improve the performance of customer churn prediction models in telecommunications companies.	Seven distinct call records (DCRs) from across the world.
[20]	Empirically comparing two techniques for customer churn: decision tree and logistic regression models.	Seventeen variables categorised into five groups: demographics (subscribers' age, gender and postcode), cost, features/marketing, services usages and customer services.
[93]	Providing a model for churn prediction for telecommunication companies.	Dataset attributes such as state, area code, phone, day minutes, night minutes and churn-status.
[115]	Building a predicative churn analysis model.	The data set included telecom customer complaint data and call quality data.
[11]	Predicting the churn behaviour of customers through various data mining techniques.	Nominal and class attributes, such as phone number, international plan, voice mail plan, number of e-mail messages, total day minutes, total international calls, customer service calls and churn.
[18]	Analysing meaning of churn management in a mobile telecommunication industry and designing a new model for churn prediction.	Name, age, gender, and area; contract data, including the time the user signed in the network; call data, including call time length, roam time length and call time length among the carriers; charging data, including ARPU and complaint information; and quantitative information to describe user level.
[116]	Suggesting counter-propagation neural networks (CPNN), classification, regression trees (CART), J48 and fuzzyARTMAP to predict customer churn and non-churn in the telecommunication sector.	Different attributes of datasets, including customer dissatisfaction, switching costs, service usage, customer-related variables and customer status.
[117]	Predicting customer churn with the help of the Apache PIG.	Average call minutes per day, customers having a number of active plans being equal to zero, customers giving negative feedback, which customers had complained, which customers had zero incoming and outgoing calls in month 4, month 5, month 6 (if these conditions all match, then these customers are likely to churn).
[113]	Examining whether sentiment/mood towards a product, as measured from a large-scale collection of tweets posted on twitter.com, is correlated or even predictive of churn-rate values.	Detail status of each customer at a certain time and number of positive, negative and neutral sentiment (mood) tweets for each month.
[24]	Using customers' social network information, along with their call log details, to predict the churn users.	Users' social network information (Pokec) to find the influential user and call log details to help improve the accuracy of prediction.
[116]	Addressing the causes of customer churn.	Factors such as differences in prices, poor customer service, billing issues and failure of network coverage.

Table 4.1: Customer Churn Variables in Previous Studies.

4.3 Methodology

To determine the possible variables that can differentiate between the behaviour of churners and non-churners, first, I collected the variables from a literature review, as explained in the section above; the review is summarised in Table 4. 1. To be built as a block, research must relate to previous knowledge [445]. Webster & Waston [446] considered a well-constructed literature review as an effective tool to develop a firm a theory. Likewise, [445] considered a literature review as an excellent tool to provide evidence of research outcomes on a meta-level and expose the areas in which research is required. Hence, I used a literature review to collect churn variables and customer satisfaction metrics.

After the literature review variable extraction, I provided the questionnaire to telecom customers. One of the most frequent tools of the quantitative method is a questionnaire [447] (section 3.1). *Questionnaire* is defined as a survey scheme that analyses a population sample to offer a numerical overview of population tendencies, sentiment or opinions [448]. The ability of a questionnaire to collect data from a large sample in a short time [449] with low costs [50], as well as its ability to represent the features of a community, were the primary reasons for using a questionnaire to measuring the importance of the customer satisfaction metrics from the customers' points of view. Additionally, the questionnaire was used to collect the churn causes and test the relation between churn behaviour and churner characteristics. More details on its construction are in sub-section 4.3.1.

After the questionnaire, I conducted informal interviews with a Zain telecom expert (a telecom business consultant) to gather more in-depth information on churning variables.

Based on the literature review, questionnaire and interview results, I developed the variables that could help us to predict customer churning.

4.3.1 Questionnaire Construction

The aims of the questionnaire were:

Qo1: To define customer satisfaction metrics from the customers' perspective.

Qo2: To understand churning causes and churner characteristics and behaviours.

Qo3: To define the differences between the telecom companies.

The questionnaire consisted of the following four parts, Appendix B:

Part A: Demographics variables: it contained three questions that addressed the following personal and socio-demographic data of the customers: age, gender and (optional) Twitter account. This part responds to the following fourth research: to predict the potential ratio of customer churn by defining the demographic variables relating to customer churn behaviour used as input variables in our prediction model.

Part B: Behaviours and characteristics of customers who change telecommunication providers: it contained seven questions to identify the characteristics and behaviours that correlate with customer churn behaviour. Furthermore, it included an open question about the reasons for changing telecommunication provider. This question provided insight for telecom companies about the reasons behind customer churn from the customers' perspective, in line with the fourth research objective.

Part C: Communication methods: it contained three questions about the method that customers used to make complaints, requests or suggestions. This section aimed to confirm that customers use the official Twitter accounts of the Saudi telecom companies for communication purposes. Furthermore, I sought to find any other communication methods used by customers of each telecommunication company.

Part D: Customer satisfaction towards telecom companies: it contained nine statements that identified the various metrics of customer satisfaction towards the telecom companies (section 2). The customer chose one item that describes the importance of customer satisfaction from their perspective. This part D reflects the first research question, which is, 'What are the traceable, measurable criteria for customers' satisfaction with telecom companies in Saudi Arabia?'.

4.3.2 Study Sample

To answer the research questions and validate the study hypotheses, it was necessary to use a large sample; [447] recommended having a large number of responses. To avoid bias in the sample, I have:

- Identified the population.
- Specified the sample frame.

- Chosen the right sampling technique.

The population in this study is STC, Mobily and Zain customers. The sample frame is an adult over 18 years old, using three Saudi telecom companies for post-paid voice service. The sample was calculated to ensure that it reflected the population features of the study. A sample size calculator[15] was used to identify the optimal sample size, with a confidence level of 95% and a confidence interval (i.e., the margin of error) of 5%, resulting in a sample size of 384. Afterwards, the probability sample technique simple random sample (SRS) was chosen. The advantage of SRS is that all of the population has the same chance of being selected [418]. Although some studies consider that a sample size of more than 200 responses is reasonable [450] , I gathered the largest number of responses possible. The total number of participants was 445. After filtering (i.e., removing incomplete answers), the remaining sample contained 437 answers. The responses and results were stored automatically when a participant finished the questionnaire.

4.3.3 Data Collection

There are different methods to distribute and collect questionnaires, such as online or by hand [451]. The questionnaire in this study was sent to participants through some social media platforms (Twitter and Whats app) and by e-mail to the staff and students at many Saudi universities to avoid bias in the outcomes and generalise results. Online distribution tools were used because they made it easier to administer the questionnaire to a wide range of participants and collect a large number of responses to ensure that the sample reflected the attributes of the community [452],[447]. Additionally, it saved time [453] and was cost-effective [454]. The questionnaire rule guide and the information section defining the research aims were included in the questionnaire cover letter.

4.3.4 Pilot Study

Before administering the whole sample, the questionnaire was piloted to confirm that the questions were clear and ensured validity and efficiency [451]. The pilot study was conducted with ten adults post-paid customers for different telecommunication companies in Saudi Arabia. Subsequently, to measure the validity and reliability of all the questionnaire elements, Cronbach's alpha test was used through the Statistical

[15] https://www.surveysystem.com/sscalc.htm

Package for Social Science (SPSS). The rate of reliability and validity for the questionnaire in terms of Cronbach's alpha [455] was 0.853, which is greater than the 0.7 cut-off point [456] and indicates the degree of consistency and clarity of questions of the questionnaire. The questionnaire was ready for distribution after it was evaluated and modified based on feedback.

4.3.5 Data Analysis

The SPSS software tool is generally used for the analysis of social science surveys [451]. Hence, SPSS was used to conduct a correlation analysis to investigate the relationships between the variables in this study.

The mean, frequencies, percentages and standard deviation are the most well-known statistical tools used in the descriptive approach [451],[457]. Additionally, in this study, the Kolmogorov-Smirnov normality test [458] was used to check the normality distribution of the sample. The aim was to choose appropriate correlation tests for both questions and respondents. The correlation analysis tests, chi-square and Kruskal-Wallis tests were applied to find out whether there were significant relationships between the variables. Furthermore, a thematic text analysis [459] was used for the open-ended question. The answers required classification before they could be analysed, which was implemented through a text analysis tool that depends on thematic text analysis (i.e., analysis looking for the occurrence of themes) [460]. This analysis was based on a coding system [460]. There are two types of thematic text analysis applied to survey data: instrumental thematic text analysis (which entails using a computer program for coding and texts interpretation from the researcher's point of view) and representational thematic text analysis (the coding is done by a human and uses text interpretation from the author's point of view) [460]. In the latter, the process unfolds as follows: gaining a good understanding of the text and what is also between the lines; looking for the theme in each fragment; and locating the code for each fragment. It is time-consuming, but it helps to avoid idiomatic ambiguity in the identification of themes [460]. In this research, manual coding was performed whenever short answers to open-ended questions were available.

A 5-point Likert-scale was used in the questionnaire to record the level of customer satisfaction concerning the services provided by a telecommunication company.

The questionnaire used a scale ranging from 1 = not important to 5 = very important, as follows:

- Very important, weighted 5.
- Important, weighted 4.
- Neither important nor unimportant (Neutral), weighted 3.
- Unimportant, weighted 2.
- Very unimportant, weighted 1.

The relative importance index (RII) [461] was used to rank the metrics based on their importance for each telecom company.

I calculated the *RII (m)* based on the equation, where *m* is the metric:

$$RII(m) = \frac{\sum i*Wi}{A*N}, \qquad (4.1)$$

where *Wi* is the weight of index *i*, with $i \in \{1\ to\ 5\}$, *A= 5* (the highest weight) and *N* is the total number of respondents for all weights.

4.3.6 Ethical and Legal Issues

Every project influences human interest differently through legal, ethical, social and professional impacts. This section addresses this Chapter's work ethical and legal issues.

Project ethics, as defined in [447], are ethical rules that are important to follow during a project for several reasons; some of them also affect a project's validity and reliability. The rules defined by [447] concerning the data collection stage are that the researcher must obtain authorisation from the target sample, the privacy and confidence of participants' information must be ensured, and questionnaire information must be saved and stored in a secure place. The process to avoid any legal and ethical issues was followed, and the ethical form was submitted to the Institutional Review Board (IRB) at Princess Nourah bint Abdulrahman University (PNU), Appendix C.

4.4 Data Analysis and Results

4.4.1 Part A: Demographic Variables

This section reports the results for each question in Part A of the questionnaire.

As shown in Table 4.2, there were 320 female respondents (73.2%) and 117 male respondents (26.8%) in this study. The options allowed us to analyse respondents as mapped over six age groups, based on the report from the Saudi Communications and Information Technology Commission [416] that categorised the users of social media in Saudi Arabia in this way.

As can be seen in Table 4.2, most respondents were between 35 and 44 years old (200 respondents, or 45.8%), while respondents over 65 years of age constituted the smallest portion of the sample (5 respondents, or 1.1%). Out of 424 respondents, 298 (70.2%) used STC as their telecom company, 97 (22.9%) Mobily, 23 (5.4%) Zain and six respondents (1.4%) other companies. As can be seen, the data is highly unbalanced; this is because the sample reflects a population – the one of Saudi Arabia – where STC customers are much more numerous than Mobily customers and there are more Mobily than Zain customers.

Age Group	Frequency	Percentage
18–24	26	5.9
25–34	112	25.6
35–44	200	45.8
45–54	66	15.1
55–64	28	6.4
65+	5	1.1
Total	437	100.0
Gender		
Female	320	73.2
Male	117	26.8
Total	437	100.0
Telecom Company		
Valid STC	298	70.2
Mobily	97	22.9
Zain	23	5.4
Others	6	1.4
Total	424	100
Missing	13	3.0
Total	437	

Table 4.2: Demographic Variables.

4.4.2 Part B: Behaviours and Characteristics of Participants who Changed their Telecommunication Company.

This section reports the results for each question in Part B of the questionnaire.

Did you change your telecommunication company before?

Over 437 respondents (311, or 71.2%) answered that they had never changed their telecommunication company, while the remaining respondents (126, or 28.8%) had done so (Figure 4.1).

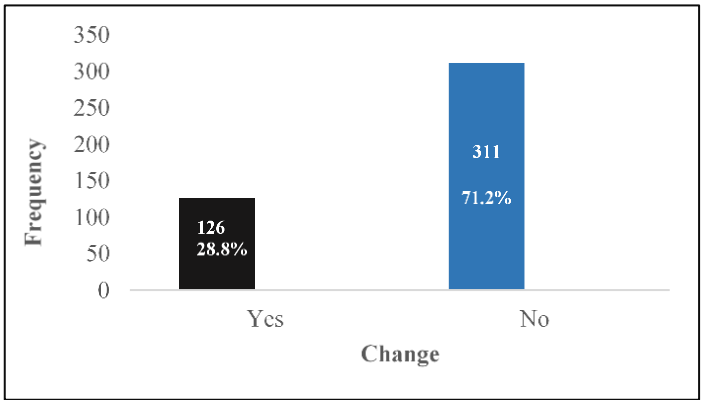

Figure 4.1: Frequency of participants who had changed telecommunication companies before.

What was the previous telecommunication company that you used as a cell phone network?

To answer, respondents had to choose one telecom company. Table 4.3 shows that the Zain company received most of the responses (45 responses, or 35.7%). STC came second with 41 responses (32.5%), while Mobily was last with 38 (30.1%).

	Frequency	Percentage
STC	41	32.5%
Mobily	38	30.1%
Zain	45	35.7%
Others	2	1.5%
Missing	311	71.1%
Total	437	100.0

Table 4.3: Previous Telecom Company.

For how long did you use the previous telecommunication company?

In Figure 4.2, 50 respondents (39.6%) indicated that they used their previous telecom company for 1 to 5 years, while 16 respondents (12.6%) used their previous telecom company for less than one year.

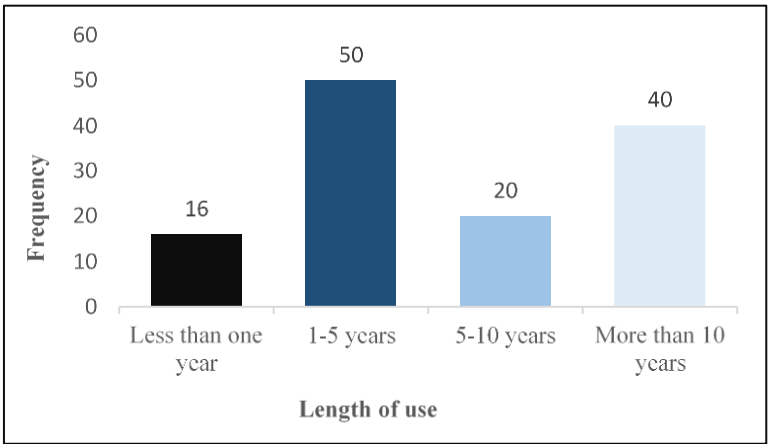

Figure 4.2: Length of using the previous telecommunication company.

Before you left the previous telecom company, did you have overdue payments?

As shown in Figure 4.3, the number of respondents who had overdue payments before leaving their previous telecom company constituted the majority, with 58 responses, or 46%, The 'no overdue payments' respondents were the next in size (44).

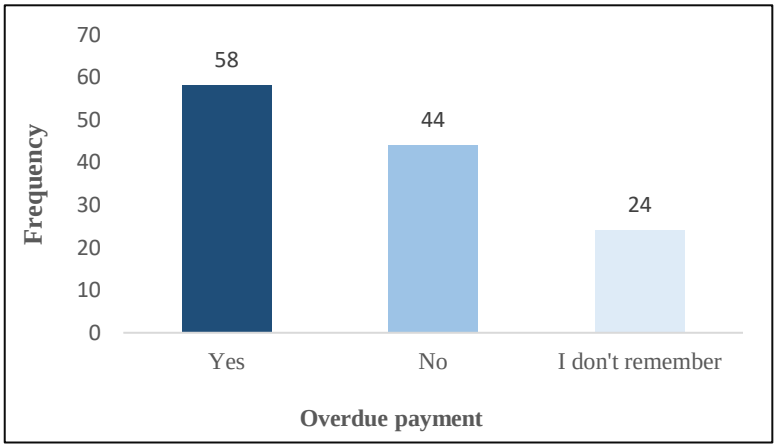

Figure 4.3: Frequency of having overdue payments.

Has one of your family members ever used your previous telecommunication company as their cell phone network?

The majority of respondents –100 responses – stated that one of their family members had used their previous telecommunication company, whereas 18 indicated that none had, and eight respondents did not know (Figure 4.4).

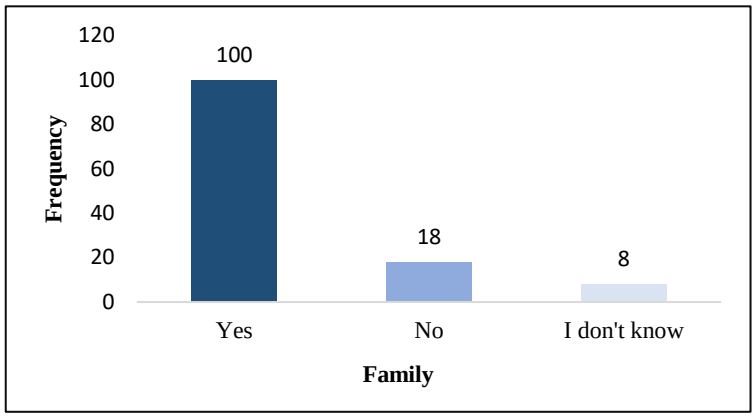

Figure 4.4: Frequency of having family members using the respondents' previous telecommunication company.

Why did you change your previous telecommunication company?

To obtain deeper insight into customer satisfaction with telecommunication company services and reasons for customer churn, this question was posed as an open question. Each respondent had to fill in the blank. These types of questions provide investigators with more information and explanations than can be captured through closed questions.

The thematic text analysis [459] was performed by a human coder through Excel, identifying nine themes that could be allocated to each response (Table 4.4). Each theme had its own code and represented one service provided by the telecom company. Responses without a stated reason (20 responses) were deleted. The responses were then divided among the three companies.

Themes	Code
Quality of voice transmission	QV
Customer service	CS
Billing price	BP
Unreasonable fees when calling someone who uses another telecom company	RF
Network coverage	NC
Internet browsing speed	BS
Technical issues	TI
Unsuccessful calls	SC
Bad offers	OF

Table 4.4: Themes and Codes for the Services Provided by Telecom Companies.

As shown in Table 4.5, STC received the highest number of responses (164). Mobily received 132 responses, making it the second-highest company in terms of the number of received responses. Zain company received the least responses (61).

Telecom Company	Frequency	Percentage
Valid STC	164	45.9
Mobily	132	36.9
Zain	61	17.08
Total	357	100
Missing	80	18.3
Total	437	

Table 4.5: The frequency and percentage of the responses based on companies.

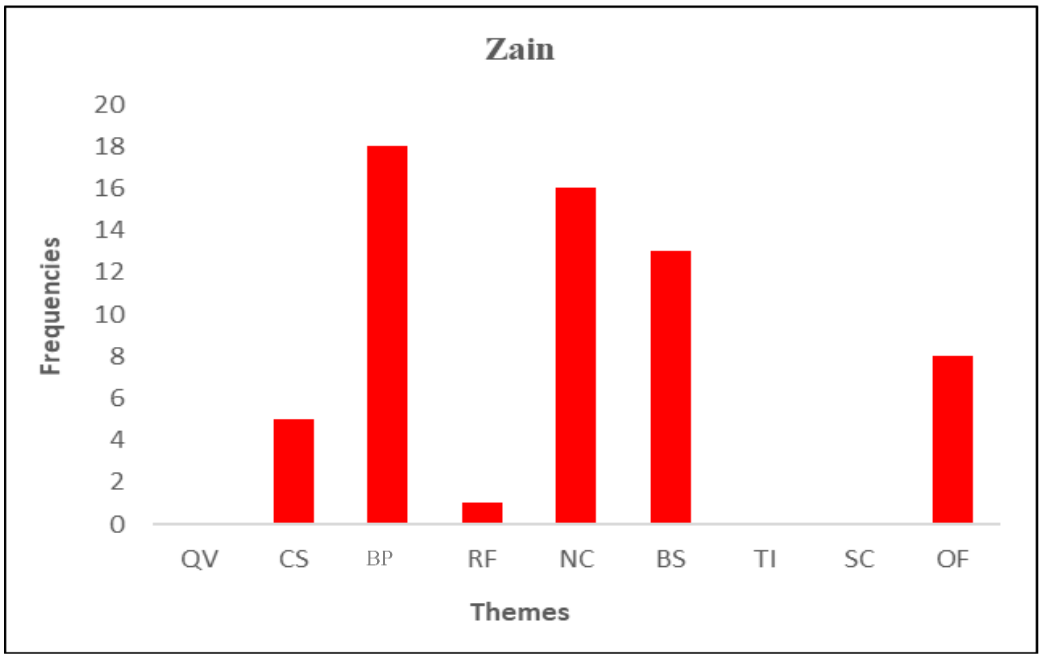

Figure 4.5: Frequency of the reasons behind Zain company's customer churn.

Figure 4.5, 'high billing price' was the most listed reason, mentioned in 18 responses, while 'bad quality of voice transmission', 'unsuccessful calls' and 'technical issues' received no responses.

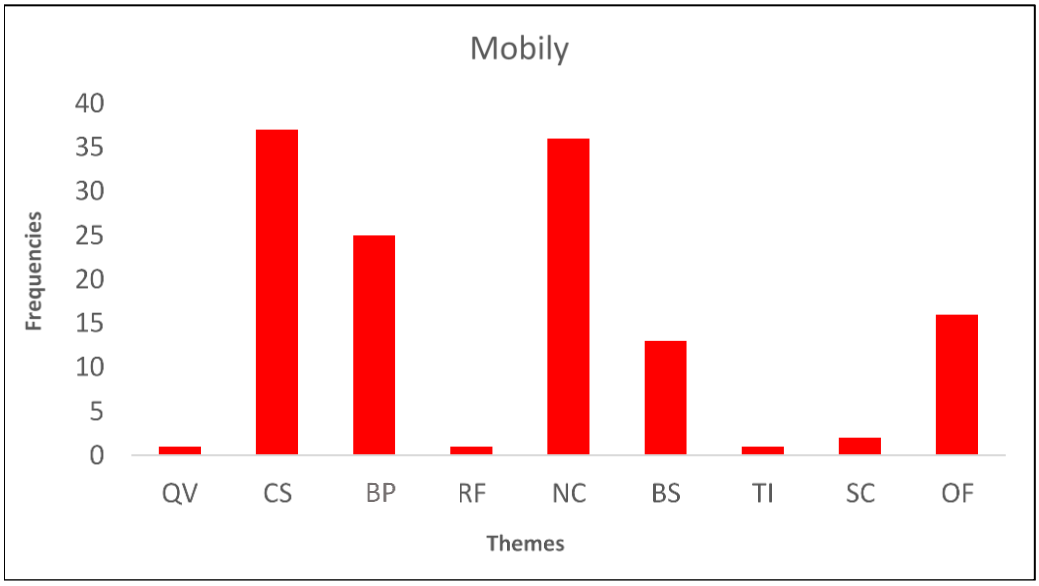

Figure 4.6: Frequency of the reasons behind Mobily company's customer churn.

As demonstrated in Figure 4.6, 'slow response of customer service' was the first churn reason mentioned (28%, 37 responses), followed by 'bad or lack of network coverage' with (27%, 36 responses), 'high billing price' with 25 responses, whereas 'bad quality of voice transmission', 'technical issues' and 'unreasonable fees when calling someone who uses another telecom company' each received one response.

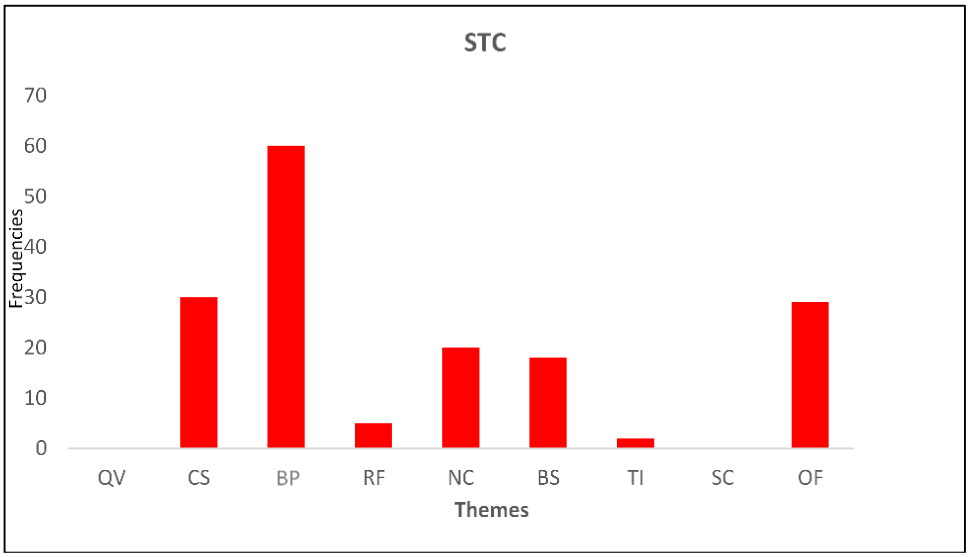

Figure 4.7: Frequency of the reasons behind STC company's customer churn.

'High billing price' was the most frequently mentioned reason (37%, 60 responses; Figure 4.7), followed by 'slow response of customer service', allocated as the second reason (18%, 30 responses), whereas 'bad quality of voice transmission' and 'unsuccessful calls' received no responses.

4.4.3 Part C: Communication Methods

This section reports the results for each question in Part C of the questionnaire.

Do you use the web or social media platforms to communicate with the telecommunication company (for example, for complaints or suggestions)?

There were 262 respondents who did not use the web or social media to communicate with their telecom company, while 174 respondents did (Figure 4.8).

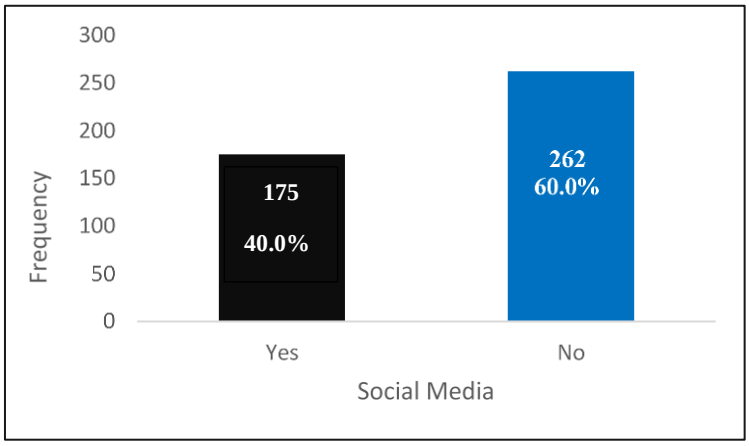

Figure 4.8: Frequency of using the web or social media platforms as communication methods to contact the telecommunication company.

What type of methods do you currently use to communicate with your telecom network (for example, for complaints, requests or suggestions)?

As shown in Table 4.6, Twitter was the most frequently used communication tool, receiving the highest number of responses (175, 40%). 'Using a telecommunication company's application to communicate' received 115 responses, whereas 59 respondents chose 'telephone'.

	Frequency	Percentage
Twitter	175	40.0%
Telecom company application	115	26.3%
Telecom company website	88	20.1%
Telephone	59	13.5%
Total	437	100%

Table 4.6: Communication Methods.

Do you think that service quality has been enhanced because of your communication with your telecom company through Twitter (for example for complaints, requests or suggestions)?

More than half of the respondents (245, 56.06%) thought that service quality was enhanced after using Twitter as a communication method with a telecom company (for example, for complaints, requests or suggestions), whereas 192 (43.94%) answered the opposite (Figure 4.9).

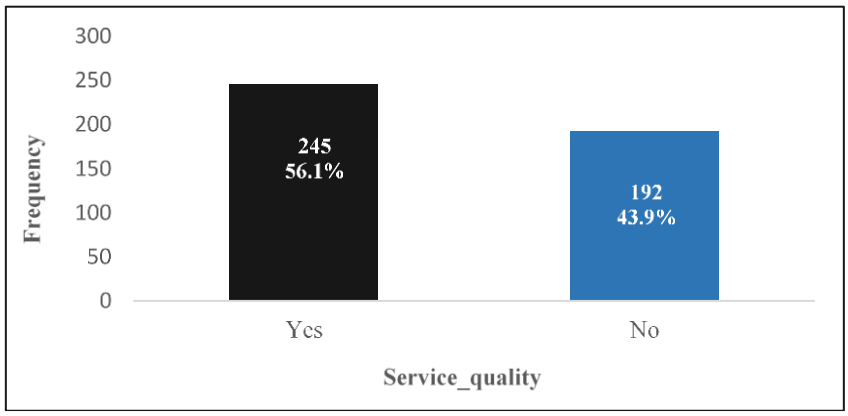

Figure 4.9: Frequency of service quality after using Twitter as a communication method with a telecommunication company.

4.4.4 Part D: Customer satisfaction metrics towards the telecom companies

This section reports the results for each question in Part D of the questionnaire. Table 4.7 shows that 98 participants considered 'good network coverage' standard to be a very important satisfaction metric, and 172 marked it an important metric. Sixty judged it as an unimportant metric and only 28 as very unimportant. Thus, this metric was considered important by the majority (270 participants) rather than unimportant (88 participants).

		Frequency	Percentage
Validity	**Very important**	98	22.4
	Important	172	39.4
	Neutral	79	18.1
	Unimportant	60	13.7
	Very unimportant	28	6.4
Total		437	100.0

Table 4.7: Frequency Table for the 'Good Network Coverage' Metric.

As shown in Table 4.8, the number of participants who considered 'good quality of voice transmission' as an important satisfaction metric was 196 – significantly more than the 35 participants who considered it as an unimportant metric.

		Frequency	Percentage
Valid	**Very important**	116	26.5
	Important	196	44.9
	Neutral	75	17.2
	Unimportant	35	8.0
	Very unimportant	15	3.4
	Total	437	100

Table 4.8: Frequency Table for the 'Good Quality of Voice Transmission' Metric.

Regarding the 'quick response provided from customer service' metric, Table 4.9 shows that 134 respondents chose it as an important satisfaction metric, and 88 chose it as a very important one. For 38 respondents, it was a very unimportant metric, and 81 selected it as an unimportant one.

		Frequency	Percentage
Valid	**Very important**	88	20.1
	Important	134	30.7
	Neutral	96	22.0
	Unimportant	81	18.5
	Very unimportant	38	8.7
	Total	437	100

Table 4.9: Frequency Table for the 'Quick Response Provided from Customer Service' Metric.

Table 4.10 shows that the number of participants who decided that the 'number of successful calls' is an important metric was higher (106 important, 100 very important) than the number of those for whom this metric was unimportant (90 unimportant, 47 very unimportant).

		Frequency	Percentage
Valid	**Very important**	100	22.9
	Important	106	24.3
	Neutral	94	21.5
	Unimportant	90	20.6
	Very unimportant	47	10.8
	Total	437	100.0

Table 4.10: Frequency Table for the 'Number of Successful Calls' Metric.

As shown in Table 4.11, 43.0% of the participants counted 'billing price' as an important metric, whereas 8.9% counted it as an unimportant metric.

		Frequency	Percentage
Valid	**Very important**	129	29.5
	Important	188	43.0
	Neutral	67	15.3
	Unimportant	39	8.9
	Very unimportant	14	3.2
	Total	437	100.0

Table 4.11: Frequency Table for the 'Billing Price' Metric.

The percentage of participants that considered 'high internet speed' to be an important standard was 24.9%, while the percentage of participants that considered it to be unimportant was the lowest (16.2%; Table 4.12).

		Frequency	Percentage
Valid	**Very important**	75	17.2
	Important	109	24.9
	Neutral	90	20.6
	Unimportant	92	21.1
	Very unimportant	71	16.2
	Total	437	100

Table 4.12: Frequency Table for the 'High Internet Speed' Metric.

Table 4.13 shows that 108 participants selected the 'reasonable fees when calling someone who uses another telecom company' metric as an unimportant metric, whereas 91 participants selected it as an important metric. It was selected as very important by 89 participants – a higher number than those who set it as very unimportant (62 participants).

		Frequency	Percentage
Valid	**Very important**	89	20.4
	Important	91	20.8
	Neutral	87	19.9
	Unimportant	108	24.7
	Very unimportant	62	14.2
	Total	437	100

Table 4.13: Frequency Table for the 'Reasonable Fees When Calling Someone Who Uses Another Telecom Company' Metric.

Table 4.14 shows that 110 participants selected 'good offers' as an unimportant metric, whereas 120 participants selected it as an important one. It was selected as very important by 89 participants, which was more than the number of those who set it as very unimportant (55).

		Frequency	Percentage
Valid	**Very important**	89	20.4
	Important	120	27.5
	Neutral	63	14.4
	Unimportant	110	25.2
	Very unimportant	55	12.6
	Total	437	100

Table 4.14: Frequency Table for the 'Good Offers' Metric.

4.4.5 Cross Tables

A correlation test analysis was used to find out whether the same variables drove any factors. A relationship was indicated between all factors where the degree of correlation coefficient above 50%. However, a lack of relationship indicated in this way does not mean that it does not exist at all [462].

- **'Did you change your telecommunication company before?' versus customer gender.**

In order to find out whether there is a relationship between the customer 'changing the telecom company before' and 'customer gender', I conducted a correlation test analysis. I used the chi-square test and *p*-value.

It has been found that there is a statistical significance at the level of significance $\alpha = 0.05$, where *p*-value = 0.007. As shown in Table 4.15, the percentage was higher for females who did not change telecom company (54%) and the males who did change telecom company (10.3%; Figure 4.10). Therefore, the null hypothesis (H0) was rejected, and the alternative hypothesis (H1) was accepted:

- **H0:** There is no significant relationship between the two variables.
- **H1:** There is a significant relationship between the two variables.

			Gender		Total	Chi-Square	*P*-Value
			Female	Male			
Change	Yes	Count	81	45	126	7.219	0.007
		% of Total	18.5%	10.3%	28.8%		
	No	Count	239	72	311		
		% of Total	54.7%	16.5%	71.2%		
Total		Count	320	117	437		
		% of Total	73.2%	26.8%	100.0%		

Table 4.15: The Correlation Between the 'Changing the Telecommunication Company' and 'Gender' Variables.

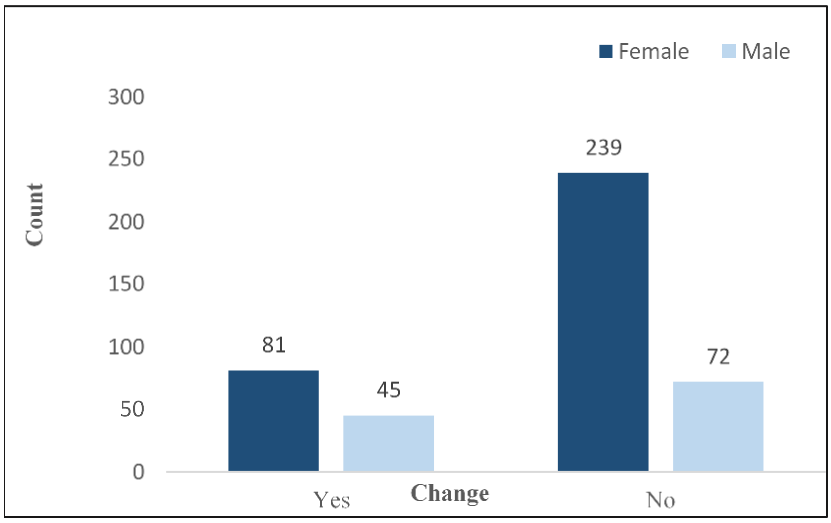

Figure 4.10: The correlation between the 'changing the telecommunication company' and 'gender' variables.

- **'Did you change your telecommunication company before?' versus customer age.**

The values of chi-square = 79.93 and *p*-value = 0.042 were obtained through a review of the correlation matrix (Table 4.16). A statistical significance at the level of significance $\alpha = 0.05$ was found, which indicates a correlation between the two variables 'changing the customer telecom company and 'customer age'. Thus, the null hypothesis (H0) is rejected, and the alternative hypothesis (H1) is accepted. As shown in Table 4.16, the 35–44 age group was the one in which the highest number of participants did not change their telecom company (31.8%), and the 65+ age group was the group with the lowest number of participants who had changed their telecom company (0%).

			Age Group						Total	Chi-Square	*P*-Value
			18- 24	25-34	35-44	45-54	55-64	65+			
change	Yes	Count	10	37	61	10	8	0	126	9.937	0.042
		% of Total	2.3%	8.5%	14.0%	2.3%	1.8%	0%	28.8%		
	No	Count	16	75	139	56	20	5	311		
		% of Total	3.7%	17.2%	31.8%	12.8%	4.6%	1.1%	71.1%		
Total		Count	26	112	200	66	28	5	437		
		% of Total	5.9%	25.6%	45.8%	15.1%	6.4%	1.1%	100.0%		

Table 4.16: The Correlation Between the 'Changing Telecommunication Company' and 'Age' Variables.

- **'Did you have overdue payments?' versus 'What telecommunication company have you used for your mobile?'**

By reviewing the correlation matrix (Table 4.17), values of chi-square = 2.195 and *p*-value = .901 were obtained, which indicates that a statistical significance at the level of significance $\alpha = 0.05$ was not found, entailing the lack of correlation between 'type of previous telecommunication company' and 'having overdue payments. Thus, the null hypothesis (H0) is accepted, and the alternative hypothesis (H1) is rejected. As shown in Table 4.17, the highest count (27) was found for Mobily customers who had overdue payments, whereas the lowest count (1) was found for Zain customers who did not remember. I ignored other responses because a limited number of participants chose the 'others' answer choice.

			What telecommunication company have you used for your mobile services?				Total	Chi-Square	*P*-Value
			STC	Mobily	Zain	Others			
Before you left the previous telecommunication company of your mobile phone, did you have counts of overdue payments?	Yes	Count	23	27	21	1	72	2.195	.901
	No	Count	12	8	23	1	44		
	I don't remember	Count	6	3	1	0	10		
						.			
Total		Count	41	38	45	2	126		
		%	32.5%	30.1%	35.7%	1.5%	100.0%		

Table 4.17: The Correlation between 'Changing the Telecommunication Company' and the 'Having Counts of Overdue Payments'.

- **'Length of using the previous telecom company' versus 'What telecommunication company have you used for your mobile services?'**

The review of the correlation matrix (Table 4.18) gave values of the chi-square = 23.02 and *p*-value = 0.006; therefore, a statistical significance at the level of significance $\alpha = 0.05$ was found, indicating a correlation between 'type of previous telecommunication company' and 'length of using the previous telecom company'. Thus, the null hypothesis (H0) is rejected, and the alternative hypothesis (H1) is accepted. As shown in Table 4.18, the highest percentage, 44.4%, was for Zain customers who had used the telecom company for 1 to 5 years. However, the lowest percentage, 4.8%, was for STC customers who had used the telecom company for 5 to 10 years. I ignored 'other' responses because of the limited number of participants that chose the 'others' answer choice.

			What telecommunication company have you used for your mobile services?				Total	Chi-Square
			STC	Mobily	Zain	Others		
How long did you use the previous telecommunication company?	Less than one year	Count	8	3	4	1	16	23.024
		%	19.5%	7.9%	8.9%	50%	12.7%	
	1-5 years	Count	14	16	20	0	50	
		%	34.1%	42.1%	44.4%		39.7%	
	5-10 years	Count	2	10	7	1	20	
		%	4.8%	26.3%	15.6%	50%	15.9%	
	More than 10 years	Count	17	9	14	0	40	
		%	41.4%	23.7%	31.1%		31.7%	
Total		Count	41	38	45	2	126	
		%	100.0%	100.0%	100.0%	100.0%	100.0%	

Table 4.18: The Correlation between the Changing the Telecommunication Company and the Length of Using the Previous Telecom Company Variables.

- **'Does one of your family members use your previous telecommunication company?' versus 'What telecommunication company have you used for your mobile services?'**

By reviewing the correlation matrix, shown in Table 4.19, values of chi-square = 10.356 and *p*-value = 0.110 were obtained. No statistical significance at the level of significance $\alpha = 0.05$ was found. This indicates that there is no correlation between the variables the 'Type of the previous telecommunication company' and the 'Having of customers' family members using the previous telecommunication company of that customer'. Thus, the null hypothesis (H0) is accepted, and the alternative hypothesis (H1) is rejected. As shown in Table 32, the highest percentage, 89.8%, was for Mobily customers having one of their family members currently using their previous telecommunication company, whereas the lowest percentage, 2.4%, was for STC customers who chose the 'didn't remember' answer choice for whether they had one their family member using their previous telecommunication company. I ignored the 'other' responses because a limited number of participants chose that option.

			What telecommunication company have you used for your mobile services?				Total	Chi-Square test	*P*-value
			STC	Mobily	Zain	Others			
Is one of your family members currently using your previous telecommunication company for their mobile phone?	Yes	Count	35	34	31	0	100	10.356	.110
		%	85.4%	89.8%	68.9%		79.4%		
	No	Count	5	3	10	0	18		
		%	12.2%	7.9%	22.2%		14.3%		
	I don't know	Count	1	1	4	2	8		
		%	2.4%	2.6%	8.8%	100%	6.3%		
Total		Count	41	38	45	2	126		
		%	100.0%	100.0%	100.0%	100.0%	100.0%		

Table 4.19: The Correlation between 'Type of Previous Telecommunication Company' and Having of Customers' Family Members Using the Previous Telecommunication Company of that Customer'.

- **'What telecommunication company have you used?' versus 'Communication method'.**

As shown in Table 4.20, 130 STC customers used Twitter as a communication method, 69 used the company application, 62 used the company website, and 37 the telephone. Ten Zain customers used Twitter as a communication method, whereas 'using the telephone' received the lowest count for Zain customers with two participants. Mobily customers used the company application as their primary communication method (39 participants), followed by Twitter as the second preferred communication method (25 participants). Furthermore, Table 4.20 shows that the values of the chi-square = 10.283 and the *p*-value = 0.036 were found as statistically significant at α = 0.05. Therefore, the null hypothesis is rejected, indicating a correlation between 'telecom company type' and 'communication method [*used to complain or make suggestions*]'. The highest percentage of those who communicate with the company through Twitter were STC customers (130 participants, 78.8%), whereas the lowest percentage of those who communicated with the company through telephone included Zain customers (two participants, 3.4%).

			Communication Method				Total	Chi-Square	*P*-Value
			Twitter	Website	Application	Telephone			
Type of Telecom Company	STC	Count	130	62	69	37	298	10.283	0.036
		% of Total	78.8	77.5	60.5	62.7	71.3		
	Mobily	Count	25	13	39	20	97		
		% of Total	15.1	16.3	34.2	34.0	23.2		
	Zain	Count	10	5	6	2	23		
		% of Total	6.0	6.3	5.3	3.4	5.5		
Total		Count	165	80	114	59	418		
		% of Total	100.0%	100.0%	100.0%	100.0%	100.0%		

Table 4.20: The Correlation between 'Type of Telecommunication Company' and 'Communication Method'.

- '**Do you think that service quality has been enhanced as a result of your communication with your telecom company on Twitter?' versus 'What telecommunication company have you used for your mobile?'**

Table 4.21 shows that the values for chi-square = 5.606 and *p*-value = 0.132 were not statistically significant at $\alpha = 0.05$. Therefore, the null hypothesis is accepted, which means that there is no correlation between 'telecom company type' and 'probability of enhancing service quality after using Twitter as a communication method'. The highest percentage, 59.1%, was for STC customers who thought that service quality was enhanced as a result of customer communication with the telecom company through social media. In contrast, the lowest percentage, 30.4%, was for Zain customers who thought that service quality was not enhanced because of it.

			What telecommunication company have you used for your mobile services?				Total	Chi-square test	*P*-Value
			STC	Mobily	Zain	Others			
Do you think that quality has been enhanced as a result of your communication with your Telecom Company in social media?	Yes	Count	176	48	16	2	242	5.606	.132
		%	59.1%	49.5%	69.6%	33.3%	57.1%		
	No	Count	122	49	7	4	182		
		%	40.9%	50.5%	30.4%	66.7%	42.9%		
Total		Count	298	97	23	6	424		
		%	100.0%	100.0%	100.0 %	100.0	100.0%		

Table 4.21: The Correlation between 'Type of Telecommunication Company' and 'Probability of Enhancing Service Quality after Communication with a Telecom Company through Social Media'.

- **The importance of customer satisfaction metrics for each telecom company.**

Before deciding on the appropriate test to analyse the correlation between the importance of the customer satisfaction metric and telecom companies, I tested the normality of the sample per metric for each telecom company (STC, Mobily and Zain) using the Kolmogorov-Smirnov Normality test [458] to check the normality distribution of the sample (Table 4.22). The purpose of this was to choose the appropriate correlation test for the questions and respondents. I found that all metrics were significant and lesser than the significant level $\alpha = 0.05$, which indicated that the sample was not normally distributed. Therefore, the non-parametric Kruskal-Wallis test was applied to test more than three independent groups not normally distributed and find out whether there were significant relationships between the variables.

Table 4.23 shows that the *p*-values for 'reasonable fees when calling someone who uses another telecom company' and 'high internet speed' are .025 and 0.03 – both < 0.05, indicating that these two metrics are significant. The 'reasonable fees when calling someone who uses another telecom company' metric was significant for Zain customers, as well as the 'high internet speed' metric.

Kolmogorov-Smirnov[a]	Good Offers			Good Quality of Voice Transmission			Quick Response Provided by Customer Service			Number of Successful Calls			Billing Price			Reasonable Fees			Good Network Coverage			Internet Speed		
	STC	Mobily	Zain	STC	Mobily	Zain	STC	Mobily	Zain	STC	Mobily	Zain	STC	Mobily	Zain	STC	Mobily	Zain	STC	Mobily	Zain	STC	Mobily	Zain
df	277	91	.000	273	89	23	274	91	22	.270	90	.000	273	90	23	272	91	23	273	90	22	270	88	22
Sig.	.000	23	.034	.000	.000	.019	.000	.000	.000	.213	23	.008	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000

Table 4.22: Normality Test.

a. Lilliefors Significance Correction

Kruskal-Wallis Test	Good Offers			Good Quality of Voice Transmission			Quick Response Provided by Customer Service			Number of Successful Calls			Billing Price			Reasonable Fees			Good Network Coverage			High Internet Speed		
	STC	Mobily	Zain	STC	Mobily	Zain	STC	Mobily	Zain	STC	Mobily	Zain	STC	Mobily	Zain	STC	Mobily	Zain	STC	Mobily	Zain	STC	Mobily	Zain
N	298	97	23	298	97	23	298	97	23	298	97	23	298	97	23	298	97	23	298	97	23	298	97	23
Mean Rank	198.8	187.9	193.8	198.8	180.0	173.6	190.3	197.7	224.0	200.0	171.9	199.5	188.3	199.7	230.3	185.9	202.5	246.7	186.7	208.2	207.7	179.0	210.7	243.9
Chi-Square	.704			3.097			2.109			5.023			3.554			7.397			3.081			11.426		
Asymp Sig.	.703			.213			.348			.081			.169			.025			.214			.003		

Table 4.23: The Kruskal-Wallis Test for the relation between the Importance of Customer Satisfaction metrics and each Telecom company.

The importance of customer satisfaction metrics versus 'What telecommunication company have you used for your mobile services?'

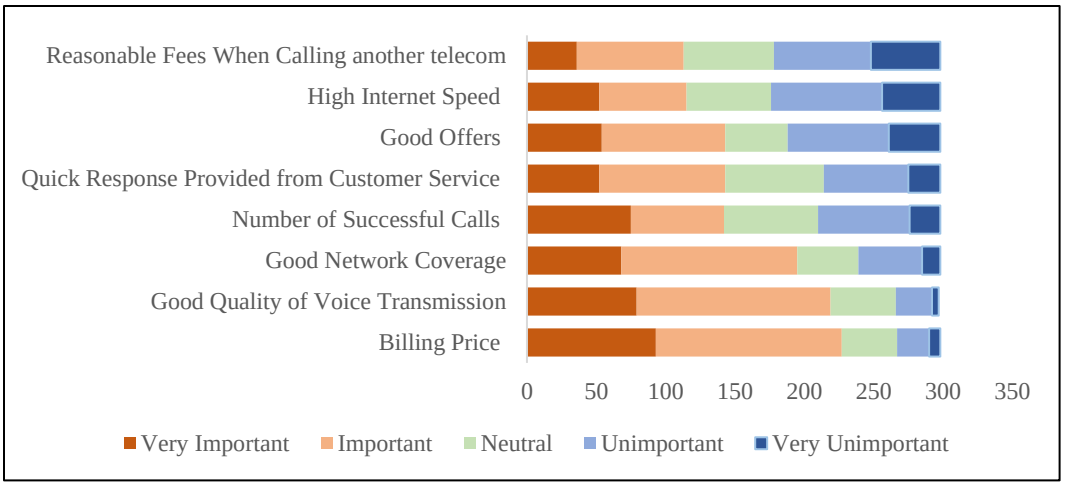

Figure 4.11: The importance of the metrics for STC customer satisfaction.

As shown in Figure 4.11, the 'good quality of voice transmission' metric was the most important metric for 140 STC customers, whereas customers allocated the 'billing price' metric the second-highest importance (134 customers). 'Billing price' was considered very important by 93 customers, whereas the 'reasonable fees when calling another telecom company' metric was considered very unimportant by 50 customers and unimportant by 70 customers. 'High internet speed' was considered very unimportant by 42 customers.

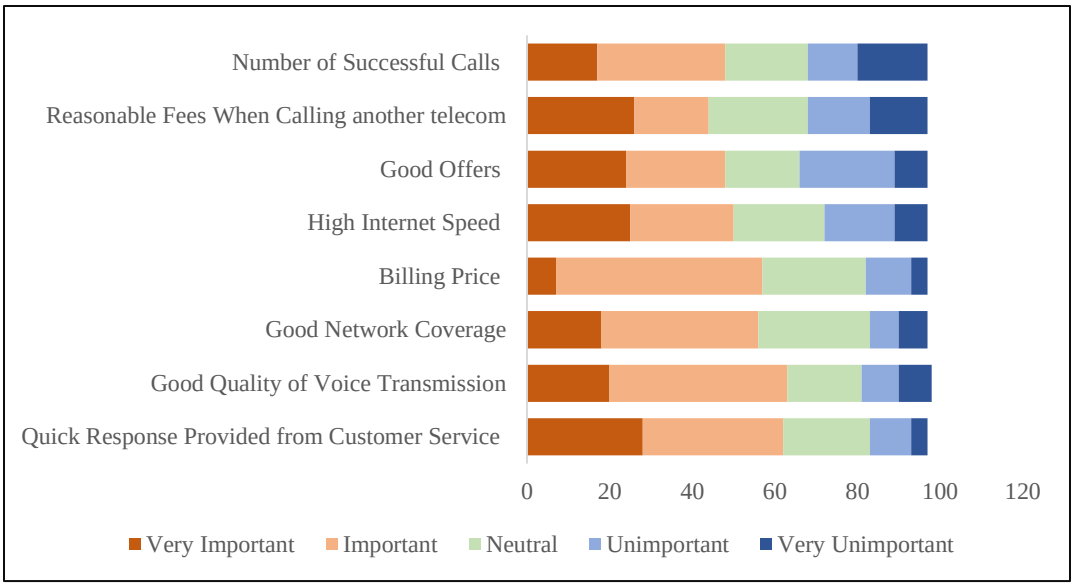

Figure 4.12: The importance of the metrics for Mobily customer satisfaction.

The highest response from Mobily customers (50 respondents) was for 'billing price' as an important metric (Figure 4.12), whereas the 'reasonable fees when calling another telecom company' metric received the lowest importance count (18 customers). The highest metric chosen as an unimportant metric was 'good offers' (23 respondents), whereas the lowest response, chosen by only seven Mobily customers, was for 'good network coverage' as an unimportant metric.

With the most responses from Zain customers (12 respondents), 'high internet speed' was chosen as a very important metric (Figure 4.13), followed by 'billing price' as important for 11 participants. 'Unimportant'

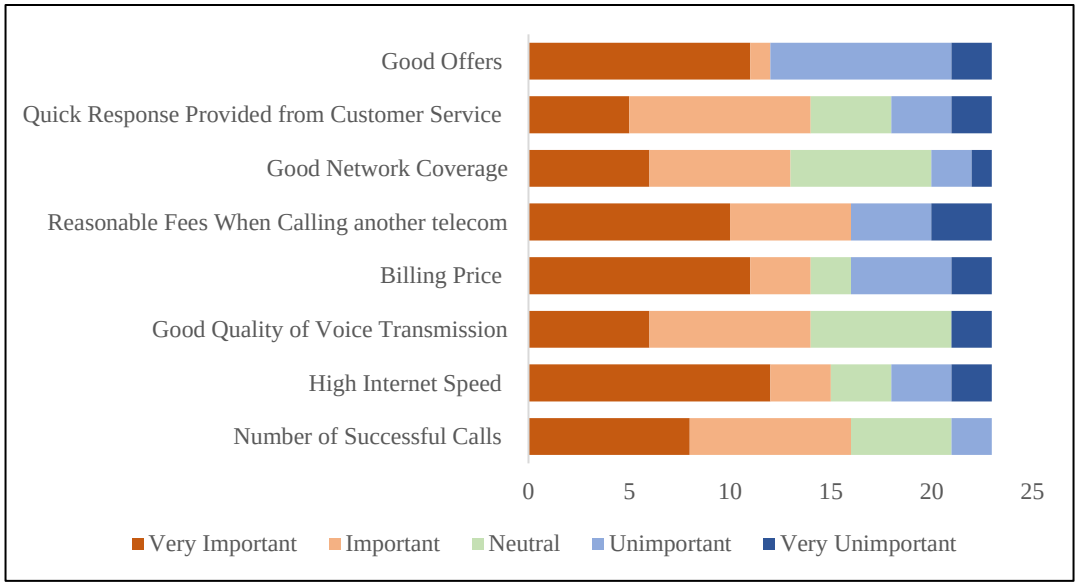

Figure 4.13: The importance of the metrics for Zain customer satisfaction.

was mostly attributed to 'good offers' (nine respondents), whereas 'good quality of voice transmission' received no responses as an unimportant metric.

As shown in Table 4.24, the *p*-value of all standards was < 0.05, indicating a statistically significant relation between 'type of telecom company' and each other metric except for 'network coverage', where the *p*-value was .184 > 0.05, which means that there was no correlation between those two variables. Furthermore, the *p*-value of 'quality of voice transmission' was .281 > 0.05, implying no correlation between the two variables. Moreover, the *p*-value of 'customer service' was .931 > 0.05, entailing no correlation between the two variables, and the *p*-value of 'number of successful calls' was .112 > 0.05, leading to no correlation between the two variables.

To rank the importance of the metrics for each telecom company from the customer's point of view, I calculated the RIIs as mentioned before and then listed them in ascending order from 1 to 8. Table 4-25 shows that the top three most important metrics for STC customers were 'billing price', 'good quality of voice transmission' and 'good network coverage', whereas the least important was 'reasonable fees when calling another telecom company'.

For Mobily customers, as shown in Table 4.26, the first three most important metrics were similar to those for STC customers for the second and third metrics: 'quick response provided from customer service', 'good

quality of voice transmission' and 'network coverage'. The least important metric was 'number of successful calls'. As shown in Table 4.27, the first and second most important metrics for Zain were 'number of successful calls' and 'high internet speed', whereas the least important was 'good offers'.

Metrics	Pearson Chi-Square	Asymp. Sig. (2-sided)
Network Coverage	11.324	.184
Quality of Voice Transmission	9.778	.281
Customer Service	3.055	.931
Number of Successful Calls	13.001	.112
Billing Price	17.806	.023
Reasonable Fees when Calling another Telecom Company	22.169	.005
Good Offers	19.760	.011
High Internet Speed	17.603	.024

Table 4.24: Chi-Square Test for the Importance of Customer Satisfaction Metrics per each Telecom Company.

Standard	RII	Rank
Billing Price	0.789	1
Good Quality of Voice Transmission	0.774	2
Good Network Coverage	0.728	3
Number of Successful Calls	0.672	4
Quick Response Provided from Customer Service	0.659	5
Good Offers	0.634	6
High Internet Speed	0.602	7
Reasonable Fees when Calling another Telecom Company	0.586	8

Table 4.25: Ranking the Importance of Customer Satisfaction Standards for STC Telecom Company.

Standard	RII	Rank
Quick Response Provided from Customer Service	0.748	1
Good Quality of Voice Transmission	0.726	2
Good Network Coverage	0.709	3
Billing Price	0.693	4
High Internet Speed	0.687	5
Good Offers	0.668	6
Reasonable Fees when Calling another Telecom Company	0.656	7
Number of Successful Calls	0.639	8

Table 4.26: Ranking the Importance of Customer Satisfaction Standards for Mobily Telecom Company.

Standard	RII	Rank
Number of Successful Calls	0.791	1
High Internet Speed	0.774	2
Good Quality of Voice Transmission	0.739	3
Billing Price	0.739	4
Reasonable Fees when Calling another Telecom Company	0.739	4
Good Network Coverage	0.730	6
Quick Response Provided from Customer Service	0.704	7
Good Offers	0.687	8

Table 4.27: Ranking the Importance of Customer Satisfaction Standards for Zain Telecom Company.

4.5 Discussion

This section discusses the results of our analysis of the responses provided by the participants. The results are then used to answer the research questions and indicate any correlation between the findings of previous studies and theories that were discussed earlier.

4.5.1 Questionnaire objectives

***Qo1:* To define customer satisfaction metrics from the customers' perspective.**

As mentioned before, I defined customer satisfaction metrics using a report by the Saudi Communications and Information Technology Commission [438], related research and the tweet annotation process. Then, I evaluated the importance of these metrics using statistical analysis for the responses obtained by the questionnaires from the customers' point of view. The metrics are 'network coverage', 'quality of voice transmission', 'customer service', 'successful calls', 'billing price, 'offers', 'reasonable fees when calling another telecom company' and 'internet speed'. I analysed the importance of each metric for each telecom company using RII, as shown in Figure 4.14.

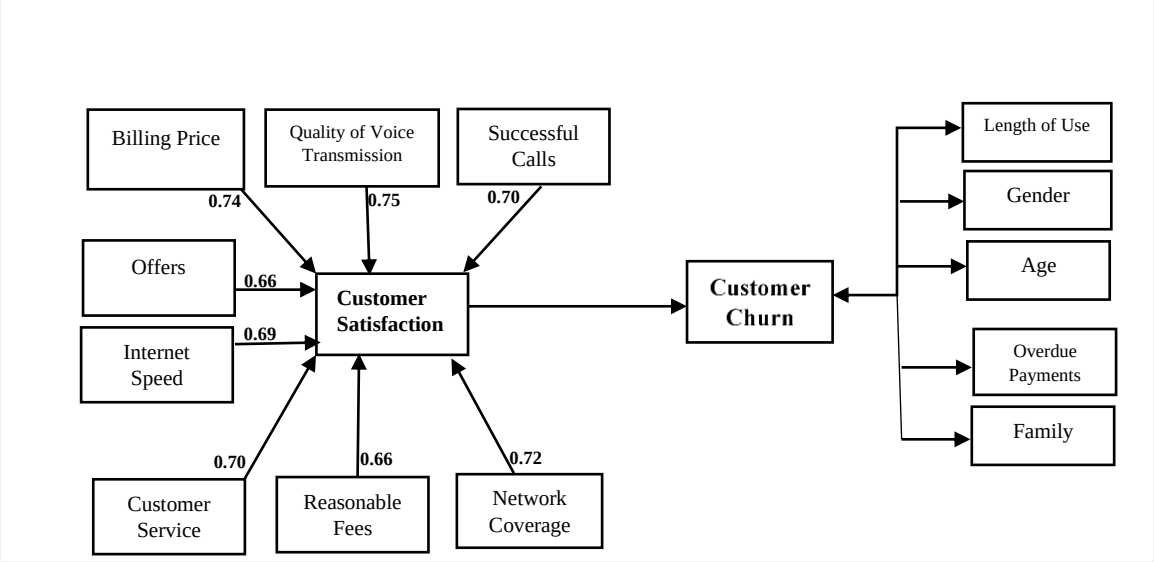

Figure 4.14: Taxonomy of the average importance of measurable metrics of customer satisfaction and their relationship with customer churn.

As shown in Figure 4.14, 'quality of voice transmission' is the most important metric from customers' point of view, followed by 'billing price'. Woo & Fock [440] argues that voice transmission quality is the most

important factor in driving customer satisfaction. However, interestingly, customers did not select 'quality of voice transmission' as a reason for churning, which proves that despite the quality of voice transmission is important for customer satisfaction, it is not a reason for churning.

Nevertheless, I found 'billing price' to be the second metric important from customers' point of views, which is compatible with findings from previous literature. Specifically, [439] found that price perceptions positively affect overall customer satisfaction. Furthermore, customers allocated this metric as the first reason for churning in the open question, which is in line with what found by [463] in his qualitative analysis – namely, that a high billing price is a crucial churn factor.

Network coverage resulted in being the third-important metric for customer satisfaction. This metric is considered a key reason explaining churn via the open question in my questionnaire. In the literature [464] it was also shown that a poor network is one of the key reasons for churning. The importance of network coverage was highlighted by internet users who faced poor network coverage, as also found by [106]. Both network coverage and the quality of voice transmission depended on the number of network towers geographically spread in Saudi Arabia for each telecom provider.

'Customer service' was found to come after 'network coverage' for customer satisfaction, and it was assessed as a crucial reason for churning following the open question. This is compatible with found by [464] – namely, that customer service is more impactful than the rest of the factors.

The least important metrics for customer satisfaction were 'reasonable fees when calling another telecom company' and 'good offers', which are related because telecom companies in Saudi Arabia offer a lot of packages at a reasonable price, especially when calling another telecom company in Saudi Arabia.

Regarding the different companies, the questionnaire analysis found that ''billing price' was the most important metric of customer satisfaction for STC customers. This result corresponds with that from the questionnaire participants in an open question for customer churn reasons, where high billing was the most frequently mentioned one (37%, 60 responses).

For Mobily, 'quick response provided from customer service' was the most important metric. The CITC report[16] regarding Mobily service quality indicators shows that 80% is the acceptable percentage for the customer service team to answer a customer call within 60 seconds. The company scored 74.67%, which indicates this service's failure rate. This result corresponds with the answers to the questionnaire's open question about customer churn reason, whereby 'slow response of customer service' was the most frequently mentioned churn reason (28.0%, 37 responses).

Zain customers ranked 'number of successful calls' first, followed by 'quality of voice transmission', likely due to the limited number of Zain network towers in Saudi Arabia, which affects the successful completion of calls. This problem was recently brought to the attention of Zain. An attempt was made to solve it by asking IHS Holding Limited – the largest cell tower operator in the European, Middle Eastern and African markets – to sell and lease back their towers to Zain[17]. This agreement should 'raise the efficiency of mobile networks'.

The least important metrics for STC customers were 'high internet speed', followed by 'reasonable fees when calling another telecom company. This is probably due to the fact that STC led the implementation of 3G and 4G technologies in Saudi Arabia and recently added 5G[18]. Furthermore, STC offers free calling to other telecom companies.

'Reasonable fees when calling another telecom company' was the least important metric for Mobily customers due to the availability of multi-offers provided by Mobily that include unlimited minutes to call the networks of other telecom companies in Saudi Arabia[19]. This result corresponds with the answer of the questionnaire participants to the open question about customer churn reasons, where 'unreasonable fees when calling someone who uses another telecom company' received one response, followed by 'number of successful calls' as the last metric in term of importance. This is explained by the fact that spread of Mobily cell towers in large areas in Saudi Arabia, as the CITC report found, entails that the geographical spread for

[16] https://www.citc.gov.sa/ar/indicators
[17] https://www.sa.zain.com/
[18] https://www.stc.com.sa/
[19] https://www.mobily.com.sa

radio coverage for Mobily is 99.4%[20]. In support of this result, the report issued from CITC about Mobily telecom service quality indicators denotes that the acceptable average for unsuccessful calls is < 2%, and the average of unsuccessful calls in Mobily company is 0.8%[21]. This result implies that the number of unsuccessful calls is not an issue for Mobily customers.

For Zain customers, 'good offers' was the least important standard because Zain company has different offers with different prices, from high to low.

***Qo2*: To understand churning causes and churner characteristics and behaviours.**

As shown in Figure 4.14, the factors that related to churning behaviour in this study are 'length of use', 'gender', 'age', 'having one family members using a customer previous telecommunication company', and 'having overdue payments'. This result is consistent with what found by [62] – namely, that both customers' age and tenure length affect churn probability.

The age and gender are key factors related to churning, as previous authors argue ([101], [114], [20], [115], [18], [116], [23], [62], [91],[116], [132], [94]). In the present study, the highest number of participants (14%) who had changed their telecommunication company belonged to the 35–44 age group, whereas the group with the lowest number of participants who had changed their telecom company (0%) was the 65 and over age group. This result is consistent with what found by [91], namely that young people below the age of 45 are more likely to churn. Furthermore, [106] found that customer age is the seventh most-important churn indicator and confirmed that young people are more likely to churn. This can be explained by the fact that young people tend to choose the most recent technology and the most up-to-date options when compared to customers of other ages. With gender, this study confirmed more females (54%) than males had not changed telecom company, whereas more males (10,3%) than males had done so.

Nevertheless, as the present study demonstrated, having a family member in the same telecom company did not have a relation with the customer churning behaviour, with 100 responses mentioning having one of their family members to still be using their previous telecommunication company. Which may be due to those

[20] https://www.citc.gov.sa/ar/indicators
[21] https://www.citc.gov.sa/ar/indicators

offers provided by telecom companies in Saudi Arabia that make calling different operators cheap or free. This result was the opposite of what found by [62],[24]. Hung et al. [62] confirmed that customers who did not often make phone calls to others within the same operator's mobile network were more likely to churn.

Regarding the length of use, I demonstrated that the highest percentage of those who left their previous telecom company was in the 1-5 years group (39.6%). Other studies used contract length as a churn predictor [20], [465], [94],[466], [62]. Balasubramanian and Selvarani [94] and [62] concluded that a customer with contract length between 25-30 months are more likely to churn.

Many studies associated 'contract length' and 'overdue bill' as potential churn predictors. Balasubramanian and Selvarani [94] found that a customer with a contract length between 25–30 months and with less than four overdue payments are less likely to churn. Furthermore, [62] concluded that churning is less likely for a customer with a contract length between 25–30 months and with less than four overdue payments within six months. In this study, I found that most respondents had overdue payments before leaving their previous telecom company (58 customers, or 46%).

Qo3: **To define the differences between the telecom companies.**

Regarding the correlation between 'type of previous telecommunication company' and 'having overdue payments', the results showed that there is no correlation between the two variables. Most customers with overdue payments before leaving a telecom company were from Mobility (27), whereas 23 were from STC and 21 from Zain. Hence, having outstanding payments is an early indicator of churning. Accordingly, previous literature [116] considers unpaid balances as a churning indicator.

However, I proved that there is a correlation between 'type of previous telecommunication company' and 'length of using the previous telecom company'; 41.4% STC customers who left it had used it for more than 10 years. Regarding Mobily customers, 42.1% of those who left it had used it for 1–5 years. Among Zain customers, 44,4% among those who left it had used it for 1–5 years. Customers' longer use of STC is due to the dominance of the STC company in the Saudi market for a long time, until the other competitors entered it, which suddenly gave customers more choice.

Concerning the relation between 'type of previous telecommunication company' and 'having customer's family member using the previous telecommunication company', the results proved that there is no correlation. There are 85.4% of the STC customers who had family members who used [the] customers' previous telecommunication company. The same applies to Mobily and Zain customers, with 90.8% and 68.9% of them, respectively, having one of their family members using their previous telecommunication company. These results explain the reason mentioned earlier.

As regards the correlation between 'telecom company type' and 'communication method [*used to complain or make suggestions*]', Table 4.20 show that there is a correlation between the two variables. 130 STC customers used Twitter as a communication method, 69 the company application, 62 the company website, and 37 the telephone. These results reflect the Saudi society as a whole, in the way it is using communication methods, as well as it reflects the impact of the internet and social media on Saudi lives. STC realised the impact on Saudi people of social media. In particular, Twitter supports different accounts that have been created to care for customers and specific businesses, some of these accounts, are @stccare_ksa which oriented for customer care 24 hours, @InspireU_STC, @dawristc, @stclive, @specialized_stc, @stc_ksa, @stcpay_ksa, @Qulitynet_q8, and @stc. Similarly, STC provided many applications specific to caring for customers, focussed on specific tasks, such as STC Pay and My Stc.

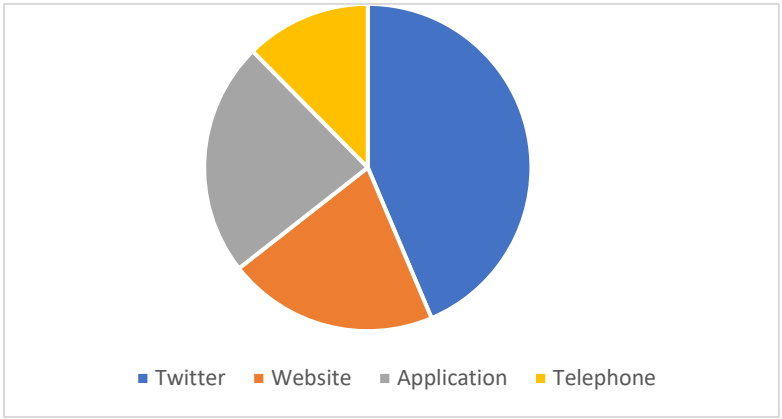

Figure 4.15: Communication methods in STC and their proportions.

However, the situation with Mobily customers is the opposite of that of STC customers, As shown in Figure 4.16, most of them used the company application as their primary communication method (39 participants), followed by Twitter as their second preferred communication method (25 participants). Mobily customers

preferred the company applications, as the first communication method and Twitter as the second communication method. This is probably due to the fact that Mobility provides many applications to serve customers, whereas they have only one Twitter account that does not provide customer care 24 hours. Furthermore, as I explained before in Table 4-26, the main metric for Mobily customer satisfaction is customer service, which is the first reason for churning, as resulted from the open question.

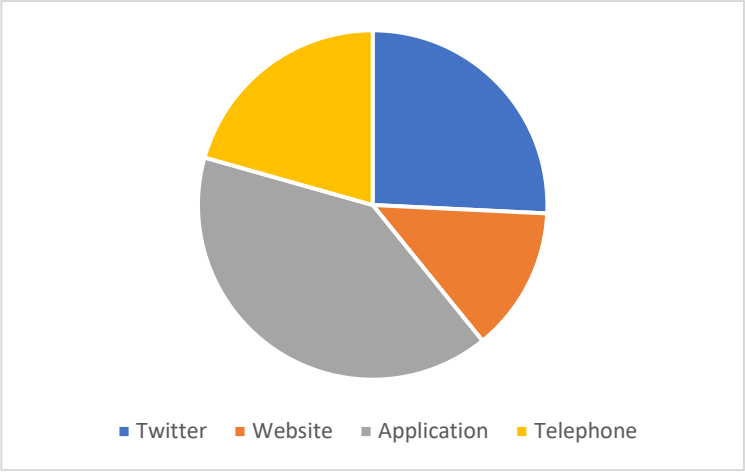

Figure 4.16: Communication methods in Mobility and their proportions.

Zain customers used Twitter as their primary communication method because the company provides one Twitter account for customer care for 24 hours. In contrast, 'using the telephone' received the lowest count for Zain customers with two participants, as can be seen in Figure 4.17.

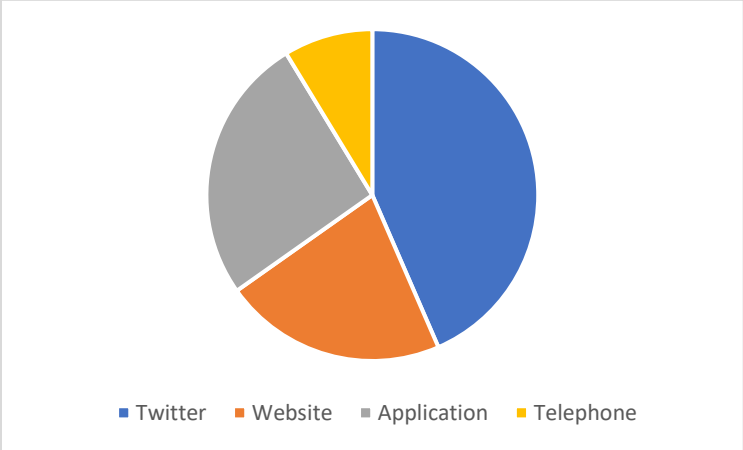

Figure 4.17: Communication methods in STC and their proportions.

Regarding the correlation between 'telecom company type' and 'probability of enhancing service quality after using Twitter as a communication method', the results denoted that there is no correlation between these two variables. The highest percentage, 59.1%, was for STC customers who thought that service quality was enhanced as a result of communicating with the telecom company through Twitter, which confirmed what I found out about Twitter being the first communication method used by STC customers. As regards Mobility customers, 50.5% of them found that using Twitter did not enhance service quality, which is compatible with what I mentioned early. In contrast, 69.6% of Zain customers remarked that service quality was enhanced by the opportunity for customer communication with the telecom company through Twitter.

4.6 Limitations

With 298 STC customers, 97 Mobily customers and 23 Zain customers, the collected data is unbalanced. The sample should reflect the relative population, in this study, STC customers in Saudi Arabia are more than Mobily customers, and Mobily customers more than Zain customers.

4.7 Summary

This chapter illustrates the creation of a metrics suite to assess customer satisfaction. The suite's implementation started with a literature review on metrics and data gathering from the Saudi Communications and Information Technology Commission [438], followed by a tweet annotation process.

The resulting metrics were then analysed using a questionnaire to investigate churn causes, characteristics and behaviours.

These metrics will be used in Chapter 6 to predict customer satisfaction percentage towards the services provided by the Saudi telecom companies and suggest recommendations for them. The collected churn characteristics and behaviours will be used in Chapter 7 to predict customer churn percentage. Thus, this chapter answers *RQ1* – **'**What are the traceable, measurable criteria for customer satisfaction with telecom companies in Saudi Arabia?' Furthermore, it contributes to achieving *RO2* – To propose recommendations to improve the services of Saudi telecom companies.

This chapter's results can potentially provide insight for decision-makers in telecom companies in Saudi Arabia to enhance their services. Furthermore, telecom companies may be able to avoid customer churn by focusing on what has been found in this study regarding customer churn reasons.

Chapter 5: Binary Classification Experiments

5.1 Introduction

Customer satisfaction is closely related to customer churn [11, 13, 14]. Twitter mining can be used as a tool to evaluate customer satisfaction. The objective of this chapter is to use Twitter mining to predict Saudi telecom customer satisfaction. This chapter answers the *RQ3* and achieves the *RO3*: to identify, based on Twitter mining, Saudi telecom companies' customers' satisfaction. I thus compare, vary and enhance several baseline and cutting-edge approaches. Starting with machine learning algorithms, with best-in-class SVM, which were used both as a baseline and enhanced with Twitter features. I used the statistical evaluation of the feature selection method. Contrary to the initial expectations in this study, although there are a significant number of prayers in the Arabic tweet corpus, the *Has-Prayer* feature had to be removed from the feature set, possibly due to both positive and negative tweets that use prayers, often including the word 'God'. Additionally, several cutting-edge techniques, i.e., two deep learning models, LSTM and GRU, with different embeddings and settings and three transformer networks, AraBERT, hULMonA and RoBERTa models were compared. The result is a new model for Arabic Sentiment Analysis in the telecommunication field. The proposed model combining the AraBERT model and Bi-GRU was demonstrated to predict customer satisfaction based on tweets successfully. The results showed that the new model is highly accurate when compared to other models. The results were triangulated with the actual result for validation. The result will be used as an input variable on the churn prediction model (Chapter 7).

5.2 Feature Engineering

5.2.1 Feature selection

Valuable feature engineering is essential in the learning process to improve accuracy [45]. Feature engineering is the primary and most complicated task [313]. One of the crucial steps in feature engineering is an appropriate feature selection. Their utility in text analysis must be examined. Feature selection entails choosing the feature subset that achieves superior performance in a classification [467]. Feature selection aims to select the minimum number of features, reduce redundancy, and increase the classification label significance [467].

Several feature selection techniques were identified in the literature: *best-subset selection*, and *forward and backwards stepwise selection* [468]. Best-subset selection is used to search for all features and improve classification accuracy. Despite being considered as the best technique [45], it is an exacting and time-consuming process, especially when there are many features. With the forward stepwise selection, model performance is observed after adding features one by one. A forward selection has its drawbacks, including the fact that each addition of a new variable may render one or more of the already-included variables non-significant. Meanwhile, the backwards stepwise selection avoids this by beginning the classification with all possible features and then eliminating one feature after another while observing the effect of each feature in terms of classification performance. Therefore, in this study, I used backwards stepwise selection for the reasons mentioned before, following [45].

A feature set could include (but is not limited to) any of the following: *syntactic*, *semantic*, *stylistic* and *genre-specific* features [425], [428]. As [249] mentioned, some constraints for selecting specific features are effortless to extract – they are uncomplicated but relevant and axiomatic for a classifier to train.

In this study the feature extractions were based on ASA literature, that is, [45], [249], [425], [428] as follows.

Term features are represented as a term-weight vector considering every term in a document as a vector. There are three available weighting schemes for term-weight vector: term frequency (TF), term presence and term frequency–inverse document frequency (TF-IDF) [469].

These features are common term features used in sentiment analysis [45]. Term presence checks the existence of a term in a document giving a term weight 1 for existence and 0 for non-existence. In contrast, the term frequency count of a term's frequency within a document and TF-IDF is the percentage of each word's frequency based on all records' frequencies.

Syntactic features are the most common features used for SA [428]. They include n-gram features [152]. N-grams are a series of the terms in a text. Unigram's process one term at the time, while bigram two terms at the time, and trigram three terms. Many studies proved that n-gram features enhance the performance of a classifier [45], [203], [425], [184].

Semantic features include using a sentiment lexicon to classify the text. This was used to annotate tweets based on positive and negative words corresponding to a sentiment lexicon. In this study, the manually built lexicon AraSTw has been used. Many studies proved that using a sentiment lexicon as a semantic feature enhances SA [430], [425], [45]. I applied the AraSTw lexicon as four features: *Has-positive word, Has-negative word, Positive Word Count, Negative Word Count.* Has-positive word and Has-negative word check the word in a tweet if it is positive/ negative in AraSTw lexicon. Positive Word Count and Negative Word Count define the number of positive words and negative words in a tweet.

Morphological features have been proposed by previous ASA studies [199], [425]. Al-Twairesh [45], which experimented with using Part-Of-Speech (POS) tags as a feature, reported that it did not enhance classification, because of the lack of POS taggers for social media. Other morphological features used in ASA include aspect, gender, mood, person, state and voice. However, [425] proved that these features drop the classifier performance by 21%. Accordingly, I did not use the morphological features in our classification models.

Language-style features are a feature set that represents the social media dialect.

- Is-Sarcastic feature: this feature is assigned via manual corpus annotation process by annotators. In this study, I asked the annotators (Chapter 3) optionally to assign true or false for Is-Sarcastic for those tweets that they think are sarcastic, following [425].

- Stylistic features: this feature set refers to the number of informal sentiment indicators on social media and some quantitative features, such as tweet-length. Refaee [425] used tweet-length features and found that this feature is beneficial to an Arabic SA. In addition, [428] stated that this feature set helps SA of Arabic forums, due to the language's rich stylistic variation nature.

Additionally, some features were selected, based on the nature of the corpus, as described below.

Affective cues features demonstrate whether there are some signals in a text; these signals reflect the user's culture and express a sentiment. The motivation for using this feature set was finding a set of simple features that can correlate with the users' culture and, at the same time, can be utilised as a means of conveying sentiments. There were many examples of *du'a'* (supplication/prayer) in the AraCust corpus. The most frequent bigrams in the AraCust corpus were a prayer, as I mentioned in Chapter 3, Table 5.1. Mourad and Darwish [430] found many Quranic verses in their corpus that expressed the writers' sentiment [425]. Therefore, the Has-Prayer feature is used to check whether a tweet has a supplication or not and the sentiment involved with that supplication following [425].

حسبي الله
Requesting Allah for a suffering

Table 5.1: An example of an affective cue feature in our corpus.

The **Tweet-topic** feature evaluates the role of the SA topic. Abdul-Mageed et al. [199] reported that this feature is beneficial for ASA. They asked annotators to manually specify one of the topics that represent a sentence in Arabic news. In addition, [425] asked the annotators to choose one topic from the following: sports, economy, politics, social/religious, Internet and other. I also processed tweets, and due to the ASA recommendations, I adopted to use this feature. In this study, I asked the annotators to select from a set of predetermined topics (Saudi telecommunication services) discussed in Chapter 4, which are Successful Calls, Internet Speed, Quality of Voice Transmission, Billing Price, Reasonable Fees when Calling another Telecom Company, Network Coverage, Customer Service and Good Offers. In this section, I present the telecommunication company services that customers considered important, and that were identified in

customer tweets to measure their customer satisfaction, following [89]. The summary of the feature sets used in this study are shown in Table 5.2 besides the term features.

Overall, in this study, the feature extractions were based on ASA literature, that is, [45], [249], [425], [428].

Feature	Value
N-gram	series of the terms
Has-Prayer	True/false
Is-Sarcastic	True/false
Tweet length	Numeric
Tweet-topic	Nominal
Has-Positive Word	True/false
Has-Negative Word	True/false
Positive Word Count	Numeric
Negative Word Count	Numeric

Table 5.2: Summary of the feature sets used in this study.

5.3 Model of Sentiment Classification

There are many sentiment classifications levels, binary and multi-way sentiment classification, conducted through a rating scheme (i.e., a 4-star rating scheme) [470]. One of the most popular one is the binary classification of sentiments, into positive vs negative. In this study, I used binary classification of sentiments into positive vs negative. Each sentiment label is indicative of customer satisfaction, Satisfied vs. Unsatisfied.

5.4 Performance Evaluation

5.4.1 Evaluation Metrics

To evaluate the performance of the model, I used four metrics suitable for binary classification F1, Accuracy (Ac) [471], Precision (Pr), and Recall (Rc). These metrics use the True Negative (TN), True Positive (TP), False Negative (FN) and False Positive (FP) [472] as follows:

- TP: the number of correctly classified positive instances.
- TN: the number of correctly classified negative instances.

- FN: the number of positive instances misclassified as negative instances.
- FP: the number of negative instances misclassified as positive instances.

The confusion matrix [473] is a tool used with binary classification; it compares the actual positive and negative and the predicted positive and negative. It uses TN, TP, FN and FP (Table 5.3).

Predictions	**Actual**	
	Satisfied	Unsatisfied
Satisfied	TP	FP
Unsatisfied	FN	TN

Table 5.3: Confusion matrix used in this study.

The micro average is suitable for binary-classes, especially if the classes are imbalanced. It totals all classes' contribution to the average metric calculation [474] and it aggregates the precision and recall of the classes. Accuracy (Acc.) is defined as the ratio of correctly classified instances, calculated as follows:

$$\frac{TP+TN}{TP+TN+FN+FP} \quad (5.1)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (5.2)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (5.3)$$

where F1 is the harmonic average among precision and recall, which is calculated as follows:

$$2 * \frac{Precision*Recall}{Precision+Recall} \quad (5.4)$$

Here, I use the four metrics for each class. In addition, I use the average of F1 (F-avg) as an indicator for the performance of the model overall, which is considered a better measure than accuracy, especially with imbalanced data [475], [476]. For a comprehensive view upon our results, I use all these metrics.

5.4.2 Evaluation Methods

Some of the successful methods for evaluating a model proposed by the literature were k-Fold Cross-Validation (CV) [436] and an independent test set.

- K-Fold CV is a widely used classification task [436], especially with ASA studies [430], [332], for the reasons given by [436]: it's simple to apply, and it avoids the bias. The k-Fold CV has a specific parameter k that denotes to equal, fixed numbers as folds. A classifier splits the dataset to a training set and test set. Based on k, a classifier splits the dataset in k folds, then each fold, in its turn, is used in the test set. In addition, in each classification process, there are different combinations of the training set. In the end, the average error is calculated as the overall score. Ten folds have been used for the dataset to obtain the best estimate of errors, as proposed by [436].

- The independent test set is used to assess the model's ability to predict in a dynamic medium like Twitter [252]. Refaee [425] and [45] are two of the ASA studies that used an independent test set as an evaluation method.

In this study, due to its flexibility and popularity, I used k-Fold CV on the AraCust to train the model, tuning the parameters until I found the most accurate predictive model.

5.5 Machine Learning Schemes

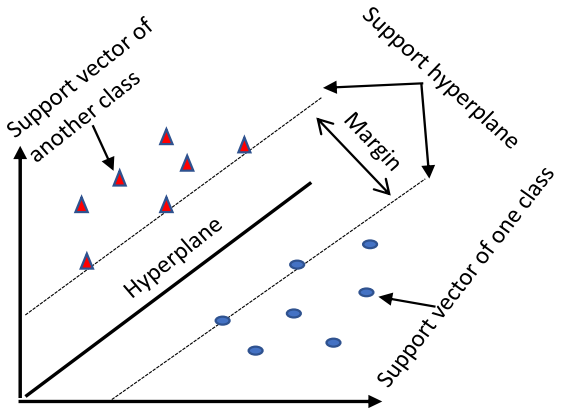

Figure 5.1: SVM architecture.

5.5.1 Baseline

First, I created the baseline to compare the model to it. The baseline includes the basic features: term features and n-gram models, following [45]. The term features include term presence, TF and TF–IDF (Table 5.4). Some studies stated that the n-gram model enhanced the classifier's accuracy and they used it as a baseline for the SA classification of tweets [39], [425], [480, 481].

I evaluated n-gram models (unigram, unigram+bigram and unigram+bigram+trigram) and term presence (feature) models to create the best model (Table 5-5). The results showed that the term presence model achieved the best F-avg, with the SVM classifier, for the three corpora Mobily, STC and Zain. This was in line with the results found by [198] and [425], which is that term presence is the best feature with binary classification due to a lack of term repetition within a short text, such as a tweet. Pang & Lee [247] noted that using term presence leads to a better performance than using TF.

Regarding the n-gram models, I found the combination of the unigram and bigram models to be the best for the STC, Mobily and Zain corpora. The result is consistent with what was found in the literature regarding the superiority of combining the unigram and bigram models over the n-gram model in both ASA [203], [481], [430], [425] and English SA [155], [184].

The rationale behind combining the unigram and bigram models is to provide more information than the unigram model alone could do, and it is less sparse [430], [155]. The baseline model for all three corpora can found in Table 5.6.

Term models/ Corpus	TF	TF-IDF	Term presence
STC	0.761	0.755	0.771
Mobily	0.749	0.750	0.815
Zain	0.701	0.767	0.855

Table 5.4: F-avg for the term models using SVM.

Corpus	Features	F-avg using SVM
STC	Term presence + unigram and bigram models	0.773
Mobily	Term presence + unigram and bigram models	0.807
Zain	Term presence + unigram and bigram models	0.853

Table 5.5: Baseline for the Three Corpora Using SVM.

Gram models/ Corpus	Unigram	Unigram+ Bigram	Unigram +Bigram +Trigram
STC	0.635	0.770	0.756
Mobily	0.746	0.799	0.764
Zain	0.799	0.850	0.729

Table 5.6: F-avg for the n-gram models using SVM.

5.5.1.1 Evaluation feature selection process

There are two attribute selection techniques common in the text classification task: Information Gain (IG) [482] and Chi-squared [437]. In this study, I used them to assess the feature selection method chosen before, following [425] and [45]. Yang and Pedersen [482] defined that the IG for a specific feature is the amount of information in the appearance or absence of this feature. Table 5.7 shows the IG for each feature for STC corpus; the features are ascendingly ordered by their IG.

Feature	**IG**
Tweet-topic	0.483
Is-Sarcastic	0.378
Has-Negative Word	0.227
Has-Positive Word	0.125
Negative Word Count	0.082
Positive Word Count	0.061
Has-Prayer	0.048
Tweet length	0.044

Table 5.7: IG for each feature.

Regarding the Chi-Square, [437] defined it as a statistical analysis that calculates the feature's independence from the class. Table 5.8 shows the Chi-square for each feature for STC corpus. After that, the features are in descending order of their Chi-square result.

Feature	**IG**
Tweet-topic	5526.24
Is-Sarcastic	4413.42
Has-Negative Word	2557.22
Has-Positive Word	1730.72
Negative Word Count	1520.85
Positive Word Count	797.96
Has-Prayer	694.55
Tweet length	620.87

Table 5.8: Chi-Square for each feature for the STC corpus.

I found that the IG and Chi-square results on STC, Mobily and Zain corpora are similar, so the results reported here are STC. I noticed from Table 5.7 and Table 5.8 that the IG and Chi-square techniques generated the same results; this result reveals the features' reliability.

5.6 Binary classification Experiments

5.6.1 Using SVM

As said, in Subsection 5.2.1 I used a back-stepwise selection to choose the best features with the SVM model on the AraCust corpus. I experimented with subsets of the initial feature set (Table 5.2). First, I calculated F-avg and the accuracy of the model with all features, then each feature was removed from the feature set one by one, with the remaining ones depicted in Table 5.9. If removing a feature decreased the classifier's performance, I kept it in the classification model. If the classifier's performance increased after removing a feature, that was interpreted as this feature was harming the classifier. If this was the case, I eliminated this feature from the feature set (see Figures 5.2, 5.3 and 5.4).

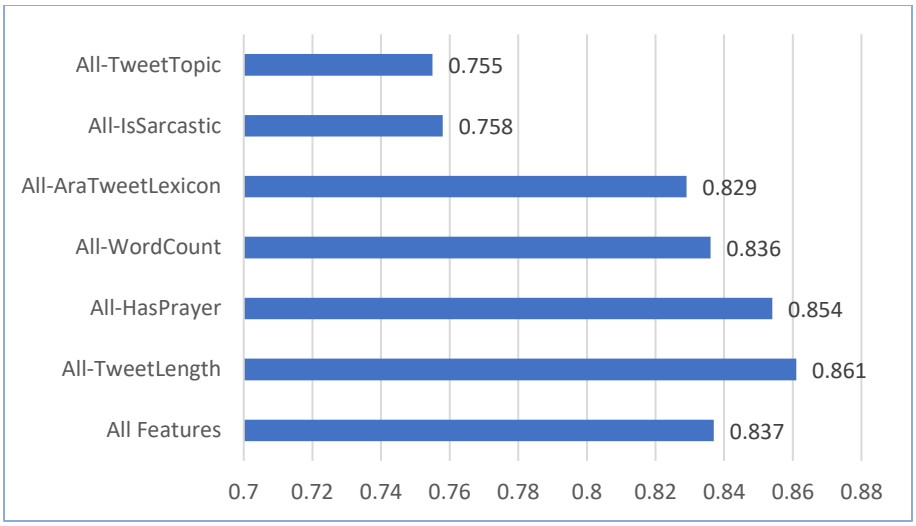

Figure5.2: The F-avg of all features used in the STC corpus and the F-avg when removing each feature.

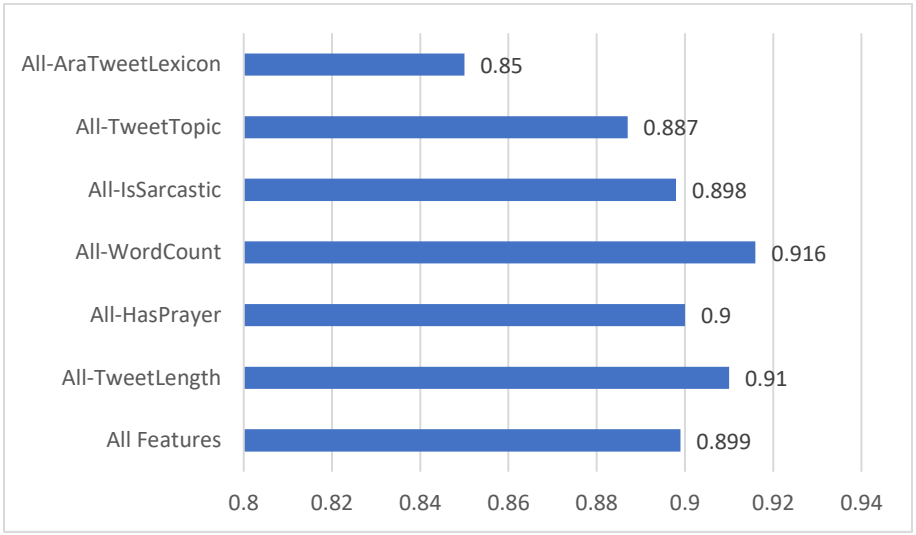

Figure 5.3: The F-avg of all features used in the Mobily corpus and the F-avg when removing each feature.

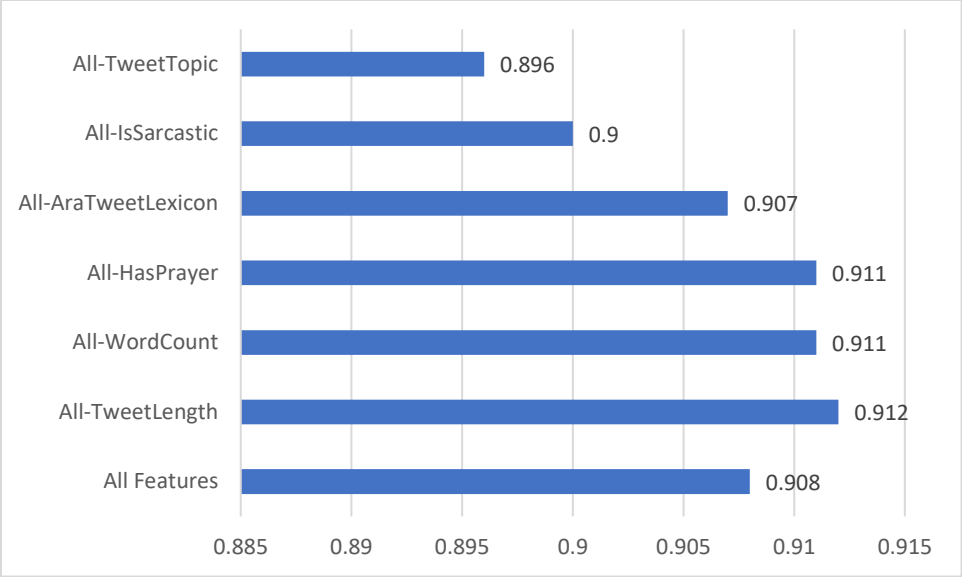

Figure 5.4: The F-avg of all features used in the Zain corpus and the F-avg when removing each feature.

5.6.1.1 Discussion

As shown above in Figure 5.2, the classifier's performance is increased when the *Tweet-Length* and *Has-Prayer* features are removed. This means that these features harmed the classifier by 11.3% and 10.4% respectively; thus, they were removed from the feature set. Meanwhile, removing the *Tweet-Topic* feature caused a higher performance drop than other features, of 2.3%. This means that this feature is essential in the features set. This is the case when removing the *Is-Sarcastic* and '*Positive Word Count* and *Negative Word Count* and '*Has-Negative Word* and *Has-Positive Word'*, so these features should also be retained. When the classifier with the remaining features (Table 5.9) was applied on AraCust, the F-avg was 0.905, which means the model performance increased than the baseline by 18.0%. It should be noted here that '*Positive Word Count* and *Negative Word Count*' and '*Has-Negative Word* and *Has-Positive Word'* features are using AraSTw lexicon.

Regarding the Mobily corpus (Figure 5.3), removing Tweet-Length, Has-Prayer and Word Count features increased the classifier's performance slightly, which means that these features harmed the model. Meanwhile, removing the other features decreased the classifier performance somewhat, so they were

retained (Table 5.9). An essential feature in feature sets is the AraSTw lexicon. After running the classifier with the remaining features, the F-avg obtained was 0.920. That is an increase over the baseline of 15.4%.

Regarding the third corpus, Zain (Figure 5.4), removing the Tweet-Length, Word Count and Has-Prayer features caused the slight increase in the model's performance. Therefore, these features were removed from the features set, because they harmed the model. Removing the Tweet-Topic feature caused a large decrease in the model performance, meaning that it is the most crucial feature. After running the classifier with the remaining features (Table 5.9), the performance increased with F-avg 0.935.

Corpus	Eliminated Features	Remaining Features	Favg	Baseline
STC	Has-Prayer. Tweet-Length	Tweet-Topic Is-Sarcastic. AraSTw Lexicon (Has-Negative Word, Has-Positive Word) Word Count (Positive Word Count, Negative Word Count)	0.900	0.773
Mobily	Tweet-Length Has-Prayer Word Count	AraSTw Lexicon Is-Sarcastic Tweet-Topic	0.920	0.807
Zain	Tweet-Length Word Count and Has-Prayer	Tweet-Topic AraSTw Lexicon Is-Sarcastic.	0.935	0.853
Average			0.918	0.811

Table 5.9: The features set of each corpus and the F-avg of the remaining Feature sets.

Table 5.9 shows the impact of the chosen feature set on the classification model. I noted that the classifier's performance was enhanced after adding the feature set; this confirms the significance of these features. The impact of the Tweet-Topic feature was the essential feature for STC and Zain corpora. The Tweet-Length feature, which was the only tweet-specific feature used, harms the corpora's classification performance. The classifier's performance on the test set decreased, consistent with the previous work by [425]. This result is due to the changing nature of Twitter over time [247]. The result of 'Has-Prayer' feature is somewhat

surprising, as it is a specific characteristic of Arabic Tweets. I think the result is due to the classifier being unable to distinguish between negative and positive tweets that used prayer because both types of tweets contain the same word "الله", which means God.

5.6.2 Using LSTM and GRU

The advantages of the deep learning model are:

- using an uncomplicated model to achieve complicated functions [314].
- using a massive volume of data (Big Data) effectively.
- dealing with the variety of data formats since it used abstract data.
- reduced demand for features extraction.
- deep learning model can learn complex features [313].

5.6.2.1 Experiment Settings

For reasons that were mentioned previously, the two most popular deep learning-based models, LSTM and GRU, have been used in this study with two different implementations: *simple LSTM and GRU* and *bidirectional LSTM and GRU* and with different parameters. This is in order to define the best model suitable to ASA and the nature of our AraCust corpus.

Model	Classification method	Embedding	Embedding size	Activation	Drop out	Optimizers	Epochs	Dense
M1	LSTM	Character	300	Sigmoid	0.5	Adam	50	2
M 2	LSTM	Word2Vec	300	Sigmoid	0.5	Adam	50	2
M 3	Bi LSTM	Character	300	Sigmoid	0.5	Adam	50	2
M 4	Bi LSTM	Word2Vec	300	Sigmoid	0.5	Adam	50	2
M 5	GRU	Character	300	Sigmoid	0.5	Adam	50	2
M 6	GRU	Word2Vec	300	Sigmoid	0.5	Adam	50	2
M 7	Bi GRU	Character	300	Sigmoid	0.5	Adam	50	2
M 8	Bi GRU	Word2Vec	300	Sigmoid	0.5	Adam	50	2
M 9	Bi LSTM	Character+Word2Vec	300	Sigmoid	0.5	Adam	50	2
M 10	Bi GRU	Character+Word2Vec	300	Sigmoid	0.5	Adam	50	2

Table 5.10: Different settings for the different models using LSTM and GRU.

Experiments were carried out with different settings and models to choose the best ASA model and the AraCust corpus (Table 5.10). Keras [483] was used for utilising deep learning models. In addition, TensorFlow [484] an open-source library, was used in a GPU environment. Two embeddings were utilised to obtain the features: character-level and Word2Vec. In the Word2Vec, the features were obtained using word representation to expose the connections between the tweets' words. On the other hand, the character-level was used to show how the sentiment affects the different characters in the tweets.

The models started with word embeddings, to represent each word in a tweet as a 300-dimensional word vector. It was then fed into the LSTM/GRU layer with this embedding, using a 128-dimensional hidden state. To avoid the model overfitting through training dropout [485]. Then the output was fed into another LSTM/GRU layer with a 128-dimensional hidden state that returns a single hidden state (Figure 5.5). Different experiments were done on 20, 40, 50, 70 and 100 epochs. The best performance was accomplished at 50 epochs. Therefore, all the experiments were conducted with 50 repetitions. The sigmoid layer was used for the classification. I applied a dense layer with two units for the two possible classes, followed by the Sigmoid activation. In addition, I used backpropagation in a default implementation bundle with the

TensorFlow library. For optimisation of the weight, Adam [486] was used, because it was shown to be efficient in computation.

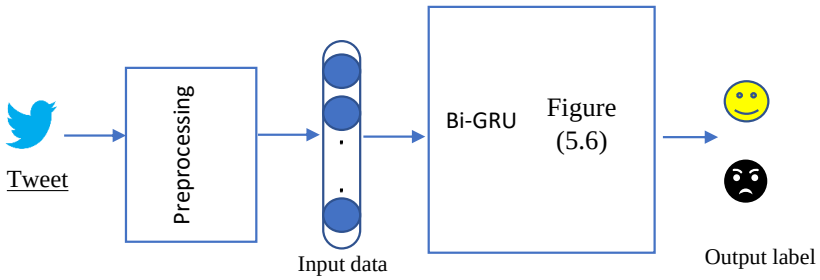

Figure 5.5: Architecture of the proposed deep learning model.

In the bidirectional LSTM or GRU (Figure 5.6), the model considers the future context of the text and the past context using joining forward and backward hidden layers [487]. I added the Keras library's attention mechanism with a context/query vector for temporal data to handle the long sequence on top of a recurrent neural network layer (LSTM or GRU) with *return_sequences=True, (1)*. The dimensions are inferred based on the output shape of the RNN.

model. Add (GRU (64, return sequ ences=True))

model. Add (AttentionWithContext ())

I used a context vector to assist the attention.

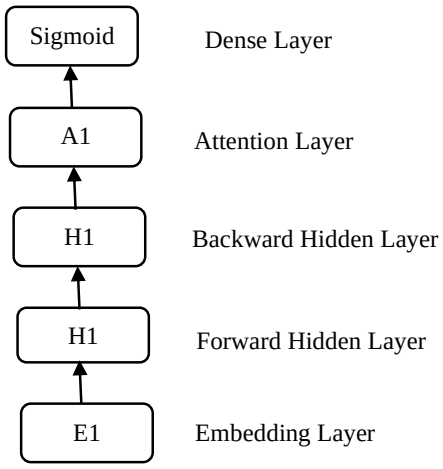

Figure 5.6: Bi-GRU/LSTM Architecture.

5.6.2.2 Results and Discussion

Figure 5.7 shows that Bi-GRU with Word2Vec (M8) performed better than all other models, at 95.16% accuracy. LSTM with character embedding achieved lower accuracy than the other models, at 94.30% accuracy. Comparing the results of the best of the two deep learning models with the baseline from the last section, SVM (Table 5.11), deep learning models have shown superiority, because of the applicability of deep learning approaches to the continuously dynamic nature of Twitter. GRU has also performed well, with a 95.16% accuracy, and better than LSTM. This may be due to GRU being less complicated than the LSTM model, which leads to it being widely used [313]. Finally, adding the bi-directional model attention mechanism enhanced the model's performance, while adding Word2Vec and the character level processing decreased the model's performance (see Fig. 7, where all models M1-M10 from table 10 are compared).

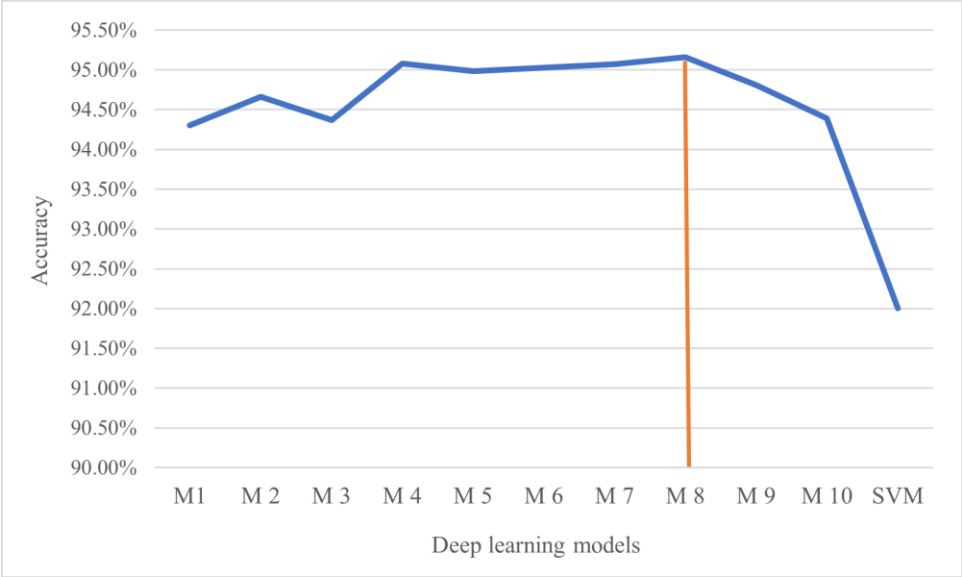

Figure 5.7: Comparing between the accuracy of deep learning models (M1-M10) with different parameters and SVM.

Model	Accuracy	F1	Recall	Precision
Bi-LSTM with Word2Vec	95.08%	0.951	0.951	0.951
Bi-GRU with Word2Vec	95.16%	0.952	0.952	0.952
SVM	93.0%	0.918	0.918	0.918

Table 5.11: M4 and M8 comparison with SVM baseline model.

I examined the statistical significance between the two models (Bi-LSTM with Word2Vec and Bi-GRU with Word2Vec) to check the differences between the two models by chance or by model skills. To check that the two models have a similar or different proportion of errors on the test set.

I used McNemar's test [488] following [489] to check the statistical significance. I found that the p-value is $0.01 < 0.05$, which means reject H0 and accept the H1.

H0: There are no differences between the models' skill.

H1: There are significant differences between the models' skill.

5.6.2.3 Comparison and Implications

After applying the M8 model to relevant data from the famous NLP competition, SemEval – specifically, the Arabic data set provided by the SemEval 2017 Task 4, Subtask A: Tweet classification according to a three-point scale of Twitter [170], the proposed model achieved 79.7% in terms of accuracy. Comparing the proposed model's result to the 58.1% results achieved by the NileTMRG team [411], which was placed first amongst the other top ten teams in Subtask A, the proposed model achieved clearly a much higher accuracy. This is promising progress in terms of ASA on Tweets.

5.6.3 Using Transformer Networks

The Transfer Learning concept depends on using a pre-training model that has some language knowledge, for a new task. Language model represents many language features, such as graded relationships [490] and sentiment direction [491].

In this section, I compared three transformer networks: a Robustly Optimized BERT Pretraining (RoBERTa) [374], and the two transfers networks designed for Arabic language, AraBERT [376] and the Universal Language Model in Arabic (hULMonA) [358]. I have chosen RoBERTa because it is shown to outperform Bidirectional Encoder Representations (BERT) in sentiment classification tasks [360], although BERT is the best in many NLP tasks [360]. As explained, this the first attempt of using RoBERTa for Arabic sentiment analysis, to the best of our knowledge. I have used Google Colab [492] for developing the experiments due to defective computer hardware.

5.6.3.1 RoBERTa Model Construction

I used the RoBERTa model, a BERT-based model with Adam optimisation [486], using the parameters seed=42 for the random weight, precision floating for GPU, and the batch size =16-64 for the maximum sequence length. After visualising the number of tokens in most tweets (Figure 5.8), the maximum tokens per tweet is 30 tokens. Therefore, all tokens were padded up to this size. After that, the model converts the word to an integer. The model used discriminative fine-tuning and gradual unfreezing. The model froze all the layers in the Neural Network except the last two layers.

RoBERTa trained over five datasets and 160GB text. To implement the sentiment analysis task, I use discriminative fine-tuning and gradual unfreezing. That means it predicts the next token, based on the present series of tokens in the sentiment corpus, with various learning rates, from 1e-02 to le-06. After that, the model unfreezes the output layer, after each epoch, layer by layer, except for the last two layers.

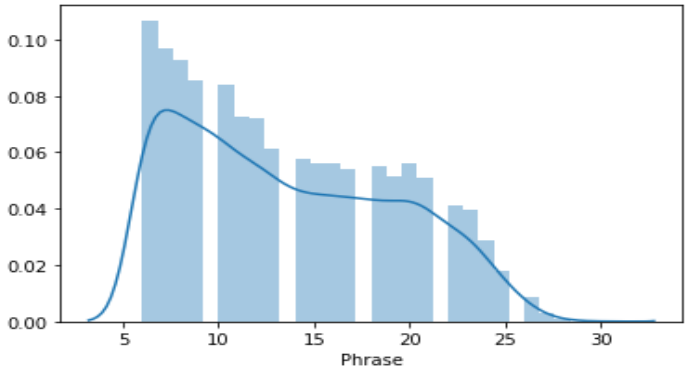

Figure 5.8: The number of tokens in most tweets.

5.6.3.2 AraBERT Model Construction

AraBERT is a BERT-based model; it is trained on different Arabic datasets. It used the BERT basic configuration [372], except a special pretraining was added before the experiment specific to the Arabic language. This tried to find the solution for the lexical sparsity in Arabic [350] which uses "ال" "Al" before the word (it is a prefix; it has no equivalent meaning in English) by using a Fast and Accurate Arabic Segmenter (Farasa) [378] to segment the word.

5.6.3.3 HULMonA Model Construction

HULMonA [358] is the first Arabic universal language model based on ULMFiT. It is pre-trained on large Arabic corpora and fine-tuned to many tasks. It consists of three stages: training AWD-LSTM model [371] on Arabic Wikipedia corpus, fine-tuning the model on a destination corpus, and for text classification, including a classification layer on the model. The results showed that hULMonA is superior in ASA.

5.6.3.4 Experiment Results, Discussion and Evaluation

When comparing the results of using RoBERTa, AraBERT, and hULMona models using the micro average of different metrics, Table 5.12 shows that the results favour the AraBERT model with 94.0% accuracy.

Model	Accuracy	F1	Recall	Precision
RoBERTa model	92.1	92.2	92.0	91.1
AraBERT model	94.0	92.6	92.1	93.0
hULMonA model	90.8	79.8	89.0	84.0

Table 5.12: Comparing between RoBERTa, AraBERT, and hULMonA Models.

To discover the reasons behind the obtained results, I analysed the three models' architecture. AraBERT outperformed the two other models because:

1. It is trained on different Arabic data sets – Modern Standard Arabic data sets and evaluated on dialectical data sets.

2. It applies a special pretraining specific to the Arabic language.

3. It uses Farasa [378]**,** a pre-processing tool directed to the Arabic language; it outperformed the state-of-the-art MADAMIRA [270].

Although hULMonA is trained on different Arabic data sets, its performance was worse than other models, because it is based on ULMFiT [359]. Additionally, it lacks the appropriate pre-processing for Arabic text. RoBERTa, which is a BERT based model, transfers each Arabic letter to a Latin character and every Latin character meets one integer number. That is the reason for the decrease in the performance of the model.

5.6.4 The proposed Model

While AraBERT performed better than other models, I propose a new model (AraBERT-GRU) – consisting of the AraBERT model combined with Bi-GRU, which achieved a high performance previously.

In the new model (Figure 5.9), the input embedded vector passed through the AraBERT [376] consisted of 12 layers, 768 in-features, 512 maximum sequence dimensions, and a total of 110M parameters. AraBERT generates the context vector. I created tensor data for training and the iterator to carry out iterations over the whole dataset. For optimisation, I used Adam optimisation. Also, a learning rate monitoring algorithm was used to change the learning rate and achieve better results.

There is a relationship between each sequential element. For that, Bi-GRU is needed to extract that relationship to classify. Once the sequential information extracted, the classifier layer classifies the sequence. Sequence information is forwarded from GRU to the next GRU and then to the final GRU. The last GRU output is fed to the SoftMax classifier layer.

I initialised the model with hidden dimensions 256- and 2-layer GRU, with the output dimensions 2, to take care of sequence information. Finally, the classifier layer classified into two classes – positive and negative. The classifier layer is a fully connected layer with a SoftMax, with a weight drop 0.25. The sequence flow of the new model is AraBERT, Bi-GRU, then classifier layer.

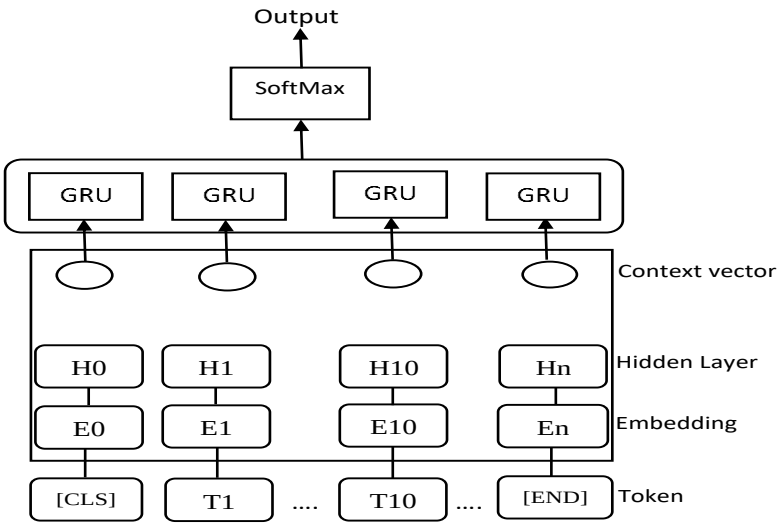

Figure 5.9: New model architecture.

5.6.4.1 *Results and discussion*

The results were impressive; the accuracy of the AraBERT before was 94.0% and after adding the Bi-GRU layer, the accuracy increased to 99.3% (Table 5.13), which means that the proposed AraBERT-GRU model achieved its goal of efficient, competitive prediction of sentiment, and thus, indirectly, customer satisfaction.

Social media is considered an easy-to-use platform. The number of internet users that use social media increased in 2019 to 2.77 billion users [80]. Therefore, there is a high probability of people using social media platforms for expressing their feelings. This may be why the new model is competitive and has obtained such a high accuracy (99.3%). The new model is suitable for a high volume of respondents and represents a cost-effective tool to monitor customer satisfaction on social media. Additionally, because of its dependence on text mining, there is also a possibility of generalising this model to different social media platforms.

Label	**Positive**	**Negative**	**Avg**
Precision	98.7	99.0	98.9
Recall	97.0	97.1	97.0
F1-Score	0.978	0.98	97.9
Accuracy	99.2	99.5	99.3

Table 5.13: Results of the new model.

In the next section, the AraBERT-GRU model has been applied on our own corpus, AraCust to predict actual customer satisfaction for the three companies.

I examined the statistical significance between the two (AraBERT and AraBERT-GRU models) to check the if there is a statical significant between the two models using McNemar's test [488] following [489]. I found that the p-value is $0.03 < 0.05$, which means reject H0 and accept the H1.

H0: There are no differences between the models' skill.

H1: There are significant differences between the models' skill.

5.7 Predict Customer Satisfaction

The third *RO* was to develop a potential model for the sentiment analysis of tweets to measure customer satisfaction using the real-time method. The application was aimed at Saudi Telecom companies STC, Mobily and Zain, as the largest providers in Saudi Arabia. In this study, a AraBERT-GRU model has been developed, which achieved a proficient result in predicting the customer satisfaction on the AraCust corpora. First, *customer satisfaction*, created based on domain expert knowledge, was calculated as follows:

$$cust\_sat = total\_ratings / (2 * num\_customers) \quad (5.1)$$

$$num\_customers = len(ratings) \quad (5.2)$$

$$total\_ratings = sum(ratings)\ (the\ summation\ of\ all\ ratings)\ rating:\ binary\ rating. \quad (5.3)$$

The corpus was then divided based on the company. The AraBERT-GRU model predicted customer satisfaction. These results showed that the predicted customer satisfaction percentage for the three companies STC, Mobily and Zain were 31.06%, 34.25% and 32.06%, respectively (all below 50%). Perhaps that was because customers tend to post a negative tweet rather than a positive tweet on Twitter, as previously observed.

5.7.1 Evaluating the New Model

This study has used a sentiment analysis to design an accurate model, by applying several approaches to measure customer satisfaction. To evaluate the proposed new model, we developed a simple questionnaire comprising two questions. The questionnaire is oriented to the customers whose tweets were mined, to evaluate the model by comparing the predicted customer satisfaction using the model, with the actual customer satisfaction, by using the questionnaire (Table 5.15).

I created an automatic tweet generator in Python (the tweet has a link to the questionnaire) to all the 20000 users whose tweets I had previously mined, but the respondents totalled just 530. The tweet generator was created using a code in Python for sending tweets that contain two things:

1. the link to the questionnaire and
2. mentions to the Twitter accounts of participants.

The code completed this procedure automatically without the need to do it myself, to save time (Figure 5.10).

```
In [26]: a=1

for a in range (1,100):
    message=" وذلك بتعبئة الاستبيان المرفق والذي يحتوي فقط على سؤالين، وسوف يتم الراجكم في السحب على هدايا قيمة "تحليل تويتراتوقع رضاء العملاء" عوكم للمشاركة في بحثي بعنوان
    "\n"+"@"+sentences[a]+" اء الرد بالموافقه او الرفض "+"\n"+"@"+"hamlaniko"
twitter1.update_status(status=message)
    a=a+1
```

Figure 5.10: Snapshot from the Python code for tweets generator.

I gave the participants the choice of answering or not. The questionnaire was built in google forms, because it is easy to build and distribute. The questionnaire had two questions: 'what is your telecom company?' and 'define your satisfaction toward your company (satisfied, unsatisfied), Appendix C. I received 530 responses. The sample was distributed between three companies, as shown in Figure 5.11.

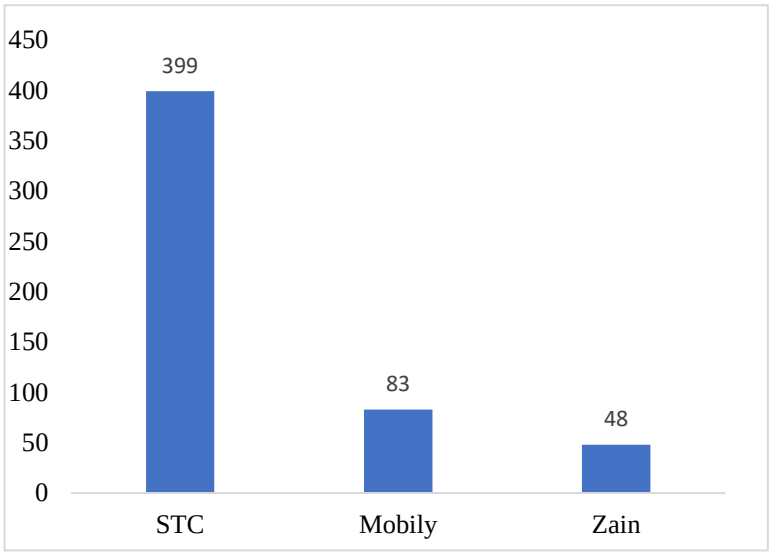

Figure 5.11: Number of participants based on telecom companies.

The unbalanced numbers of participants between the three companies reflects the real distribution of the users of the Saudi telecom companies. The number of unsatisfied and satisfied users for STC, Mobily and Zain companies is shown in Figure 5.12.

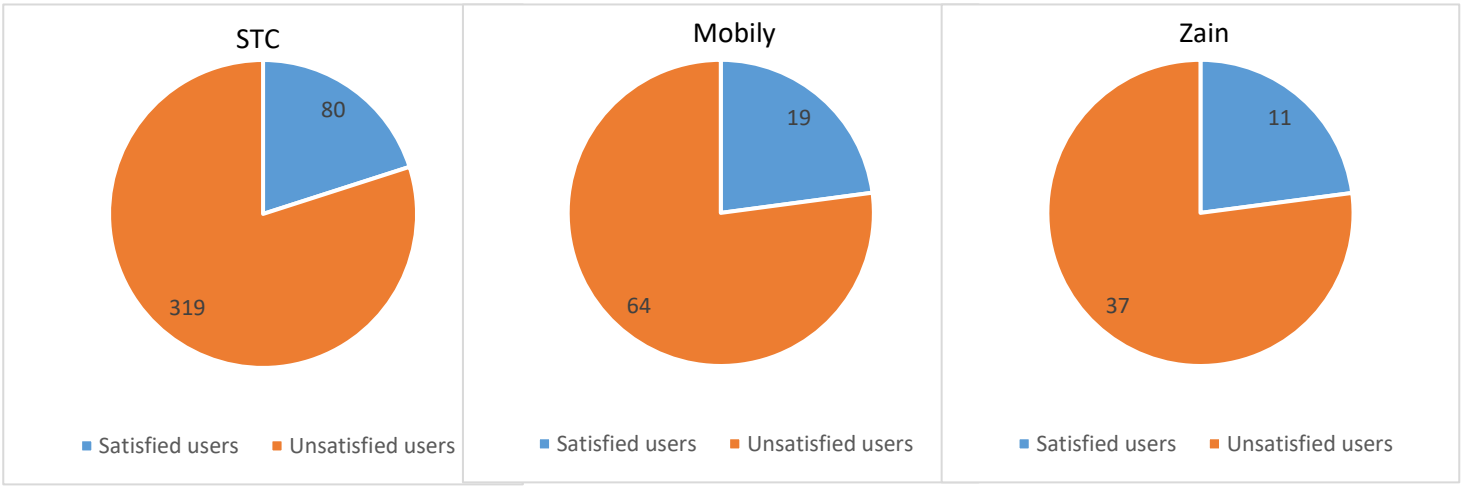

Figure 5.12: Number of satisfied and unsatisfied users for STC, Mobily and Zain companies.

In Table 5.14, it is shown that the AraBERT-GRU model achieved the goal of predicting the customer satisfaction of telecom companies based on the Twitter analysis.

These results can provide insights for the decision-makers in these companies regarding the percentage of customer satisfaction and help to improve the services provided by these companies. These results should encourage the decision-makers to consider using Twitter analyses for measuring customer satisfaction and to include them as a arguably reliable method for evaluating their marketing strategies.

Company	Predicted Customer's Satisfaction	Actual Customer's Satisfaction
STC	31.06%	20.1%
Mobily	34.25%	22.89%
Zain	32.06%	22.91%

Table 5.14: Percentage of predicted customer's satisfaction vs. actual customer's satisfaction.

5.8 Summary

This chapter explained the experiments completed using ASA to achieve the *RO3* and answer the *RQ3*, which defines the customer satisfaction percentages on the three Saudi telecom companies STC, Mobily and Zain. The experiments started with applying SVM as a baseline model with many feature selections experiments to choose the best feature sets. I used the statistical evaluation of the feature selection method. Contrary to

the initial expectations in this study, although there are a significant number of prayers in the Arabic tweet corpus, the *Has-Prayer* feature had to be removed from the feature set, possibly due to both positive and negative tweets that use prayers, often including the word 'God'.

Next, I compared two deep learning models LSTM and GRU, with different embeddings and settings on AraCust to define the best model for AraCust corpus and Arabic dialect characteristics. After that, three transfer networks designed for Arabic language AraBERT, hULMonA and RoBERTa models were utilised on AraCust to define the best performance suitable to the corpus and the dialect Arabic characteristics. Finally, the proposed model combining the AraBERT model and Bi-GRU predicted customer satisfaction for the three companies.

The results showed that the new model is highly accurate when compared to other models. Moreover, there is the possibility to generalise the proposed model for use on different social media platforms – a simple questionnaire distributed to the customers to calculate the same companies' customer satisfaction percentage.

In the next chapter, customer satisfaction experiments will further predict the customer churn for the telecom company that offered the historical data to this study and compare it to the customer churn percentage obtained from the company.

Chapter 6: Multi-Way Arabic Sentiment Analysis

6.1 Introduction

Usually, past studies considered binary sentiment analysis (positive or negative) instead of multi-way sentiment analysis (MWSA), which means classifying the dataset into several (>2) classes [493]. This is due to the complexity of multi-way sentiment analysis for machine learning to predict [494]. Some studies attempt to alleviate this complexity by using a rating scale (from poor to excellent) as multi-way sentiment analysis [494], [69]. Hence, this Chapter attempts to alleviate this matter by focusing on MWASA assigning one of the multi-level classes to tweets - from 'Strongly negative' to 'Strongly positive.' The reason for doing this is that, whilst there is a lack of studies that use multi-way analysis, many studies discussed and recommended this type of classification [493].

This Chapter achieves objective *RO2*, which is to propose recommendations to improve Saudi telecom companies' services. This chapter also answers research question *RQ2*: *what type of services for customers of telecom companies in Saudi Arabia are mentioned in tweets, and what is the customer sentiment about these services?* The first part of the research question is answered by manually creating and annotating a gold-standard corpus, as explained in Chapter 3. For the second part of the research question, SA needs to be performed on this corpus.

Thus, to address the second part of the research question and find the best possible sentiment classifier for this problem, this study compares *flat classification* and *hierarchical classification structures*, using (MWASA) classifier on the same data set. As a result of this comparison, a hierarchical classification structure is proposed for MWASA in this Chapter. The hierarchical classifier structure consists of four-level binary classifiers, and this Chapter shows how this addresses the multi-way sentiment analysis and raises the MWSA classifier performance.

6.2 Related Research

Starting with Sentiment Analysis in general, outside the Arab language sphere, usually, studies consider binary sentiment analysis (positive or negative) instead of MWSA, the latter meaning classifying the data set to classes greater than two [493], [495]. The reason for this is that the complexity of predicting the multi-way sentiment analysis for machine learning is higher than that of the binary sentiment analysis [494]. Some studies used a rating scale (from low to excellent) as multi-way sentiment analysis [494], [496], [69].

Bickerstaffe and Zukerman [496] proposed a hierarchical classifier for the multi-way classification, taking into consideration the inter-categories' similarity. The proposed classifier depended on SVM for removing unrelated features. The results proved the efficiency of their proposed classifier for movie reviews with three or four-star ratings.

Due to the significance of the MWSA problem, SemEval-2017[22] [497], included it in Task 4, and SemEval-2018 [498], included it in Task 1. The task 4 in 2017 [497], is about sentiment analysis on Twitter. They included MWSA in the classification and quantification subtasks. The classification subtask focused on prediction of the sentiment of the tweet towards a given topic, via two classes (positive and negative), three classes (positive, neutral and negative) and five classes (strongly positive, positive, neutral, negative and strongly negative). The OMAM [499] team achieved the top ranks in subtask C, topic-based polarity classification (the only task relevant to the current research). The OMAM system achieved 0.9431 macro average mean absolute error (MAEM) and 0.6461 standard mean absolute error (MAEμ).

As already stated throughout this thesis, ASA is less represented by this body of research [45]. Relevant for the current Chapter is that especially studies proposing MWASA are very few and far between. Due to the importance of MWSA, some studies have applied it to different Arabic datasets. One of the earliest studies that addressed the MWSA problem [230] presented a Large-Scale Arabic Book Reviews (LABR) dataset collected from the Good Reads website and performed two tasks on the dataset: sentiment polarity classification and rating classification. LABR consisted of 63,000 Arabic book reviews written in MSA and

[22] SemEval are annual natural language (NLP) competitions posted for the NLP community at large to solve: SemEval | International Workshop on Semantic Evaluation

DA. In the rating classification, each review was rated on a scale from 1 to 5 stars. For the sentiment polarity classification, the review was labelled as positive, if the rating is 4 or 5; or negative, with 1 or 2. They used Multinomial Naive Bayes (MNB), Bernoulli Naive Bayes (BNB) and SVM implemented in Python. According to their results, SVM performed better than other classifiers in the unbalanced setting, with accuracy/weighted F1 measure 0.503/0.491. One subsequent LABR study is [202].

In addition, Elnagar et al. [500] introduced a Hotel Arabic-Reviews Dataset (HARD) containing 370,000 hotel reviews from the Booking website in MSA and Gulf Dialectal Arabic (GDA). They carried out three experiments on their dataset: polarity classification using a supervised approach, rating classification using a supervised approach and polarity classification using the lexicon-based approach. Their results showed that the supervised approach outperformed the lexicon-based approach in polarity classification with 94%-97% against 89%. However, in the rating classification, they obtained a lower accuracy than 75%.

Another related research [501] developed the Aara system for opinion mining for 815 comments from local Saudi newspapers. The comments were written in Arabic, both formal and colloquial. The specific dialect was Najdi Arabic, used in Riyadh (the capital city of Saudi Arabia). They used an NB classifier to classify Arabic reviews into four categories (strongly negative, negative, positive, strongly positive). Their system achieved 82% accuracy and 84.5% macro-averaged F-score.

Nabil et al. [273] introduced an Arabic social sentiment analysis dataset (ASTD), collected from Twitter. They applied MNB, BNB, SVM, stochastic gradient descent (SGD) and KNN on four sentiment classes (objective, subjective positive, subjective negative, and subjective mixed). The results showed that SVM performs better than other classifiers.

From the above studies, it is obvious that MWASA is an interesting and relevant topic and that it still has a wide scope for improvement. There is a lack of studies that used multi-way analysis, although many studies used this type of classification [493]. As a result, I have decided to apply MWASA on my corpus for telecom companies. I consider the multi-way sentiment analysis as assigning one of the multi-classes for tweets, in a range starting from ‘Strongly negative’ to ‘Strongly positive’ as recommended by [502].

Some studies proposed the Hierarchical Classifiers (HCs) instead of Flat Classifiers (FCs) [502], [503], [504] to address MWSA. HCs are a type of well-defined classification that works hierarchically, where the classification output matches one or more classification labels [503]. HCs classify the classes sequentially in a hierarchical way and consider the relations between the classes. Hao et al. [505] explained the HC structure as classifying the document starting from root level to sub-categories, until the classification extends to the leaf category at the leaf node level. HCs are widely used with massive and mixed data [503]. HCs were proposed to alleviate issues with FCs, which classify the data at one level (leaf nodes) regardless, where the relations between the classes could affect the classifier performance on a big dataset [503]. Despite this drawback of FCs, however, some researchers prefer them, because they are less complicated and more straightforward [502], [505].

The advantages of using the hierarchy classification demonstrated in the literature include effective learning, by subdividing the classification problem into sub-problems [506]; keeping the characterising ability regarding the sole classifier, due to the fact that a HC is flexible and customisable [507]; and the fact that it is able to complete the classification of a large problem faster and more efficiently than flat classification, because HC divides the large problem into simpler subproblems [505]. In addition, some studies have proved the higher performance of HC when compared to FC [503], [507], [505], [504].

Angiani et al. [507] detected emotions using flat and hierarchical classifiers. The HC that they used had three levels, and on each level, there was a binary classifier. The first classifier classified a tweet as subjective or objective. The second one classified the subjective tweet as positive or negative. The third classifier classified a positive tweet as one of three positive emotions and a negative tweet as one of three negative emotions. The authors applied Naive Bayes Multinomial to 10,000 tweets. The best accuracy achieved by HC was 45.17%.

Hao et al. [505] consider the Binary Hierarchical Classifier as a virtual tree of category classification, organising the classes on a tree that the multi-class classification addresses by some binary classifiers [508].

There is a lack of studies that proposed a hierarchical classifier in Arabic sentiment analysis [470], [502], [504]. Per to our knowledge, [470] the first who proposed hierarchical structure for MWASA problem. They applied two HCs and compared them with FC. They used Large Scale Arabic Book Reviews (LABR) dataset

[230]. They used many classifiers SVM, Naive Bayes, Decision Trees, and KNN. The two HCs that they proposed are 2-level HC and 3-level HC. The 2-level HC start with ternary classifier classified the text to positive, negative or neutral. The second classifier classified the positive text to weak or strong and the third classifier classified the negative text to strong or weak. While the 4-level HC composed of 4 binary classifiers starting with classifying the majority class with others sentiment class. The results proved that the hierarchical classifiers improved the classification by c50% more than flat classifiers.

Followed by [502] proposed six different Hierarchical Classifiers to solve the MWSA problem. They applied a supervised approach (corpus-based) which included SVM, NB, KNN and DT. The first hierarchical classifier has two levels, the first one classifies the text to (negative, neutral, positive), the second level classifies the text to (strong, weak). The second hierarchical classifier has 4 binary classifiers classify one label against the rest of the labels. The third hierarchical classifier has two levels, the first level classifies to (strong positive, strong negative), then it classifies to (weak positive, neutral and weak negative). The fourth hierarchical classifier has four classifiers, the first one classifies to (neutral, not neutral), the second one classifies to (weak positive, weak negative), the third one classifies to (weak positive, strong positive), the fourth one classifies to (weak negative, strong negative). The fifth hierarchical classifier has four classifiers starting by classifying the text to the majority sentiment label from the other sentiment labels. The sixth hierarchical classifier has two level classifiers: the first one is a flat classifier that classifies a text to all sentiment classes. When comparing the results between the FC and HC, the best accuracy and Mean Square Error (MSE) for FC were 45.77% and 1.61, and the best accuracy and MSE for an HC were 72.64% and 0.53, respectively.

In addition, [504] compared between 2-level, 3-level and 4-level binary HC using different techniques. Their dataset was Hotel Arabic Reviews Dataset (HARD) [500]. The 2-level binary HC composed of two binary classifiers, the first one classified the text to positive, negative or neutral. The second classifier classified the positive text to strong or weak and the negative text to strong or weak. In the 3-level binary HC, the first classifier classified the text to neutral or not neutral. The second classifier classified the text to positive, or negative. The last classifiers classified the text to (weak or strong) positive/negative. The 4-level binary HC start by classifying the text to strong positive or other sentiment labels, taking in consideration that the

majority in the data set are strong positive. They achieved the best results with Random Forest and Decision Tree, while the worst results with SVM and NB. The best result got by Decision Tree for the three-level binary HC with 99% for Accuracy and F1.

6.3 Corpus Collecting and Annotating

To create the corpus, as explained in Chapter 3, I asked the annotators to annotate the 20,000 tweets with the pre-defined telecom services from my list that matched the one mentioned in a tweet. This list of telecom services was extracted from customer satisfaction metrics defined by the Saudi Communications and Information Technology Commission [438] and related researches. The annotators were asked to annotate with labels from this list of pre-defined services and the sentiment towards that particular service, if existent. They could identify more than one service in a single tweet. The annotators used the five-way sentiment analysis scale (Strongly Positive, Positive, Neutral, Negative, Strongly Negative). They found 4,380 tweets that mentioned one or two services. They listed the services mentioned in the tweets: Network Coverage, Phone Network, Quality of Voice Transmission, Customer Service, Successful Calls, Billing Price, Good Offers, Reasonable Fees when calling another Telecom Company, Browsing Speed, and Hiring Section. The Network Coverage and Phone Network were subsequently merged after analysing the initial sets of labels with an expert, as they pointed to the same service. Additionally, I merged Internet speed and Browsing Speed. I excluded the Hiring Section, because it was out of scope for this research. After that, I listed the final list of services for which I will identify the sentiment as follows: Network Coverage, Quality of Voice Transmission, Customer Service, Number of Successful Calls, Billing Price, Good Offers, Reasonable Fees when calling another Telecom Company, and Internet Speed. Each sentiment label considers the degree of customer satisfaction towards that specific telecom service. Tables 6.1, 6.2, and 6.3 show the number of tweets mentioning each service in each company in the corpus, which I called *AraCust1*.

Services\Company	**# Positive**	**#Strongly Positive**	**#Neutral**	**#Negative**	**#Strongly Negative**	**Total**
Network Coverage	44	1	0	50	0	95
Quality of Voice Transmission	50	4	1	50	5	110
Customer Service	160	9	0	320	10	499

Successful Calls	60	3	0	69	2	134
Billing Price	35	0	0	45	3	83
Reasonable Fees when calling another Telecom Company	65	1	0	65	3	134
Good Offers	100	0	1	114	10	225
Internet Speed	130	17	0	319	1	467
Total	**644**	**35**	**2**	**1032**	**34**	**1747**

Table 6.1: The number of tweets in AraCust1 for each category in the STC company.

Services\Company	# Positive	#Strongly Positive	#Neutral	#Negative	#Strongly Negative	Total
Network Coverage	120	21	0	250	50	441
Quality of Voice Transmission	38	1	0	38	2	79
Customer Service	25	10	0	80	20	135
Successful Calls	54	7	0	58	5	124
Billing Price	14	7	0	20	2	43
Reasonable Fees when calling another Telecom Company	40	21	0	50	12	123
Good Offers	50	4	1	70	6	131
Internet Speed	50	7	0	108	3	168
Total	**391**	**78**	**1**	**674**	**100**	**1244**

Table 6.2: The number of tweets in AraCust1 for each category in the Mobily company.

Services\Company	# Positive	#Strongly Positive	#Neutral	#Negative	#Strongly Negative	Total
Network Coverage	48	8	1	52	20	129
Quality of Voice Transmission	100	20	0	90	36	246
Customer Service	122	30	1	122	57	332
Successful Calls	30	15	0	70	12	127
Billing Price	73	8	0	77	9	167
Reasonable Fees when calling another Telecom Company	112	4	0	100	16	232
Good Offers	94	5	0	95	4	198
Internet Speed	115	4	2	220	68	409
Total	**694**	**94**	**4**	**826**	**222**	**1840**

Table 6.3: The number of tweets in AraCust1 for each category in the Zain company.

6.4 Evaluation Metrics

To compare the performances of HC and FC models, I used five metrics suitable for multi-way classification: micro averages of Precision (Pr), Recall (Rc), F1, Accuracy (Ac), and Mean Square Error (MSE) [509], as motivated by [502]. The micro average is suitable for multi-classes, especially if the classes are imbalanced, and the micro average totals all classes' contribution to the average metric calculation [474]. It aggregates the precision and recall of the classes.

MSE is calculated as shown [470]:

$$\text{MSE}= \frac{1}{n}\sum_{i=1}^{n}(yi - yi)^2 \qquad (6.1)$$

Where *y'i* is predicted value within *n* predictions and *yi* is the true value.

The reason for using MSE is that the four metrics, Pr, Rc, F1, and Ac, do not consider the relationship the relation between the classes in a hierarchical classification [510]. However, MSE considers the gap between the real category and the predicted category, making it more appropriate for MWSA. Nevertheless, for a comprehensive set of results, I have used all of the other metrics as well.

6.5 Model Construction

6.5.1 Flat Classification

The classifier classifies the tweets into five-way (Strongly Positive, Positive, Neutral, Negative, Strongly Negative) on one level (Figure 6.1).

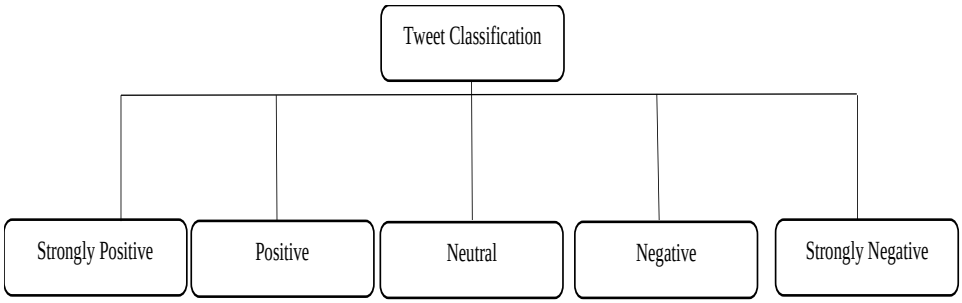

Figure 6.1: Flat classification structure of Tweets.

I have carried out the experiments using the proposed classifier described in Chapter 5. The flat experiment started by using Python libraries, such as Panda, Keras and Sklearn, reading the data file of the two types, with one service label (Table 6.4) and two service labels (Table 6.5). The file has four columns: tweet text, mentioned services, sentiment label and mentioned company. The row represents each tweet.

	Tweet	Label	Service	Company
0	الفاتورة غير معقولة مرتفعه جدا جدا	Strongly Negative	Billing Price	STC
1	وتظام الرد فيه مشكلة وما استطعت الوصول لاي موظف	Negative	Customer Service	STC

Table 6.4: Sample of the data set with one service label.

Tweet	Label-service1	Label-service2	Service	Company	Num_service
والله نتكم مجنون و السعر الحلو	Positive	Positive	Internet Speed, Billing Price	Zain	2
تغطيه وانترنت زي الخرا	Negative	Negative	Network Coverage, Internet Speed	Zain	2

Table 6.5: Sample of the data set with two service labels.

I defined the number of services per tweet as follows in the Python code snippets provided, where *df* means the data frame, *Num_ser* is the number of services:

```
df['Num_ser'] = df. Service. apply (lambda x: len(x.split(',')))
```

Creating a data frame which has one service:

```
df_ = df[df['Num_ser']==1]
```

Creating a data frame which has 2 services:

```
df2 = df[df['Num_ser']>1]
```

I then generated 8 column data frames, one column for each service:

```
table = pd. pivot_table (df_, values='Label', index=df_. index,
columns=['Service'])
```

After that, I converted the data frame (*df2*) with all services as columns filled with 0 except for the services that were mentioned in the tweet; I filled these with the sentiment, where 2 is for Strongly Positive, 1 for Positive, 0 for Neutral, -1 for Negative and -2 for Strongly Negative (Table 6-6).

```
services = table.columns[0:-2].tolist()
```

for *i* in services:

```
 df2[i] = np.ones((len(df2)))
```

Adding the label:

for i in serv2:

```
df2[i] = df2['Label']
```

Next, this processed data was used for applying on it the various models.

6.5.2 Hierarchical Classification

For multi-way Sentiment Analysis, hierarchical classifiers containing many classification layers are recommended for ASA, following [502] and [504], Section 6.2. Addi et al. [504] proposed the 3-level binary classification because it is less complicated than other hierarchical structures and does not consider the data balance. In [502] and [504], they used different hierarchical structure techniques, but they did not apply deep learning models. Here, I will apply the deep learning model with the hierarchical structure and see how it works with MWASA and hierarchical structure. Here, three layers of classification were used, as shown in Figure 6.2. The first layer used a binary classifier to classify the tweets as 'Has the sentiment'; 'Neutral' means that a tweet does not include an opinion about the target, any one of the telecom company services. Then, in the second layer, the binary classifier classifies the tweets as Positive or Negative. There are two binary classifiers in the third layer; the third classify the tweets as Strongly Positive or Positive, and the fourth classifies the tweets as Strongly Negative or Negative. Four data frames were created via a Python transcript – one for each model:

The first data frame had 2 classes: ‘neutral’ or ‘has sentiment’.

```
label_dict = {'Neutral':0, 'negative':1,’Strongly Negative':1,
'Positive':1,'Strongly Positive':1}
df1 = df [df.Label.isin(['Neutral', 'Negative',' Strongly Negative',
'Positive', ‘Strongly Positive'])]
```

The second data frame was for further classifying the output of the ‘has sentiment’ class as negative, or positive. It did not affect the neutral class output.

```
label_dict = {‘Strongly Positive':1,’Positive’: 1,'Strongly
Negative':0,’Negative’:0}
df2 = df [df. Label.isin([‘Strongly Positive',’Positive’,'Strongly
Negative',’Negative’])]
df2.Label = df2.Label. apply (lambda x: label_dict[x])
```

The third data frame was only applied to the previous outcome of the positive class, and further differentiated between strongly positive and positive.

```
label_dict = {‘Strongly Positive':1,’Positive’: 0}
df3 = df [df. Label.isin([‘Strongly Positive',’Positive’])]
df3.Label = df2.Label. apply(lambda x: label_dict[x])
```

The fourth data frame performed the same for the outcome of the negative class, resulting in an output of: strongly negative or negative.

```
label_dict = {‘Strongly Negative':1,’Negative’:0}
df4 = df [df. Label.isin(['Strongly Negative',’Negative’])]
df4.Label = df2.Label.apply(lambda x: label_dict[x])
```

Then the four models model1, model2, model3 and model4 were trained on the four datasets, respectively, df1, df2, df3 and df4.

```
model1=build_2_input_gru_classifier(bidirectional=True,add_attention=Tr
ue,trained_embedding=True,num_classes=2)
model2=build_2_input_gru_classifier(bidirectional=True,add_attention=Tr
ue,trained_embedding=True,num_classes=2)
model3=build_2_input_gru_classifier(bidirectional=True,add_attention=Tr
ue,trained_embedding=True,num_classes=2)
model4=build_2_input_gru_classifier(bidirectional=True,add_attention=Tr
ue,trained_embedding=True,num_classes=2)
model1.load_weights ('Model1.hdfs')
model2.load_weights ('Model2.hdfs')
model3.load_weights ('Model3.hdfs').
model4.load_weights ('Model4.hdfs')
```

After that, the hierarchal classifier was built and applied for merged data frames.

```
df = pd. concat([df1,df2,df3,df4])
```

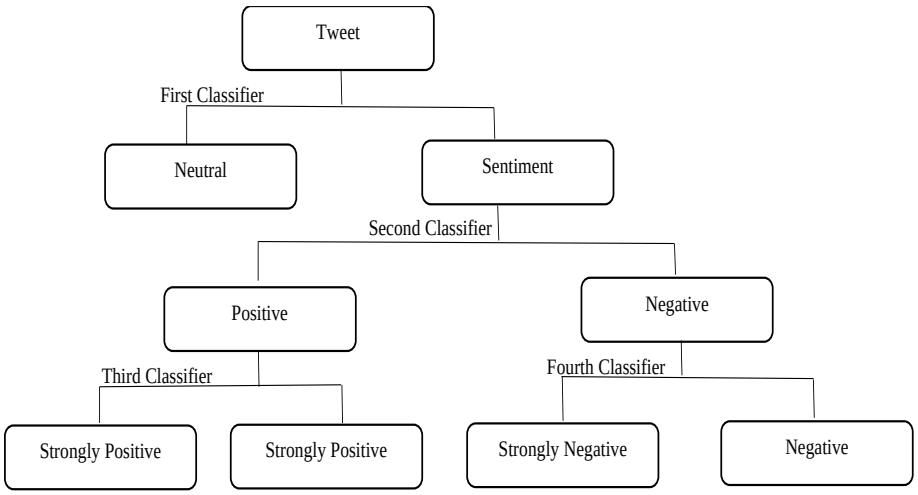

Figure 6.2: Hierarchical classification structure of the tweets.

6.6 Experimental Results

Table 6.7 shows that there is improvement using HC, with 83.17% accuracy. The MSE obtained using HC was 0.31. That means that the hierarchical structure improves model performance.

	Ac	F1	Pr	Rc	MSE
HC	83.17%	0.80	0.80	0.80	0.31
FC	70.4%	0.70	0.70	0.70	0.35

Table 6.7: Comparing between Hierarchical classification and Flat classification.

6.7 Comparing with similar study

At present, only three studies have used hierarchical structure for MWASA. The experiment run here was in 2019, and only two studies preceded it. In comparison with these studies found and mentioned in the related research, results are as follows. The best accuracy and Mean Square Error (MSE) in [502] for HC6 which is 2-level classifier are 72.64% and 0.53, respectively. While the best accuracy and MSE for [470] was for 4-level hierarchical classifier using KNN with 57.8% for accuracy and 0.96 for MSE. Regarding the [504] the best accuracy and F1 were 99.00% using decision tree in 3-level binary classification.

So, my HC structure outperforms their HC structure with 83.71% and 0.31 for accuracy and MSE.

Recently, in 2020, another study was published that outperformed mine [504]. This is possibly due to their database, which is based on Modern Standard Arabic, which is easier to process than Dialectal Arabic (as in my work). Thus, results still emphasise the effectiveness of my approach to solving the MWASA problem.

6.8 Customer Satisfaction Toward the Services

To answer the second research question, *RQ2 – what type of services for customers of telecom companies in Saudi Arabia are mentioned in tweets, and what is the customer sentiment about these services?* – I applied the proposed hierarchical model on the *AraCuts1*. Table 6.8 shows the average F1 score for all services, over 10 training sessions.

Service	FI score
Billing Price	0.96
Quality of voice transmission	0.98
Customer Service	0.88
Internet Speed	0.88
Network Coverage	0.96
Good Offers	0.88
Reasonable fees when calling someone uses another telecom company	0.85
Successful Calls	0.91
Average F1 score on all services	**0.91**

Table 6.8: F1 average for each service.

After that, I applied the HC model to calculate the customer satisfaction percentage towards each company's service, based on the equation (5.1) in Chapter 5, Figure 6.3.

Calculating Customer satisifaction

In [322]:
```
def calculate_customer_satistaftion(ratings):
    num_customers = len(ratings)
    total_ratings = sum(ratings)
    cust_sat = total_ratings/(4*num_customers) # we used 4 as neutral is 0
    return cust_sat
```

In [323]:
```
def extrapolate(pred_prop):
    ratings_list = []
    for i in range(8):
        ratings  = pred_prop[i][:,1]*2 + pred_prop[i][:,2]*4
        ratings_list.append(ratings)
    return ratings_list
```

In [327]:
```
### Predicted Customer Satisfaction
pred_prop = model.predict(X)
ratings_list = extrapolate(pred_prop)
sat_df = pd.DataFrame(columns=services,index=companies)
for s in range(len(services)):
    ratings = ratings_list[s]
    for company in companies:
        ratings_comp = ratings[table.Company==company]
        company_sat = calculate_customer_satistaftion(ratings_comp)
        sat_df.loc[company,services[s]] = company_sat
sat_df
```

Figure 6.3: Calculating customer satisfaction using HC.

The code shown in Figure 6.3 calculates the customer satisfaction percentage using the *calculate_customer_satisfaction* function that used four labels – ignoring the neutral label because its weight is 0 – using the HC model through *pred_prop* as input for the extrapolating function, to calculate each service's customer satisfaction percentage through the *for* loop.

Tweet	Label	Service	Company	Num _ser	Bill Price	Call quality	Customer Service	Hiring section	Internet Speed	Network Coverage	Offers	Reasonable fees	Successful Call	Internet Speed
عندكم ضعف الشبكه والابراج	2	Network Coverage, Internet	Zain	2	0	0	0	0	0	2	0	0	0	2
شكرا ، الشبكة تحسنت والانترنت	2	Network Coverage, Internet	Zain	2	0	0	0	0	0	4	0	0	0	4

Table 6.6: The data frame after the services were filled with 0 (except for the services mentioned in the tweet).

Services\Company	Customer Satisfaction percentage (%)		
	STC	Mobily	Zain
Network Coverage	47.37	31.97	48.06
Quality of voice transmission	49.09	49.37	48.78
Customer Service	33.87	25.93	45.92
Successful Calls	47.01	49.19	35.43
Billing Price	42.17	48.84	48.50
Reasonable fees when calling another telecom company	49.25	49.59	50.00
Good Offers	44.44	41.22	50.00
Internet Speed	31.48	33.93	29.10

Table 6.9: Customer Satisfaction of the STC, Mobily and Zain customers toward the services.

6.9 Discussion

RQ2*: *What type of services for customers of telecom companies in Saudi Arabia are mentioned in tweets, and what is the sentiment of customers about these services?

The listed services that are mentioned in the tweets are as follows: Network Coverage, Phone Network, Quality of voice transmission, Customer Service, Successful calls, Billing Price, Good Offers, Reasonable fees when calling someone uses another telecom company, Browsing Speed and Hiring section. The Network Coverage and Phone Network were merged, as they pointed to the same service. Additionally, I merged Internet Speed and Browsing speed. I excluded the hiring section because it is out of scope for this research.

Table 6.9 shows that the average customer satisfaction percentage towards the three companies' service is below 50%. The STC customer satisfaction percentage is between 31.48% for Internet Speed and 49.25% for Reasonable Fees when calling another Telecom Company. That is under 50%, consistent with the customer satisfaction percentage overall towards the company, which is 31.06%.

Regarding the Mobily company, the customer service scored lower satisfaction with 25.93%, although the Reasonable Fees when calling another Telecom Company service received higher satisfaction rates of 49.59%.

For the Zain company, the lowest satisfaction percentage of 29.10% is for Internet Speed service. In addition, it scored a higher satisfaction percentage than other companies with 50.0% for Reasonable Fees when calling another Telecom Company and Good Offers.

***RO2*: To propose recommendations to improve the services of Saudi telecom companies.**

To achieve the second research objective, I used the tableau[23] software to visualise the service importance versus customer satisfaction towards the service due to the enormous potential of tableau and its easy use.

[23] https://www.tableau.com

The service importance was obtained from the questionnaire analysis in Chapter 4. I visualised this correlation to highlight potential recommendations for the decision-makers of the three telecom companies.

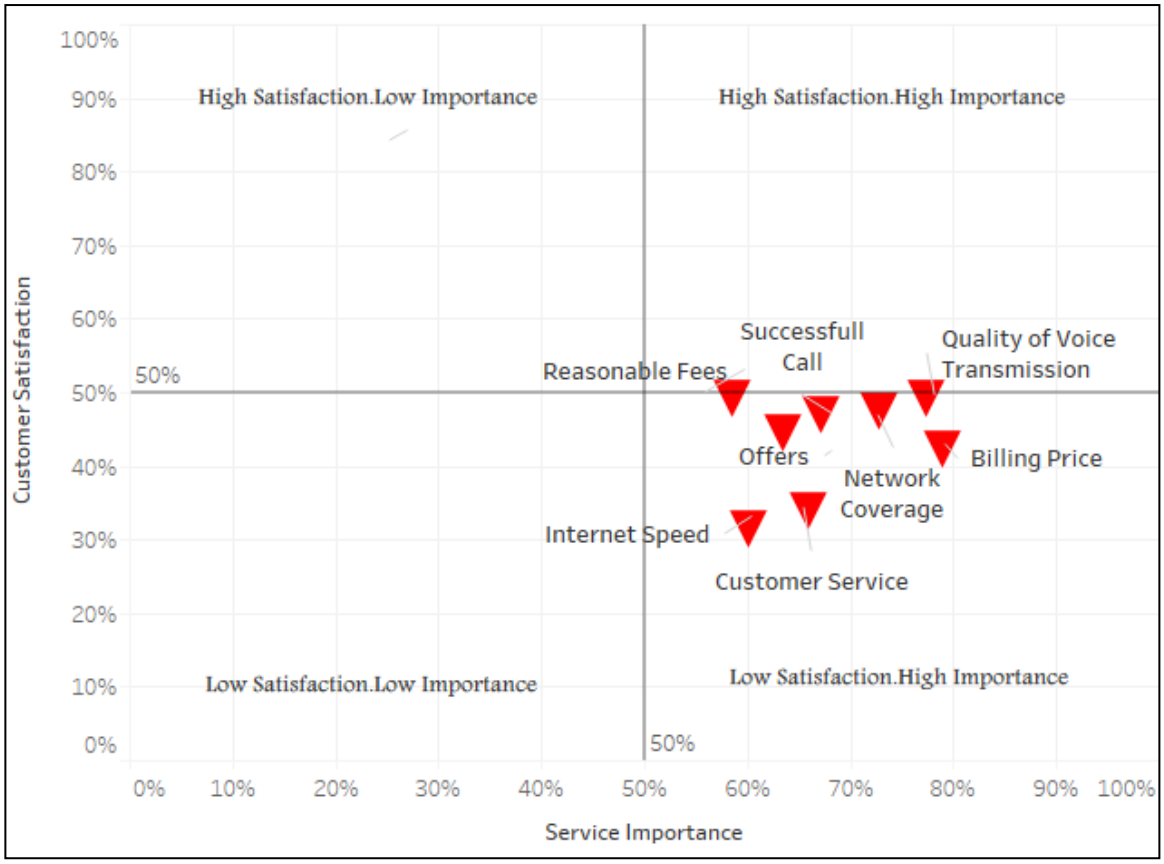

Figure 6.4: The importance versus customer satisfaction for STC customers.

As shown in Figure 6.4, all the STC company services are placed on Low satisfaction and High importance area. All the services are important from the customers' point of view, yet the customer satisfaction towards these services is lower than 50%. Therefore, STC decision-makers need to pay attention to all these services, especially Internet Speed, where customer satisfaction is 31.48% and its service importance rating is much higher at 60.2%. In addition, Customer Service scored 33.87% customer satisfaction, and 65.62% service importance. That means these services are highly important to STC customers, but the satisfaction was low. Billing Price and Quality of Voice Transmission scored 78.9% and 77.4% as the most important services,

while the satisfaction was 42.17% and 49.09%; this is better than the satisfaction towards other standards, but there is room for improvement.

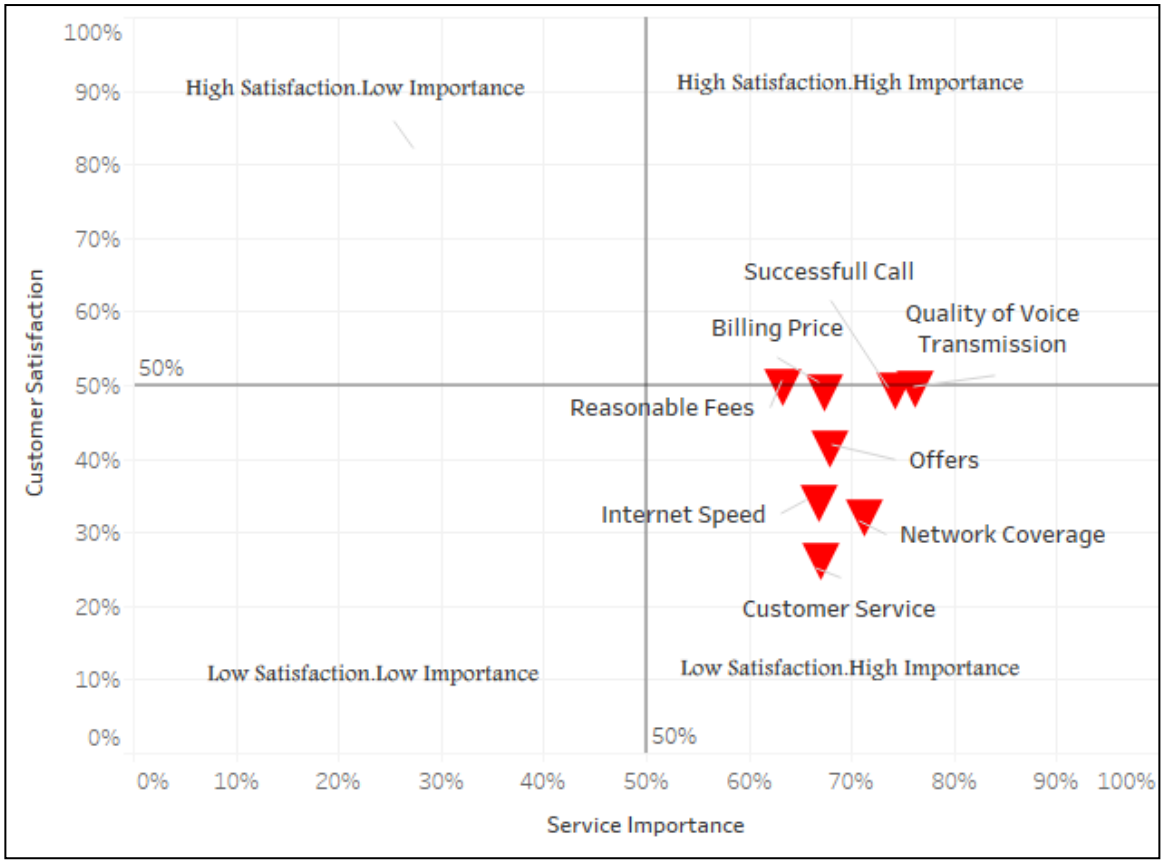

Figure 6.5: The importance versus customer satisfaction for Mobily customers.

Figure 6.5 shows four of the services – Customer Services, Network Coverage, Internet Speed, and Offers – located in the area Low satisfaction/High importance for the Mobily company. They scored the lower satisfaction rate of 25.93%, 31.97%, 33.93% and 41.22%. This means that the Mobily company's decision-makers need to improve these services because their importance is much higher than customer satisfaction. This explains the answers to the open question in the questionnaire (Chapter 4): **Why did you change from the previous telecommunication company you have used for your mobile phone?** The Slow response of Customer Service was the reason most mentioned. Successful Calls, Reasonable Fees when calling another Telecom Company, Billing Price, and Quality of Voice Transmission are located in the borderline between Low Satisfaction areas, High Importance and High Satisfaction, and High importance, indicating that the service is important customer satisfaction percentage towards the service are close. Reasonable Fees

when calling another Telecom Company achieved the highest customer satisfaction percentage of 49.59%. This finding is consistent with the answers to the open question: **Why did you change from the previous telecommunication company you have used for your mobile phone or Internet access?** The Unreasonable Fees when calling another Telecom Company was mentioned just by one participant as the reason for changing the company. In addition, the bad Quality of Voice Transmission received just one response as the reason for changing the company; its importance from a customer point of view is the highest percentage of all the services (72.6%), and the customer satisfaction towards this service is 49.37%. This service is good in Mobily company, but there is also room for improvement. The high Billing Price and Bad/Lack of Network coverage received the same number of responses as the answers to the open question mentioned above – 28.0% responses and 27.0% responses. Billing price scored 69% for the service importance and 48.84% for the customer satisfaction towards it from the customers' point of view. The Network Coverage scored 71% for service importance and 31.97% for customer satisfaction. Therefore, the Mobily company's decision-makers need to improve these two services to increase customer satisfaction, especially given the importance of these services.

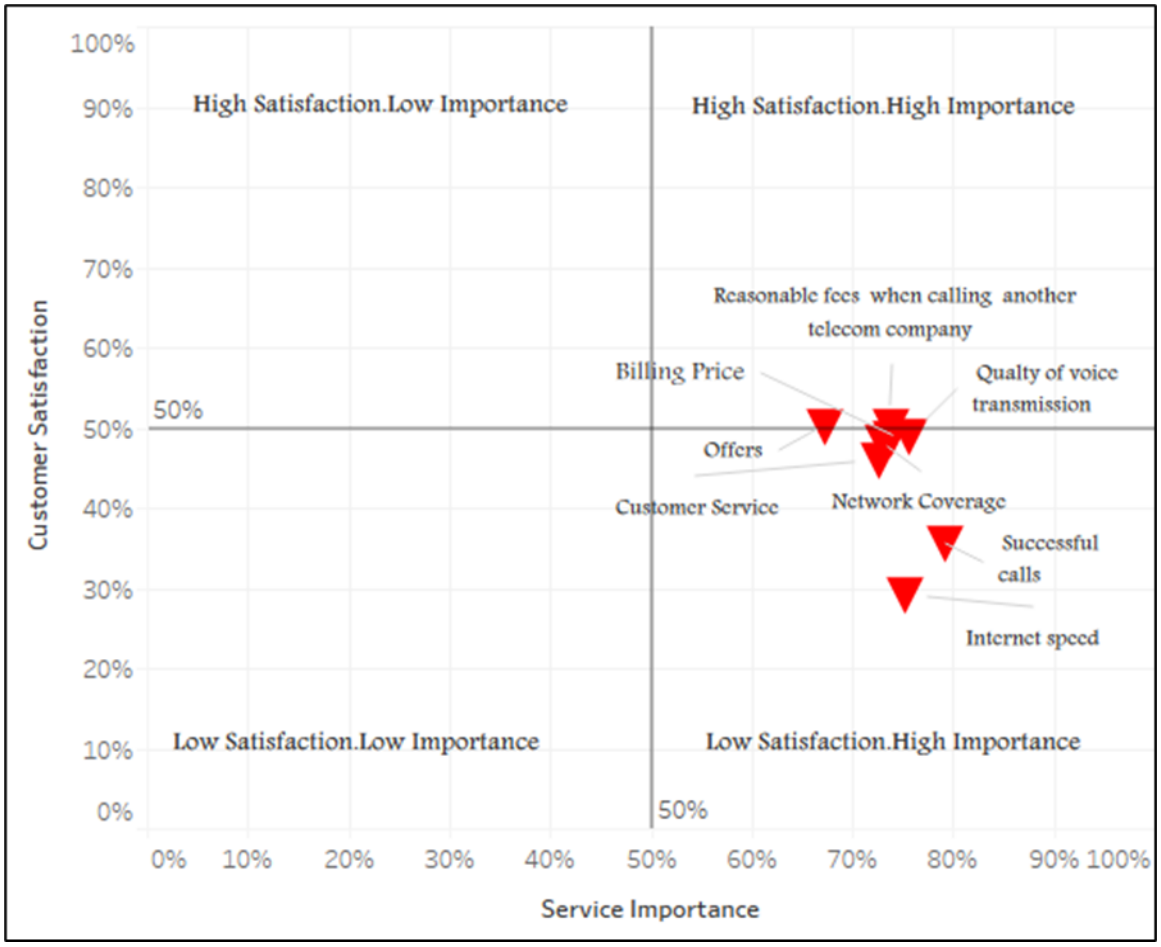

Figure 6.5: The importance VS. customer satisfaction for Zain customers.

Figure 6.6 shows the results for the Zain Company. Internet Speed received a Higher Importance service rating from the customers with 77.5% and the lowest customer satisfaction with 29.10%. This means this service needs to improve more than any other services in the Zain company. They need to plan for more improvement because 21.3% of the responses indicated this as a reason for changing the company in the open question in the questionnaire. In addition, the Customer Service and Successful Calls are situated in the Low Satisfaction area. At the same time, High Importance, with 70.4% importance and 45.92%, is allocated to Customer Satisfaction for Customer Service. Successful Calls scored the highest service importance rating with 79%, 35.43% for customer satisfaction and no responses to change the company in the open question in the questionnaire. As has been stated before in Chapter 4, the Zain company has attended to this poor service and is trying to raise mobile networks' efficiency by agreeing to increase their network infrastructure.

Offers scored a High Importance service rating from the customers' point of view, with 67%, and it received the highest customer satisfaction with 50.0%. In addition, the *Quality of voice transmission* service was rated with a high importance service with 74%, customer satisfaction 48.78% and no responses as a reason for changing the company. The improvement of this service depends on Zain's agreement to sell their towers to IHS Holding Limited (the largest cell tower operator in the markets of Europe, Middle East and Africa) and lease them back. *Reasonable Fees when calling another Telecom Company, and Billing Price* were rated at the same importance level of 74%. In comparison, the customer satisfaction towards the *Reasonable fees* service is a little bit higher at 50% than the satisfaction towards the *Billing Price* – 48.50%. Both services need improvement. This result is inconsistent with the questionnaire participants' responses about the reason for changing the company question. High Billing price was the reason most mentioned, with 29.5% responses.

In contrast, just one response mentioned the Unreasonable Fees as the reason for changing. *Network coverage* received 73% as an important service, 48.06% for customer satisfaction and 26.2% responses as the reason for changing. This problem has been recognised by the Zain company, and they agreed to increase their network infrastructure.

6.10 Summary

This Chapter proposes a hierarchical structure and compares it with the flat classification structure. The results showed that using the proposed hierarchical classifier improves the outcome, based on several different measures, for the complex problem of Multi-Way Arabic Sentiment Analysis compared to FC. The best accuracy and MSE for FC are 70.4% and 0.35, respectively. For HC, the best accuracy and MSE are 83.17% and 0.31.

Chapter 7: Customer Churn

7.1 Introduction

With the rising growth of the telecommunication industry, the customer churn problem has grown in significance as well. One of the most critical challenges in the data and voice telecommunication service industry is retaining customers, thus reducing *customer churn* by increasing *customer satisfaction*. To overcome the *delayed feedback* problem of traditional customer satisfaction questionnaires, new methods to extract real-time customer satisfaction feedback are needed. Therefore, this study offers a new approach to using social media mining to predict customer churn in the telecommunication field. This represents the first work using Twitter mining to predict churn in Telecom industries. A proposed SentiChurn model was developed to predict customer churn. The selection of the model's inputs was based on a literature review, questionnaire, and interview with an expert.

The newly proposed method proved its efficiency based on various standard metrics, and secondly based on a comparison with the ground-truth real outcomes provided by a telecom company. The proposed SentiChurn model proved its efficiency firstly based on various standard metrics; average precision for our model was 93.0%, the average recall was 97.0%, the average F1-score was 95.0%, and the model accuracy was 95.8%, and secondly based on a comparison with the ground-truth real and recent outcomes provided by a telecom company as 27% of customer churn rate. SentiChurn model predicted the customer churn for the same period as 31.6%, which is close to the actual rate. This Chapter answers the *RQ4*: Is it possible to predict the customer churn of telecommunication companies in Saudi Arabia by analysing customers' tweet?

7.2 Related Research

7.2.1 New Customer Churn Model Variables

I show here how our customer churn variables have been chosen. These data and parameters are presented here as gathered from three sources, sequentially: literature review, questionnaire and interviews with the telecom company experts (Table 7.1) and the customer satisfaction rate obtained from customer tweet mining. I can further divide the variables into two types: *independent variables* (predictors), all the variables

collected as inputs for the prediction model, and *dependent variable*, which represents the model outcome of the churn status variable. This section explains in detail where and why I use these specific variables based on literature.

Using customer demographics (age and gender) as churn predictors in the churn prediction model is common in the literature [101, 114], [20], [115], [18], [116], [23], [62], [91], [116], [132, 511]. Olle and Cai [91] found that young people below forty-five years of age are more likely to churn. Similar results were found by [94], [62]: customers between forty-five and forty-eight years old are more likely to churn.

Many researchers have studied the impact of a family or a friend leaving the same telecom company on a customer's churn decision [62], [24]. That is because of the increase in call price between two customers with different voice provider.

Consistent with this result, [511] showed that customers are more likely to churn if they have a social relationship with another customer who intends to or has already churned from the telecom company. This finding denotes that a company is at risk of churning if a customer's relationship leaves the company. Moreover, [444], [132], [466] used calling behaviour and network interaction (call length and number of calls) as churn predictors.

Some studies have realised the impact of social network information on churn prediction. For instance, [24] predicted customer churn by using customer information and their social network information. Their dataset was from the Pokec social network[25], and the call details of customers issued from the network for six months. They found that combining social network information with call log details improved the churn prediction. The same results were obtained by [78], who studied the impact of the social network on the prediction of customer churn. They combined call details from a social network with information about the customers. Moreover, [444] used a relational learner to increase the performance of the churn prediction model. They analysed calling behaviour and network interaction.

Different studies used the contract length as a churn predictor [20], [465], [94], [466], [62]. Balasubramanian and Selvarani [94] and [62] concluded that customers with contract lengths between twenty-five and thirty

[25] http://snap.stanford.edu/data/soc-pokec.html

months are more likely to churn. Many studies are related to contract length and overdue bills as churn predictors. Balasubramanian and Selvarani [94] found that customers with contract lengths between twenty-five and thirty months and four overdue bills are more likely to churn. In agreement with this result, [62] concluded that churning happens more for customers with contract lengths between twenty-five and thirty months and who have more than four overdue payments within six months. Mohanty and Rani [116] chose five attributes to predict churning, one of which also includes unpaid balances.

Most studies analysed using the customer call details as the primary churn predictor [511].

Coussement et al. [114] assessed the categorical and continuous data transformation in the performance of the churn prediction model. Their dataset was from a European telecommunication company. Some of the variables they selected were the number of minutes for outgoing calls and contacts with the call centre. In addition, [20] compared some techniques used in churn modelling. Their dataset was from a UK mobile telecommunication company. They included several variables, one of which was call usage detail.

In 2016, [93] proposed a model for churn prediction for telecommunication companies. They used historical records related to the telecom company. The attributes included phone and call details. Forhad et al. [115] applied the rule-based classification to predict whether a customer is likely to churn or not. Their dataset contained customer information such as call details (billing information and length of calls).

Furthermore, [11] applied different data mining techniques to predict customer churning. They applied their methodology to the online dataset from Kaggle. They used fourteen attributes, including call details, customer service calls and phone number. Chen et al. [18] built a churning prediction model for a mobile telecommunication company. They used two datasets: customer information and statistical data, which contained call length and complaint information. Mohanty and Rani [116] assessed many techniques to predict customer churning and used the dataset from an Indian telecommunication company. They chose five attributes to predict churning: customer dissatisfaction and satisfaction, switching costs, quality of services, service usage in terms of used minutes in calls, call details, and unpaid balances. They also used customer-related variables, such as customer gender, customer status or whether a customer is an active user. Tiwari et al. [117] concluded that customers with no active plans and no incoming and outgoing calls within six months

are likely to churn.

In addition, [119] predicted customer churning in the telecommunication industry based on rough set theory. They used historical data on a publicly available dataset and found some essential attributes in the customer churn prediction, such as evening minutes, customer service calls and day minutes. Keramati et al. [465] proposed a prediction model for a customer churn by using different data mining techniques. They used customer information, such as contract length, customer complaints and call details.

Hudaib et al. [512] used three hybrid models over two stages: data clustering and churning prediction. They collected the three-month call data of customers of a Jordanian telecommunication company. Wei and Ghiu [63] predicted customer churn according to the call details and contract information gathered from interviews with telecom experts. Singh and Singh [23] proposed a model for predicting high-value customers and customers' churner. They used customer information, such as age, sex and call details. Numerous studies recognised the importance of including customer complaints as an attribute in their churn prediction model [465], [117], [116], [18], [115], [64], [20], [513], [91], [23], [466], [132], [106], [116, 138].

After reviewing the literature, I listed the most common techniques in Table 3. As shown in the literature, decision trees and logistic regression are the most common techniques used in churning prediction models. A decision tree offers a graphical representation of the relations between churning variables [98]. CART or CHAID are examples of the algorithms used to develop a decision tree [514]. Both logistic regression and decision tree are effective and easy techniques to predict churning and analyse the characteristics that cause a churn [93], [515], [96], [94]. However, there are some disadvantages in using a decision tree, such as being affected by the complex relations between the variables [516]. The following technique commonly used in the literature is a neural network, which has some limitations, including its need for an extensive dataset and extensive time consumption in training [93]. Support vector machine and naïve Bayes were likewise used.

Num.	Customer Churn Variables	Description	Type of Variable	Range
1	Age	Age group has been identified	Ordinal variable	18–24, ‘1’
				25–34, ‘2’
				35–44, ‘3’
				45–54, ‘4’
				55–64, ‘5’
				65+, ‘6’
2	Gender	Male or Female	Binary variable	Male, ‘0’
				Female, ‘1’
3	Has a relation at the same telecom company	Does the customer have a family member who used the same telecom provider as he/she did?	Binary variable	Yes, ‘1’
				No, ‘0’
4	Overdue bill	Does the customer have an unpaid bill?	Binary variable	Yes, ‘1’
				No, ‘0’
5	Long period	Contract length in month from start day of contract until June 2017	Ordinal variable	$\underline{\geq}$ 1, ‘1’
				1 $\underline{\geq}$ 5, ‘2’
				5 $\underline{\geq}$ 10, ‘3’
				10+, ‘4’
6	New customer	Has the customer used a telecom provider recently?	Binary variable	Yes, ‘1’
				No, ‘0’
7	Inactive	Is the customer active?	Binary variable	Yes, ‘1’
				No, ‘0’
8	Low data	Does the customer have low data usage?	Binary variable	Yes, ‘1’
				No, ‘0’
9	Low talk	Does the customer make few phone calls?	Binary variable	Yes, ‘1’
				No, ‘0’
10	No Internet & Talk & SMS	Does the customer not use the Internet, phone calls and short message service?	Binary variable	Yes, ‘1’
				No, ‘0’

11	No value-added service	Does the customer use any of the non-core services?	Binary variable	Yes, '1'
				No, '0'
12	Customer satisfaction	Percentage of customer satisfaction from Twitter analysis	Continuous variable	
13	Churn status	Does the customer churn?	Binary variable	Churner/Non-churner Churner, '1'
				Non-churner, '0'

Table 7.1: Details of The Customer Churn Variables.

7.3 Methodology

The two types of known customer churning are voluntary and involuntary [517]. The decision of a customer to move to another telecom company on their own is called voluntary, while a customer ceasing telecom company services for reasons outside their influence, such as death or change of the customer's job, is called involuntary [65]. Usually, the literature is interested in voluntary customer churning because it describes the relationship between a customer and a company. There are two types of customer payment schemes: post-paid and pre-paid [518]. Post-paid customers receive a monthly bill for company services, while a pre-paid customer is charged in advance for company services.

In this study, a *churner* is defined as a post-paid customer who voluntarily leaves the company and stops telecom services within our time window. By contrast, a non-churner in our study is a post-paid customer who remains with the company within our time window.

Data mining refers to knowledge discovery from a large database [519]. The three most common data mining methodologies used to develop data mining models are Knowledge Discovery Databases (KDD) [520], Cross-Industry Standard Process for Data Mining (CRISP-DM) [521], [4] and Sample, Explore, Modify, Model, Assess (SEMMA), which was created by the SAS Institute (Inc SI. SAS version 9.1., 2005). The literature review indicated that KDD and CRISP-DM are more widely used than SEMMA [522], [523]. Although KDD includes nine phases and CRISP-DM has six phases, their phases are equivalent [522].

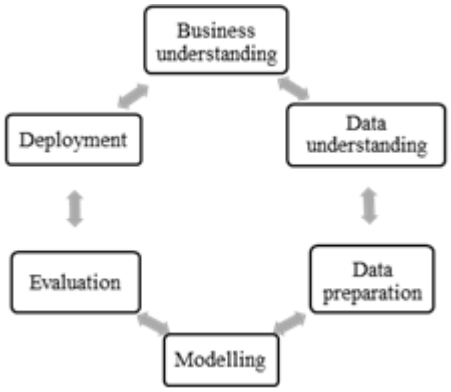

Figure 7.1: The CRISP-DM approach (based on [4]).

I adopted some of the steps of CRISP-DM [4] that suit our task to develop our churn prediction model (SentiChurn model, Figure 7.2) because CRISP-DM is appropriate for a business domain [524]. The six phases of CRISP-DM are shown in Figure 7.1.

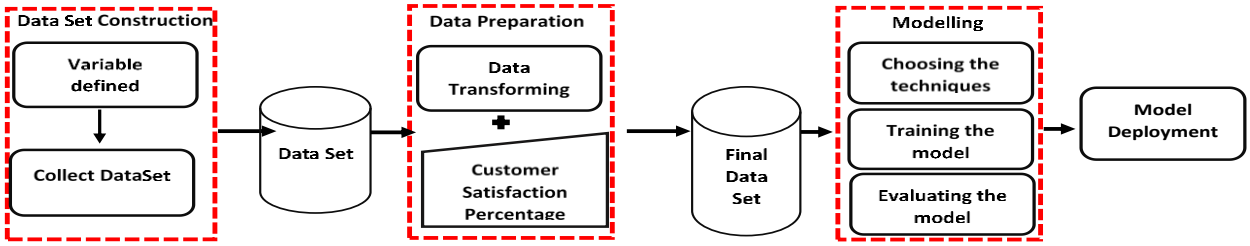

Figure 7.2: Our SentiChurn Model Approach.

Defining the variables entails collecting the variables from the sources (Figure 7.2). To determine the possible variables that can differentiate between the behaviour of churners and non-churners, variables were collected from three sources (Table 7.1). First, I collected variables from the literature review. Some variables found in the review were disregarded because of the difficulty of obtaining them from telecom companies due to privacy concerns, such as name, phone number and code, call details and billing information. This is the case with many prediction model systems in other countries [117]. Next, I conducted a survey via questionnaire with the telecom customers. The questionnaire aimed to test the relationship between the collected variables and churning behaviour from a customer's perspective. Afterwards, I conducted an informal interview with a Saudi telecom expert (a telecom business consultant) to show here

the collected variables and question him about other variables from the company's point of view. The telecom company divides its customers into segments based on their own selected set of variables and calculates the churn rate for each segment quarterly, half-yearly and annually. They propose that the variables for one segment have higher churn rates half-yearly because a higher churn rate must be obtained to train the prediction model. Based on the literature review results, questionnaire, and interview, I collected some variables that could help us predict customer churning and differentiate between churners and non-churners. The company provided us with historical data from two years ago to maintain customer privacy about their current customers. According to its request, the company name has been withheld and is called in the rest of the document 'the Company'.

Data preparation includes data description, data transformation and initialization of the dataset model (Figure 7.2). In training the model, an appropriate data mining algorithm (G. Modelling) the model in the training set is trained to address the problem.

In model evaluation, the model is evaluated on the test set by using the performance measures. In the model deployment stage, the prediction result is presented to the company for evaluation from a real-world and company perspective.

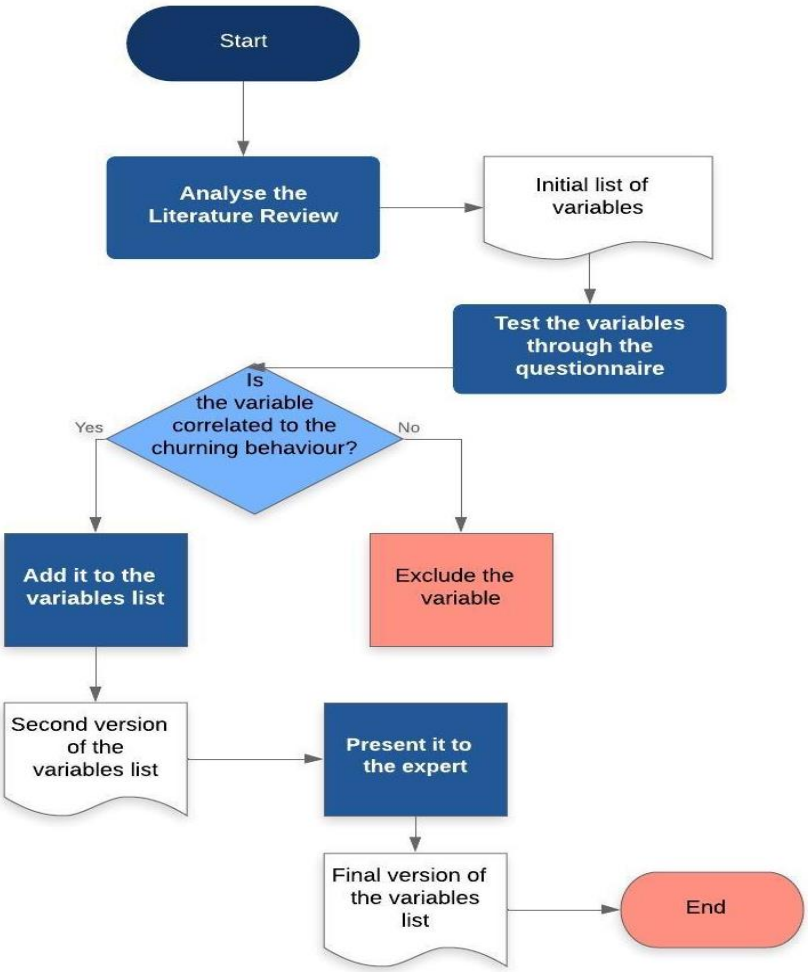

Figure 7.3: Workflow to develop the customer churn variables.

7.3.1 Data Set Construction

The dataset has been constructed from historical data that the Company provided and the customer satisfaction rate measured through Twitter mining [336]. I collected a sample of 100,000 customers' data from the Saudi telecom Company. From this figure, 27,000 were churners while 73,000 were non-churners. These historical data of customers were collected randomly within six months, from January 2017 to June 2017. Earlier studies differed in setting the time window for churning analysis and prediction. For instance, [113] proved that a customer mood on Twitter could be a predictor for churning three months later. In addition, [512] collected the three-month call data of customers from a Jordanian telecommunication

company. Their results agreed with those found by [525] that two to three months is a sufficient time window to prepare a strategy for retaining customers and preventing churning.

On the contrary, [91] stated that four months is needed to predict a customer churning based on his/her dissatisfaction. However, [13] increased this to a five-month collection of tweets as a dataset to predict their customer growth model. Other studies set even six months as the time window for churn prediction [94], [20], [24], [62]. Tsai and Lu [92] found that a customer should be with a company for six months or longer to have an accurate prediction model.

Thus, our selected time window is adequate to conform to even the strictest previous studies. I take [92]'s suggestions into account, as I agree that a customer could become resentful but may take a more extended period to carry out the churning action. Thus, I can consider that, as our dataset is from January 2017 to June 2017, the churning can only be estimated between July and December 2017 (Figure 7.4).

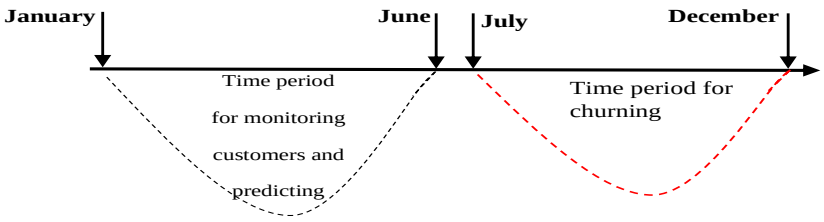

Figure 7.4: Time window of the prediction.

7.3.2 Historical Data Set Preparation

The variable data type is transformed in the dataset preparation step, and the binary data are normalised. The goal of data preparation is to help the model deal with data easily [114]. The binary variable is normalised to '1' for 'yes' and '0' for 'no' and '0' for 'Male' and '1' for 'Female'. Regarding the continuous variables,

such as age and long period as a customer, I transform them into categories as an ordinal variable and then assign them by sequential numbering starting from 1. The final collected variables and their types that will be used as inputs for our prediction model are listed in Table 7.1. The dataset captures the features of the population under study. The outcome from this step is the final dataset that will be used to train the model (Figure 7.5).

Out[183]:

	Age	Gender	Unpaid_Bill	Has_Family	long_period	Inactive	Low_data	Low_talk	No_Int_Talk_SMS	No_Vas	CS	New_Customer	Churn_status
0	4	0	1	0	4	0	0	0	0	0	32	0	0
1	6	0	1	0	4	0	1	0	1	0	32	0	0
2	2	1	0	1	2	1	1	1	1	1	32	0	1
3	3	0	1	0	4	0	0	0	0	0	32	0	0
4	5	0	1	0	4	0	0	0	0	0	32	0	0

Figure 7.5: Final Data set after Preparation.

7.4 Modelling

7.4.1 Performance Evaluation Metrics

There are useful metrics that should be used to assess the model's performance and compare it with a benchmark. Numerous churning prediction studies used specific performance metrics, such as precision, recall, F1, accuracy, confusion matrix, specificity, sensitivity, area under the curve (AUC) and receiver operating characteristic curve (ROC). A confusion matrix is a tool used with binary classification; it compares the actual Positive and Negative and the predicted Positive and Negative. It uses the True Negative (TN), True Positive (TP), False Negative (FN) and False Positive (FP) [472] as follows:

- FP: indicates that our model predicts the customer is a churner, but the customer is a non-churner.
- FN: indicates that our model predicts the customer is a non-churner, but the customer is a churner.
- TP: indicates our model correctly predicts the customer is a churner.
- TN: indicates our model correctly predicts the customer is a non-churner.

There are other metrics used in addition to TP, TN, FN and FP, such as sensitivity, specificity and accuracy. The weakness of an accuracy measure originates from overusing the sensitivity and specificity measures [14]. Sensitivity is equal to recall. Meanwhile, specificity is the ratio of the negative correctly predicted as shown in the following equation:

$$\text{Specificity} = \frac{\text{TN}}{(\text{TN} + \text{FP})} \qquad (7.1)$$

High sensitivity is more preferred than high specificity in telecom providers because the cost of an untrue classification of a non-churner is less than the cost of an untrue classification of a churner [14].

Some churning prediction studies prefer to evaluate model performance by using ROC and AUC because these curves' ability to remain the same with imbalanced data, even if the positive and negative instances are changing [14].

ROC is a two-dimensional curve drawn to show the relation between TP, the churner correctly predicted, and FP, the non-churner incorrectly predicted as a churner [526]. The best model performance occurs when the ROC is close to (0,1). A better model performance also has a higher AUC.

Moreover, I used a cross-entropy/logarithmic loss (log loss) as a loss function; both calculate the same in the classification problem. The loss function is an error metric to measure uncertainty. It is one of the measures used for evaluating the performance of a binary classifier from the probability estimation between 0 and 1. Log loss penalises both types of errors, especially those predictions where the confidence is inaccurate. If the log loss is closer to zero, then this indicates the good performance of the model.

Using the log loss provides us with an accurate view of our performance model based on the prediction of probabilities, not only the output.

$$Hp(q)^n = \frac{1}{N} \sum_{i=1}^{n} y_i \log((P(y_i) + (1 - y_i) \log\big(1 - P(y_i)\big) \qquad (7.2)$$

where N is the number of items on the training set; $\frac{1}{N}$ is the probability of each class; log is the natural logarithm; y is the binary label, which is either 0 or 1; and P(y*)* is the probability predicted of the class.

7.4.2 SentiChurn Churn Modelling Technique

I used the proposed model that explained in detail in Chapter 5.

7.4.3 Training the Model

Given that an overlap exists between a churner and a non-churner, the threshold 'cut-off' must be defined. Usually, the threshold is set as fifty per cent. Any probability right of the threshold has the most specificity, while any probability left of the threshold has the most sensitivity, as shown in Figure 7.6.

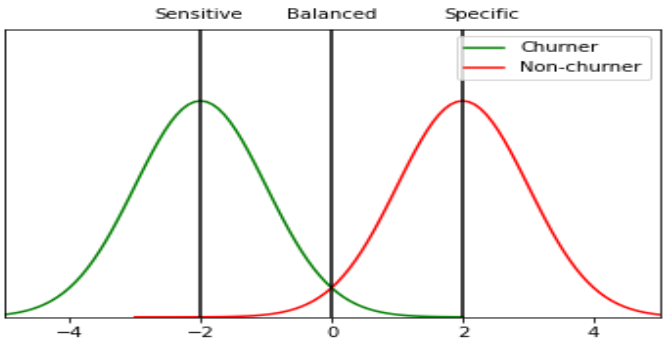

Figure 7.6: Threshold setting between churner and non-churner.

The dashed line in Figure 7.7 is the threshold. Any probability under the threshold means higher sensitivity, with more churners correctly predicted and better model performance, whereas any probability under the threshold means higher specificity, with more non-churners incorrectly predicted and worse model performance. The closer curve to the top left corner (0,1) denotes the better prediction power of the model. The ROC of the class' 'churner' and 'non-churner' is 0.97; this denotes the power of our prediction model performance.

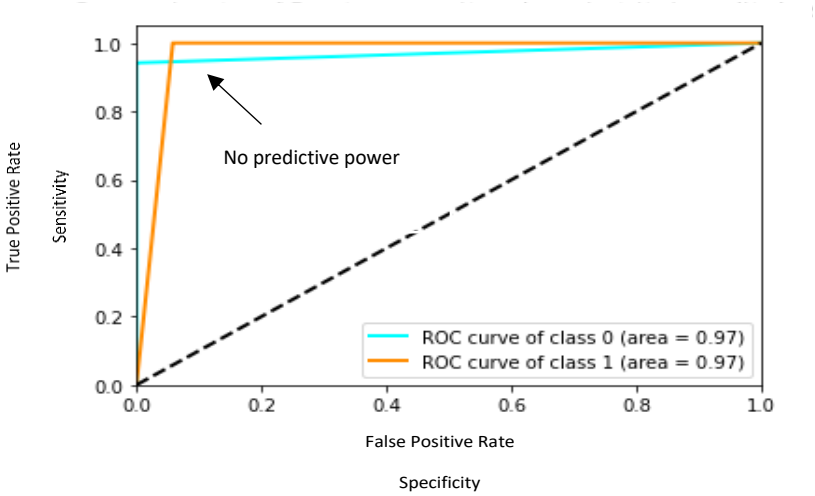

Figure 7.7: ROC result for SentiChurn model.

	Precision	Recall	F1-score
Non-Churner	1.00	0.94	0.97
Churner	0.87	1.00	0.93
macro average	0.93	0.97	0.95
weighted average	0.96	0.96	0.96

Table 7.2: The Classification report.

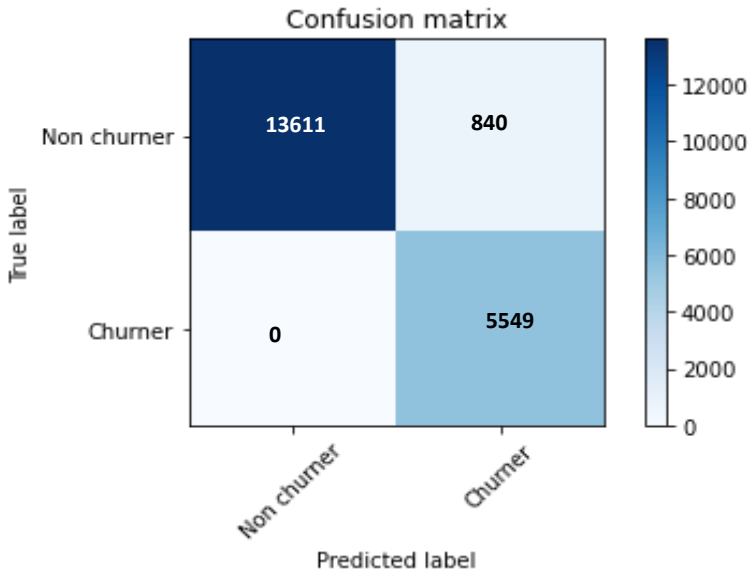

Figure 7.8: Confusion Matrix.

The classification report in Table 7.2 denotes the performance model, where the average metrics precision for both classes is 0.93, the average recall for both classes is 0.97, the average F1-score for both classes is 0.95, and the model accuracy is 95.8%.

In the confusion matrix (Figure 7.8), 13,611 non-churner customers were correctly predicted as non-churners by our model. Furthermore, 5,549 churner customers were correctly predicted as churners by our model, 840 non-churner customers were predicted as churners by our model and no churners were predicted as a non-churner customer by our model.

The log loss score is 0.1, which means our model is fine. Figure 7.9 shows the probability distribution (x) with the log loss(y) and the distribution between the actual and predicted values.

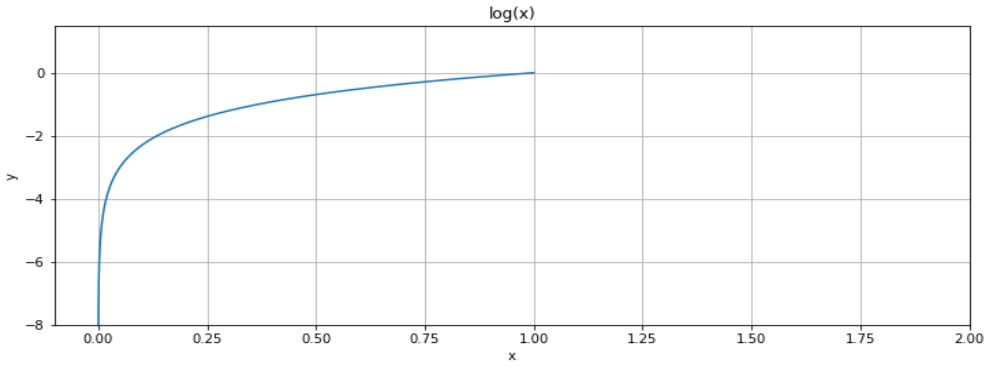

Figure 7.9: Log loss score versus Probability distribution.

7.4.4 Evaluating the Model

I evaluate the model by using the performance evaluation metrics and validating the percentage of customer churn than our model predicted versus that provided by the company.

The company presented a customer churn percentage of 27% from January 2017 to June 2017. The model predicted the customer churn for the same period as 31.6%, which is close to the actual percentage.

The model predicts the customer churn percentage based on the following equation:

$$cust\_churn = total\_churner/num\_customers \times 100 \qquad (7.3)$$

Where total_churner is the total number of churners in the dataset, and num_customers is the total number of all the customers in my dataset.

After validating the customer churn percentage by using the historical data of customers and the customer satisfaction percentage predicted by Twitter mining, I answered the RQ4, 'Is it possible to predict the customer churn of telecom companies in Saudi Arabia by analysing customers' tweets?'

7.5 Summary

In this chapter, I proved the third research hypothesis: "The customer churn of telecom companies in Saudi Arabia can be predicted by analysing customer satisfaction in Twitter" by building our prediction model. The selection of the model's inputs was based on a literature review, questionnaire, and interview with an expert. The proposed SentiChurn model proved its efficiency firstly based on various standard metrics; average precision for our model was 93.0%, the average recall was 97.0%, the average F1-score was 95.0%, and the model accuracy was 95.8%, and secondly based on a comparison with the ground-truth real and

recent outcomes provided by a telecom company as 27% of customer churn rate. SentiChurn model predicted the customer churn for the same period as 31.6%, which is close to the actual rate. The next chapter will conclude the recommendation for the company based on the last experiments.

Chapter 8: Conclusions and Future Work

This Chapter reviews the main contributions of the thesis concerning the research hypothesis and raises questions regarding future work.

8.1 Thesis Summary and Contributions

The flexibility in mobile communications allows customers to switch from one service provider to another quickly, making customer churn one of the most critical challenges for the data and voice telecommunication service industry. Churn prediction models are required to avoid facing significant losses. Customer satisfaction is a popular topic in marketing literature, as much research correlates customer satisfaction with customer loyalty [11, 13, 14]. Customers who are satisfied with a company's services make a company more valuable because the cost of attracting new customers is five times greater than retaining existing customers [15], [11], [14].

An evaluation of associated literature presented in Chapter 2 shows that many studies have been carried out to create useful models to predict churners. While these models suffer in being applied post-factum and lack real-time analytics to target customers efficiently [23], there is a demand for such improved customer churn models. The delay can cause drops in market position, particularly for companies with a vast customer community spreading across various time zones; monitoring daily data is challenging. Therefore, real-time methods are needed to solve the delayed acquisition of feedback and create efficient maintenance strategies.

The thesis hypotheses were: *If measurable criteria for customer satisfaction are defined, they could extract services that do not meet the expectations of customers. Customer satisfaction with telecom companies in Saudi Arabia can be monitored by analysing microblogging sites. The customer churn of telecom companies in Saudi Arabia can be predicted by analysing microblogging sites.*

To examine these hypotheses and accomplish the objectives as presented in Section 2.1, several steps were essential. This work's main aim has been to capture user satisfaction with telecom companies by mining microblogging sites and using these insights to apply social media mining techniques, to develop a useful

churn prediction model, taking into consideration language and location factors, and to use those data to make recommendations for these telecom companies.

This research aims to contribute to the Arabic sentiment analysis community by addressing the deficiency of Saudi dialect corpora and sentiment lexicon for SA. Developing natural language processing tools for Arabic text requires an understanding of the unique internal structure of Arabic [38].

To achieve the aims of this research, a deductive approach is used. Deductive research involves adopting hypotheses and testing them in a causal manner [50] to explore the research variables' relationships. A literature review determined that there are distinct gaps in analysing of social media sentiments to predict customer churn. Accordingly, this research addressed the relationship between the variables of a Twitter sentiment analysis, customer satisfaction and customer churn prediction using a systematic review of the literature. The results showed a need for an Arabic sentiment analysis tool and resources freely available and specific to Dialectical Arabic. In addition, I identified a **requirement for a real-time based churn prediction model** that takes into consideration the language and location factors; the above investigation point is the first contribution of this study.

To address the Arabic sentiment analysis requirement, I have constructed, cleaned, pre-processed and annotated a 20,0000 Gold Standard Corpus (GSC) to create **AraCust, the first Telecom GSC for Arabic Sentiment Analysis DA**, as explained in Chapter 3. AraCust contains Saudi dialect tweets, processed from a self-collected Arabic tweets dataset, and it has been annotated for sentiment analysis. Additionally, AraCust's power is illustrated by performing an exploratory data analysis to analyse the features sourced from the AraCust corpus's nature to choose the suitable ASA methods. In addition, using a corpus-based approach, the **AraSTw lexicon** has been manually created (Chapter 3). I evaluated the AraSTw lexicon using a simple lexicon-based approach, where negative and positive words were counted to define each tweet sentiment. I used one internal data set, our AraCust corpus, and one publicly available Arabic dataset created from the Twitter Arabic Sentiment Tweet Dataset (ASTD). The AraSTw lexicon outperformed the accuracy on ASTD by 44.7%. Also, it outperformed by 1.11% the accuracy on AraCust. The AraCust corpus and AraSTw lexicon have been released online for free to the research community. These resources are the second contribution of this study.

To make recommendations to the telecom companies, I first extracted a taxonomy of metrics based on the current literature, and then I evaluated them using a questionnaire with the telecom customers. The questionnaire aimed to test the metrics and the relationship between the collected variables and churning behaviour from a customer's perspective, as explained in Chapter 4.

Secondly, Twitter sentiment analysis has been examined using shallow machine learning, deep learning models and transformer networks. The machine learning algorithm, SVM, was used to test both baseline and corpus-based features. Then, I used the two most popular deep learning-based models, LSTM and GRU, with two different implementations: simple LSTM and GRU, and bidirectional LSTM and GRU, with a different setting. This was defined as the most appropriate model for the ASA and the telecommunication corpus. After that, I utilised three different transformer networks, RoBERTa, AraBERT and hULMonA.

Based on the results, I developed a new prediction model that fills the detected gaps in the ASA literature and fits the telecommunication field. The proposed model proved its effectiveness for Arabic sentiment analysis and churn prediction. The taxonomy of metrics has been used to measure telecom services' satisfaction and visualise the importance of the metrics (services) to set the recommendations.

8.2 Answers to the Research Questions

In the following, each research question is separately discussed in terms of how this thesis answered it.

RQ1. What are the traceable, measurable metrics for customers' satisfaction with telecom companies in Saudi Arabia and how can they be combined for visualisation?

I defined customer satisfaction metrics using a report by the Saudi Communications and Information Technology Commission [438], related research (Chapter 2) and the tweet annotation process (Chapter 3). Then, I evaluated the importance of these metrics using statistical analysis for the responses obtained by the questionnaires from the customers' point of view. The final metrics are: 'network coverage', 'quality of voice transmission', 'customer service', 'successful calls', 'billing price, 'good offers', 'reasonable fees when calling another telecom company' and 'internet speed'. I analysed the importance of each metric for each telecom company using RII. The taxonomy of the measurable metrics of customer satisfaction and their relationship with customer churn is shown in Chapter 4.

What type of services for customers of telecom companies in Saudi Arabia are mentioned in tweets, and what is the sentiment of customers about these services?

The listed services mentioned in the tweets are Network Coverage, Phone Network, Quality of voice transmission, Customer Service, Successful Calls, Billing Price, Good Offers, Reasonable Fees when calling someone who uses another Telecom Company, Browsing Speed, and Hiring section. The Network Coverage and Phone Network were merged as they pointed to the same service. Additionally, I merged Internet Speed and Browsing Speed. I excluded the hiring section because it is out of scope for this research.

I applied the proposed hierarchical model on the AraCust1 to calculate the CS percentage toward each company's service, Chapter 6. After that, I used the tableau software to visualise the service importance versus customer satisfaction toward the service. The service importance was obtained from the questionnaire analysis in Chapter 4. I visualised this correlation to set the recommendations for the decision-makers of the three telecom companies.

Can we automatically measure and make automatic predictions about customers' satisfaction with telecom companies in Saudi Arabia using Twitter?

Several ASA experiments were carried out on AraCust (Chapter 5), starting by applying SVM as a baseline model to choose the best feature sets with many feature selections experiments. Next, I compared two state-of-the-art deep learning models, LSTM and GRU, with different embeddings and settings, on AraCust, to define the best model for AraCust corpus and Arabic dialect characteristics. After that, three transformer networks, AraBERT, hULMonA and RoBERTa models, were utilised on AraCust to define the best performance suitable to the corpus and the dialect Arabic characteristics. Finally, I developed a model combining the AraBERT model and Bi-GRU to predict customer satisfaction for the three companies.

The results proved that the prediction model is highly accurate when comparing with other models, and it achieves the prediction aim of accurately comparing with the actual results provided by the telecommunication customers.

Is it possible to predict the customer churn of telecom companies in Saudi Arabia by analysing customers' tweets?

I proposed a **new variable input to the churn prediction model (SentiChurn)**, customer satisfaction percentage, calculated using SA, as shown in Chapter 7. The proposed model proved its efficiency based on various standard metrics: the average precision for the model was 93%, the average recall was 97%, the average F1-score was 95%, and the model accuracy was 95.8%. The efficiency was based on comparing the actual and recent outcomes provided by a telecom company as 27% of customer churn rate. The SentiChurn model predicted the customer churn as 31.6%, which is close to the actual rate.

8.3 Limitations and Future work

One of the limitations of this research was that the proposed method depended on one social media platform only, Twitter. This may mean that, as not all the telecommunication subscribers used social media to express their feelings towards their telecom company, some customers are therefore not included in this study as they may have used other means to communicate with or complain about their telecom companies. The reason for this was one of practicality – at the start of this research, other social media platforms, were considered; however, their unavailability for research purposes limited my choices.

In addition, STC customers are greater in number than other telecommunication companies' customers in Saudi Arabia because it is the primary (most popular) Saudi telecommunication company in Saudi Arabia, making the dataset unbalanced. However, balancing techniques were applied to counter this issue to the extent possible.

Next, the availability of customer data is limited to one company due to other telecommunication companies' privacy issues, so the SentiChurn model is applied only to one telecommunication company.

The future work includes:

- **Expanding the Arabic dialect resources**

  This work aimed to build further the gold standard Saudi Corpus and Saudi lexicon for ASA, as presented in Chapter 3. This is recommended by some studies that have identified the need for comprehensive Arabic dialect resources to combine their morphological analysis and tokenisation into one process. Doing so may resolve the Arabic tokenisation issue and improve the process when conducting SA and opinion mining [28], [34].

- **Better handing for Arabic language in ASA tool**

I recommend creating a specific corpus for the language valence shifters and negation with their ranks. This recommendation demands expert Arabic linguists' association with a computer specialist to give the best ASA corpus.

8.4 Broader applicability of this work

In this work, the ASA model was considered when proposing solutions for the CS and CC problem in the telecom sector. Different fields, such as education, have different features, making applying my approach is interesting because the proposed approach based on text-mining. Investigations could be made into how this work can be generalised to other fields and what changes or improvements are needed to enhance this work's recommended solutions to be enhanced.

Bibliography

1. Landis, J.R. and G.G. Koch, *An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers.* Biometrics, 1977. **33**(2): p. 363-374.
2. Comission, C.I.T., *Individuals and families report: Results of a market survey* 2019: Saudi Arabia. p. 1-102.
3. PRISMA. *PRISMA: TRANSPARENT REPORTING of SYSTEMATIC REVIEWS and META-ANALYSES*. 2015 [cited 2019 30 Nov]; Available from: http://prisma-statement.org/PRISMAStatement/FlowDiagram.aspx.
4. Chapman, P., J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth, *CRISP-DM 1.0: Step-by-step data mining guide.* SPSS inc, 2000. **9**: p. 13.
5. Kim, M.-K., M.-C. Park, and D.-H. Jeong, *The effects of customer satisfaction and switching barrier on customer loyalty in Korean mobile telecommunication services.* Telecommunications policy, 2004. **28**(2): p. 145-159.
6. Duwairi, R.M. and I. Qarqaz. *Arabic sentiment analysis using supervised classification*. in *2014 International Conference on Future Internet of Things and Cloud*. 2014. IEEE.
7. Sivo, S.A., C. Saunders, Q. Chang, and J.J. Jiang, *How low should you go? Low response rates and the validity of inference in IS questionnaire research.* Journal of the association for information systems, 2006. **7**(1): p. 17.
8. Qamar, A.M., S.A. Alsuhibany, and S.S. Ahmed, *Sentiment classification of twitter data belonging to saudi arabian telecommunication companies.* International Journal of Advanced Computer Science and Applications (IJACS), 2017. **1**: p. 395-401.
9. Najadat, H., A. Al-Abdi, and Y. Sayaheen, *Model-based sentiment analysis of customer satisfaction for the Jordanian telecommunication companies.* 2018 9th International Conference on Information and Communication Systems, ICICS 2018, 2018: p. 233-237.
10. Fitri, F.S., M. Nasrun, and C. Setianingsih. *Sentiment analysis on the level of customer satisfaction to data cellular services using the naive bayes classifier algorithm*. in *2018 IEEE International Conference on Internet of Things and Intelligence System (IOTAIS)*. 2018. IEEE.
11. Ali, M., A.U. Rehman, S. Hafeez, and M.U. Ashraf. *Prediction of Churning Behavior of Customers in Telecom Sector Using Supervised Learning Techniques*. in *2018 International Conference on Computer, Control, Electrical, and Electronics Engineering (ICCCEEE)*. 2018. IEEE.
12. Bhatnagar, V., *Data mining and analysis in the engineering field*. 2014: IGI Global.
13. Ranjan, S., S. Sood, and V. Verma. *Twitter Sentiment Analysis of Real-time Customer Experience Feedback for Predicting Growth of Indian Telecom Companies*. in *2018 4th International Conference on Computing Sciences (ICCS)*. 2018. IEEE.
14. Li, P., S. Li, T. Bi, and Y. Liu. *Telecom customer churn prediction method based on cluster stratified sampling logistic regression*. in *In International Conference on Software Intelligence Technologies and Applications & International Conference on Frontiers of Internet of Things*. 2014. IET.
15. Kotler, P., *Marketing management*. 2009: Pearson education.
16. Deng, Z., Y. Lu, K.K. Wei, and J. Zhang, *Understanding customer satisfaction and loyalty: An empirical study of mobile instant messages in China.* International journal of information management, 2010. **30**(4): p. 289-300.

17. De Caigny, A., K. Coussement, and K.W. De Bock, *A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees.* European Journal of Operational Research, 2018. **269**(2): p. 760-772.
18. Chen, Y.-b., B.-s. Li, and X.-q. Ge. *Study on predictive model of customer churn of mobile telecommunication company*. in *2011 Fourth International Conference on Business Intelligence and Financial Engineering*. 2011. IEEE.
19. Mahajan, V. and R. Mahajan, *Variable Selection of Customers for Churn Analysis in Telecommunication Industry.* International Journal of Virtual Communities and Social Networking (IJVCSN), 2018. **10**(1): p. 17-32.
20. Hassouna, M., A. Tarhini, T. Elyas, and M.S. Abou Trab, *Customer Churn in Mobile Markets: A Comparison of Techniques.* International Business Research, 2015. **8**(6).
21. Berson, A., S. Smith, and K. Thearling, *Building data mining applications for CRM*. 1999: McGraw-Hill Professional.
22. Kentrias, S., *Customer relationship management: The SAS perspective.* Retrieved March, 2001. **24**: p. 2011.
23. Singh, I. and S. Singh, *Framework for Targeting High Value Customers and Potential Churn Customers in Telecom using Big Data Analytics.* International Journal of Education and Management Engineering, 2017. **7**(1): p. 36-45.
24. Pagare, R. and A. Khare. *Churn prediction by finding most influential nodes in social network*. in *2016 International Conference on Computing, Analytics and Security Trends (CAST)*. 2016. IEEE.
25. Market.us. *Global Number of Daily Active Social Networking Sites Users in Million*. 2020 [cited 2020 1/11]; Available from: https://market.us/statistics/social-media/.
26. Marcus, A., M.S. Bernstein, O. Badar, D.R. Karger, S. Madden, and R.C. Miller, *Processing and visualizing the data in tweets.* ACM SIGMOD Record, 2012. **40**(4): p. 21-27.
27. Medhat, W., A. Hassan, and H. Korashy, *Sentiment analysis algorithms and applications: A survey.* Ain Shams engineering journal, 2014. **5**(4): p. 1093-1113.
28. Ravi, K. and V. Ravi, *A survey on opinion mining and sentiment analysis: tasks, approaches and applications.* Knowledge-Based Systems, 2015. **89**: p. 14-46.
29. Sohangir, S., D. Wang, A. Pomeranets, and T. Khoshgoftaar, *Big Data: Deep Learning for financial sentiment analysis.* Journal of Big Data, 2018. **5**(1): p. 3.
30. Abdulrahman A., A.-J. *Designation Of (STC), (Mobily) And (Zain) As Dominant Service Providers In The Wholesale Mobile Call Termination Services Marke*t. 2010; Available from: https://www.citc.gov.sa/ar/Decisions/Pages/283-1431.aspx.
31. Saudi InformationTechnology Commission, *Communication Service Provider Rating Index*. 2019.
32. STC, *Investor Relations*. 2020.
33. Analyzer, C., *State of Social Media*. 2018.
34. Masmoudi, A., M.E. Khmekhem, Y. Esteve, L.H. Belguith, and N. Habash. *A Corpus and Phonetic Dictionary for Tunisian Arabic Speech Recognition*. in *LREC*. 2014.
35. Syiam, M.M., Z.T. Fayed, and M. Habib, *An intelligent system for Arabic text categorization.* International Journal of Intelligent Computing and Information Sciences 2006. **6**(1): p. 1-19.
36. Mubarak, H. and K. Darwish. *Using Twitter to collect a multi-dialectal corpus of Arabic*. in *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*. 2014.

37. Masmoudi, A., M.E. Khmekhem, Y. Esteve, L.H. Belguith, and N. Habash. *A Corpus and Phonetic Dictionary for Tunisian Arabic Speech Recognition*. in *In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*. 2014. Reykjavik, Iceland: European Language Resources Association (ELRA).
38. Albraheem, L. and H.S. Al-Khalifa. *Exploring the problems of sentiment analysis in informal Arabic*. in *Proceedings of the 14th international conference on information integration and web-based applications & services*. 2012. ACM.
39. Al-Twairesh, N., H. Al-Khalifa, A. Al-Salman, and Y. Al-Ohali, *Arasenti-tweet: A corpus for arabic sentiment analysis of saudi tweets.* Procedia Computer Science, 2017. **117**: p. 63-72.
40. Farghaly, A. and K. Shaalan, *Arabic natural language processing: Challenges and solutions.* ACM Transactions on Asian Language Information Processing, 2009. **8**(4): p. 14.
41. Shoukry, A. and A. Rafea. *Sentence-level Arabic sentiment analysis*. in *2012 International Conference on Collaboration Technologies and Systems (CTS)*. 2012. IEEE.
42. Al-Ayyoub, M., A.A. Khamaiseh, Y. Jararweh, and M.N. Al-Kabi, *A comprehensive survey of arabic sentiment analysis.* Information processing management, 2019. **56**(2): p. 320-342.
43. Gamal, D., M. Alfonse, E.-S.M. El-Horbaty, and A.-B.M. Salem, *Twitter Benchmark Dataset for Arabic Sentiment Analysis.* International Journal of Modern Education Computer Science, 2019. **11**(1): p. 33-38.
44. Assiri, A., A. Emam, and H. Al-Dossari, *Towards enhancement of a lexicon-based approach for Saudi dialect sentiment analysis.* Journal of Information Science, 2018. **44**(2): p. 184-202.
45. Al-Twairesh, N., *Sentiment analysis of Twitter: a study on the Saudi community*. 2016, PhD Thesis, King Saud University, Riyadh, Saudi Arabia.
46. Habash, N.Y., *Introduction to Arabic natural language processing.* Synthesis Lectures on Human Language Technologies, 2010. **3**(1): p. 1-187.
47. Gadalla, H., H. Kilany, H. Arram, A. Yacoub, A. El-Habashi, A. Shalaby, K. Karins, E. Rowson, R. MacIntyre, and P. Kingsbury, *CALLHOME Egyptian Arabic Transcripts LDC97T19*, in *Linguistic Data Consortium*. 1997: Philadelphia.
48. Maamouri, M., A. Bies, T. Buckwalter, M.T. Diab, N. Habash, O. Rambow, and D. Tabessi. *Developing and Using a Pilot Dialectal Arabic Treebank*. in *LREC*. 2006. Genoa, Italy.
49. News, B. *Saudi social media users ranked 7th in world*. 2020 [cited 2016 15 Jun]; Available from: https://www.arabnews.com/saudi-arabia/news/835236.
50. Collis, J., R. Hussey, D. Crowther, G. Lancaster, M. Saunders, P. Lewis, A. Thornhill, A. Bryman, E. Bell, and J. Gill, *Business research methods*. 2003, Palgrave Macmillan, New York.
51. Almuqren, L. and A.I. Cristea. *Twitter Analysis to Predict the Satisfaction of Telecom Company Customers*. in *HT (Extended Proceedings)*. 2016.
52. Koeze, E. and N. Popper, *The Virus Changed the Way We Internet*, in *The New York Times*. 2020.
53. Asaari, M. and N. Karia. *Business strategy: customer satisfaction among cellular providers in Malaysia*. in *The European Applied Business Research Conference Proc., Venice, Italy*. 2003.

54. Yoo, D.-K. and S.-W. Suh, *The effect of medical service quality and perceived risk on customer satisfaction, repurchase intention, and churn intention as to hospital sizes.* Korea Serv Manage Soc, 2009. **10**(3): p. 97-130.
55. Hassan, R.S., A. Nawaz, M.N. Lashari, and F. Zafar, *Effect of customer relationship management on customer satisfaction.* Procedia economics and finance, 2015. **23**: p. 563-567.
56. Tam, J.L., *Customer satisfaction, service quality and perceived value: an integrative model.* Journal of marketing management, 2004. **20**(7-8): p. 897-917.
57. Khan, S. and S. Afsheen, *Determinants of customer satisfaction in telecom industry a study of telecom industry peshawar KPK Pakistan.* Journal of Basic Applied Scientific Research, 2012. **2**(12): p. 12833-12840.
58. Tse, D.K. and P.C. Wilton, *Models of consumer satisfaction formation: An extension.* Journal of marketing research, 1988. **25**(2): p. 204-212.
59. Oliver, R.L., *Whence consumer loyalty?* Journal of marketing, 1999. **63**(4_suppl1): p. 33-44.
60. Gustafsson, A., M.D. Johnson, and I. Roos, *The effects of customer satisfaction, relationship commitment dimensions, and triggers on customer retention.* Journal of marketing, 2005. **69**(4): p. 210-218.
61. Lin, C.-S., G.-H. Tzeng, and Y.-C. Chin, *Combined rough set theory and flow network graph to predict customer churn in credit card accounts.* Expert Systems with Applications, 2011. **38**(1): p. 8-15.
62. Hung, S.-Y., D.C. Yen, and H.-Y. Wang, *Applying data mining to telecom churn management.* Expert Systems with Applications, 2006. **31**(3): p. 515-524.
63. Wei, C.-P. and I.-T. Chiu, *Turning telecommunications call details to churn prediction: a data mining approach.* Expert systems with applications, 2002. **23**(2): p. 103-112.
64. Dulhare, U.N. and I. Ghori. *An efficient hybrid clustering to predict the risk of customer churn*. in *2018 2nd International Conference on Inventive Systems and Control (ICISC)*. 2018. IEEE.
65. Leech, G., *Corpus annotation schemes.* Literary and Linguistic Computing, 1993. **8**(4): p. 275-281.
66. Lyu, K. and H. Kim, *Sentiment analysis using word polarity of social media.* Wireless Personal Communications, 2016. **89**(3): p. 941-958.
67. Tundjungsari, V., *Business Intelligence with Social Media and Data Mining to Support Customer Satisfaction in Telecommunication Industry.* International Journal of Computer Science and Electronics Engineering, 2013. **1**(1).
68. Sohangir, S., D. Wang, A. Pomeranets, and T.M. Khoshgoftaar, *Big Data: Deep Learning for financial sentiment analysis.* Journal of Big Data, 2018. **5**(1).
69. Collins, C., S. Hasan, and S. Ukkusuri, *A novel transit rider satisfaction metric: Rider sentiments measured from online social media data.* Journal of Public Transportation, 2013. **16**(2): p. 2.
70. Salampasis, M., G. Paltoglou, and A. Giachanou, *Using Social Media for Continuous Monitoring and Mining of Consumer Behaviour.* IJEB, 2014. **11**(1): p. 85-96.
71. Miranda, M.D. and R.J. Sassi. *Using sentiment analysis to assess customer satisfaction in an online job search company*. in *International Conference on Business Information Systems*. 2014. Springer.
72. Mostafa, M.M., *More than words: Social networks' text mining for consumer brand sentiments.* Expert Systems with Applications, 2013. **40**(10): p. 4241-4251.

73. Tsakalidis, A., S. Papadopoulos, A.I. Cristea, and Y. Kompatsiaris, *Predicting elections for multiple countries using Twitter and polls.* IEEE Intelligent Systems, 2015. **30**(2): p. 10-17.
74. Kampakis, S. and A. Adamides, *Using Twitter to predict football outcomes.* arXiv preprint 2014.
75. Bollen, J., H. Mao, and X. Zeng, *Twitter mood predicts the stock market.* Journal of computational science, 2011. **2**(1): p. 1-8.
76. AbdulGhani, N., S. Hamid, I.A.T. Hashem, and E. Ahmed, *Social media big data analytics: A survey.* Computers in Human Behavior, 2019. **101**: p. 417-428.
77. Kennedy, H., *Perspectives on sentiment analysis.* Journal of Broadcasting Electronic Media, 2012. **56**(4): p. 435-450.
78. Verbeke, W., D. Martens, and B. Baesens, *Social network analysis for customer churn prediction.* Applied Soft Computing, 2014. **14**: p. 431-446.
79. Kumar, S., F. Morstatter, and H. Liu, *Twitter data analytics*. 2014: Springer.
80. Statista. *Leading Countries based on number of Twitter users as of July 2020*. 2020 [cited 9-2-2020; Available from: https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/.
81. Al-Jenaibi, B., *The Twitter revolution in the Gulf countries.* Journal of Creative Communications, 2016. **11**(1): p. 61-83.
82. Homburg, C. and A. Giering, *Personal characteristics as moderators of the relationship between customer satisfaction and loyalty—an empirical analysis.* Psychology Marketing, 2001. **18**(1): p. 43-66.
83. Jahanzeb, S. and S. Jabeen, *Churn management in the telecom industry of Pakistan: A comparative study of Ufone and Telenor.* Journal of Database Marketing Customer Strategy Management, 2007. **14**(2): p. 120-129.
84. Woodside, A.G., L.L. Frey, and R.T. Daly, *Linking sort/ice anlity, customer satisfaction, and behavioral intention.* Journal of health care marketing, 1989. **9**(4): p. 5-17.
85. Hadden, J., A. Tiwari, R. Roy, and D. Ruta, *Churn prediction: Does technology matter.* International Journal of Intelligent Technology, 2006. **1**(2): p. 104-110.
86. Mihelis, G., E. Grigoroudis, Y. Siskos, Y. Politis, and Y. Malandrakis, *Customer satisfaction measurement in the private bank sector.* European Journal of Operational Research, 2001. **130**(2): p. 347-360.
87. Anastasia, S. and I. Budi, *Twitter sentiment analysis of online transportation service providers*, in *2016 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*. 2016, IEEE. p. 359-365.
88. Vidya, N.A., M.I. Fanany, and I. Budi, *Twitter sentiment to analyze net brand reputation of mobile phone providers.* Procedia Computer Science, 2015. **72**: p. 519-526.
89. Kang, D. and Y. Park, *based measurement of customer satisfaction in mobile service: Sentiment analysis and VIKOR approach.* Expert Systems with Applications, 2014. **41**(4): p. 1041-1050.
90. Opricovic, S., *Multicriteria optimization of civil engineering systems.* Faculty of Civil Engineering, 1998. **2**(1): p. 5-21.
91. Olle, G.D.O. and S. Cai, *A hybrid churn prediction model in mobile telecommunication industry.* International Journal of e-Education, e-Business, e-Management e-Learning, 2014. **4**(1): p. 55.
92. Tsai, C.-F. and Y.-H. Lu, *Customer churn prediction by hybrid neural networks.* Expert Systems with Applications, 2009. **36**(10): p. 12547-12553.

93. Dalvi, P.K., S.K. Khandge, A. Deomore, A. Bankar, and V. Kanade. *Analysis of customer churn prediction in telecom industry using decision trees and logistic regression*. in *2016 Symposium on Colossal Data Analysis and Networking (CDAN)*. 2016. IEEE.
94. Balasubramanian, M. and M. Selvarani, *Churn Prediction in Mobile Telecom System Using Data Mining Techniques.* International Journal of Scientific Research Publications, 2014. **4**(4): p. 1-5.
95. Kamalraj, N. and A. Malathi, *A survey on churn prediction techniques in communication sector.* International Journal of Computer Applications, 2013. **64**(5): p. 39-42.
96. Dahiya, K. and K.J.I.J.A.R.C.S.S.E. Talwar, *Customer churn prediction in telecommunication industries using data mining techniques-a review.* 2015. **5**(4): p. 417-433.
97. Mishra, A. and U.S. Reddy. *A comparative study of customer churn prediction in telecom industry using ensemble based classifiers*. in *2017 International Conference on Inventive Computing and Informatics (ICICI)*. 2017. IEEE.
98. Burez, J. and D. Van den Poel, *CRM at a pay-TV company: Using analytical models to reduce customer attrition by targeted marketing for subscription services.* Expert Systems with Applications, 2007. **32**(2): p. 277-288.
99. Xia, G.-e., W.-d.J.S.E.-T. Jin, and Practice, *Model of customer churn prediction on support vector machine.* 2008. **28**(1): p. 71-77.
100. Dolatabadi, S.H. and F. Keynia. *Designing of customer and employee churn prediction model based on data mining method and neural predictor*. in *2017 2nd International Conference on Computer and Communication Systems (ICCCS)*. 2017. IEEE.
101. Oyeniyi, A., A. Adeyemo, A. Oyeniyi, and A. Adeyemo, *Customer churn analysis in banking sector using data mining techniques.* Afr J Comput ICT, 2015. **8**(3): p. 165-174.
102. Fei, T.Y., L.H. Shuan, L.J. Yan, G. Xiaoning, and S.W. King, *Prediction on Customer Churn in the Telecommunications Sector Using Discretization and Naïve Bayes Classifier.* Advance Soft Compu., 2017. **9**(3).
103. Farquad, M.A.H., V. Ravi, and S.B. Raju, *Churn prediction using comprehensible support vector machine: An analytical CRM application.* Applied Soft Computing, 2014. **19**: p. 31-40.
104. Clark, J., I. Koprinska, and J. Poon. *A neural network based approach to automated e-mail classification*. in *Proceedings IEEE/WIC International Conference on Web Intelligence (WI 2003)*. 2003. IEEE.
105. Zhang, R., W. Li, W. Tan, and T. Mo. *Deep and shallow model for insurance churn prediction service*. in *2017 IEEE International Conference on Services Computing (SCC)*. 2017. IEEE.
106. Ahmad, A.K., A. Jafar, and K. Aljoumaa, *Customer churn prediction in telecom using machine learning in big data platform.* Journal of Big Data, 2019. **6**(1): p. 28.
107. Vafeiadis, T., K.I. Diamantaras, G. Sarigiannidis, and K. Chatzisavvas, *A comparison of machine learning techniques for customer churn prediction.* Simulation Modelling Practice and Theory, 2015. **55**: p. 1-9.
108. De Bock, K.W. and D. Van den Poel, *An empirical evaluation of rotation-based ensemble classifiers for customer churn prediction.* Expert Systems with Applications, 2011. **38**(10): p. 12293-12301.
109. Idris, A., A. Khan, and Y.S. Lee, *Intelligent churn prediction in telecom: employing mRMR feature selection and RotBoost based ensemble classification.* Applied intelligence, 2013. **39**(3): p. 659-672.

110. Coussement, K., S. Lessmann, and G. Verstraeten, *A comparative analysis of data preparation algorithms for customer churn prediction: A case study in the telecommunication industry.* Decision Support Systems, 2017. **95**: p. 27-36.
111. Amin, A., F. Al-Obeidat, B. Shah, A. Adnan, J. Loo, and S. Anwar, *Customer churn prediction in telecommunication industry using data certainty.* Journal of Business Research, 2019. **94**: p. 290-301.
112. Dasgupta, K., R. Singh, B. Viswanathan, D. Chakraborty, S. Mukherjea, A.A. Nanavati, and A. Joshi. *Social ties and their relevance to churn in mobile telecom networks*. in *Proceedings of the 11th international conference on Extending database technology: Advances in database technology*. 2008.
113. Napitu, F., M.A. Bijaksana, A. Trisetyarso, and Y. Heryadi. *Twitter opinion mining predicts broadband internet's customer churn rate*. in *2017 IEEE International Conference on Cybernetics and Computational Intelligence (CyberneticsCom)*. 2017. IEEE.
114. Coussement, K., S. Lessmann, and G.J.D.S.S. Verstraeten, *A comparative analysis of data preparation algorithms for customer churn prediction: A case study in the telecommunication industry.* 2017. **95**: p. 27-36.
115. Forhad, N., M.S. Hussain, and R.M. Rahman. *Churn analysis: Predicting churners*. in *Ninth International Conference on Digital Information Management (ICDIM 2014)*. 2014. IEEE.
116. Mohanty, R. and K.J. Rani. *Application of Computational Intelligence to predict churn and non-churn of customers in Indian Telecommunication*. in *2015 International Conference on Computational Intelligence and Communication Networks (CICN)*. 2015. IEEE.
117. Tiwari, A., R. Sam, and S. Shaikh. *Analysis and prediction of churn customers for telecommunication industry*. in *2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC)*. 2017. IEEE.
118. Dasgupta, S., C.H. Papadimitriou, and U.V. Vazirani, *Algorithms*. 2008: McGraw-Hill Higher Education.
119. Amin, A., S. Anwar, A. Adnan, M. Nawaz, K. Alawfi, A. Hussain, and K. Huang, *Customer churn prediction in the telecommunication sector using a rough set approach.* Neurocomputing, 2017. **237**: p. 242-254.
120. Yıldız, M. and S. Albayrak, *Customer churn prediction in telecommunication with rotation forest method.* DBKDA, 2017: p. 35.
121. Kumar, A.S. and D. Chandrakala, *An Optimal Churn Prediction Model using Support Vector Machine with Adaboost.* International Journal of Scientific Research in Computer Science, Engineering Information Technology, 2017. **2**(1): p. 225-230.
122. Azeem, M., M. Usman, and A.C.M. Fong, *A churn prediction model for prepaid customers in telecom using fuzzy classifiers.* Telecommunication Systems, 2017. **66**(4): p. 603-614.
123. Yu, R., X. An, B. Jin, J. Shi, O.A. Move, and Y. Liu, *Particle classification optimization-based BP network for telecommunication customer churn prediction.* Neural Computing Application, 2018. **29**(3): p. 707-720.
124. Khan, M.R., J. Manoj, A. Singh, and J. Blumenstock. *Behavioral modeling for churn prediction: Early indicators and accurate predictors of custom defection and loyalty*. in *2015 IEEE International Congress on Big Data*. 2015. IEEE.

125. Yanfang, Q. and L. Chen. *Research on E-commerce user churn prediction based on logistic regression*. in *2017 IEEE 2nd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*. 2017. IEEE.
126. Au, W.-H., K.C. Chan, and X. Yao, *A novel evolutionary data mining algorithm with applications to churn prediction.* IEEE transactions on evolutionary computation, 2003. **7**(6): p. 532-545.
127. Cao, K. and P.-j. Shao. *Customer churn prediction based on svm-rfe*. in *2008 International Seminar on Business and Information Management*. 2008. IEEE.
128. Burez, J. and D. Van den Poel, *Handling class imbalance in customer churn prediction.* Expert Systems with Applications, 2009. **36**(3): p. 4626-4636.
129. Jadhav, R.J. and U.T. Pawar, *Churn prediction in telecommunication using data mining technology.* International Journal of Advanced Computer Science Applications, 2011. **2**(2).
130. Verbeke, W., D. Martens, C. Mues, and B. Baesens, *Building comprehensible customer churn prediction models with advanced rule induction techniques.* Expert systems with applications, 2011. **38**(3): p. 2354-2364.
131. Abbasimehr, H., M. Setak, and M. Tarokh, *A neuro-fuzzy classifier for customer churn prediction.* Int J Comput Appl, 2011. **19**(8): p. 35-41.
132. Shaaban, E., Y. Helmy, A. Khedr, and M. Nasr, *A proposed churn prediction model.* International Journal of Engineering Research Applications, 2012. **2**(4): p. 693-697.
133. Brandusoiu, I. and G. Toderean, *Churn prediction in the telecommunications sector using support vector machines.* Margin, 2013. **1**: p. x1.
134. Kim, K., C.-H. Jun, and J. Lee, *Improved churn prediction in telecommunication industry by analyzing a large network.* Expert Systems with Applications, 2014. **41**(15): p. 6575-6584.
135. Gürsoy, U.Ş., *Customer churn analysis in telecommunication sector.* İstanbul Üniversitesi İşletme Fakültesi Dergisi, 2010. **39**(1): p. 35-49.
136. Chu, B.-H., M.-S. Tsai, and C.-S. Ho, *Toward a hybrid data mining model for customer retention.* Knowledge-Based Systems, 2007. **20**(8): p. 703-718.
137. Mahajan, V., R. Misra, and R. Mahajan, *Review of data mining techniques for churn prediction in telecom.* Journal of Information and Organizational Sciences, 2015. **39**(2): p. 183-197.
138. Chathuranga, L., R. Rathnayaka, and H. Arumawadu, *New Customer Churn Prediction Model for Mobile Telecommunication Industry*, in *11TH INTERNATIONAL RESEARCH CONFERENCE* 2018.
139. Qureshi, S.A., A.S. Rehman, A.M. Qamar, A. Kamal, and A. Rehman. *Telecommunication subscribers' churn prediction model using machine learning*. in *Eighth International Conference on Digital Information Management (ICDIM 2013)*. 2013. IEEE.
140. Lazarov, V. and M. Capota, *Churn prediction.* Bus. Anal. Course. TUM Comput. Sci, 2007. **33**: p. 34.
141. binti Oseman, K., N.A. Haris, and F. bin Abu Bakar, *Data mining in churn analysis model for telecommunication industry.* Journal of Statistical Modeling and Analytics, 2010. **1**(19-27).
142. Larivière, B. and D. Van den Poel, *Predicting customer retention and profitability by using random forests and regression forests techniques.* Expert Systems with Applications, 2005. **29**(2): p. 472-484.
143. Lu, N., H. Lin, J. Lu, and G.Zhang, *A customer churn prediction model in telecom industry using boosting.* IEEE Transactions on Industrial Informatics, 2012. **10**(2): p. 1659-1665.

144. Glady, N., B. Baesens, and C. Croux, *Modeling churn using customer lifetime value.* European Journal of Operational Research, 2009. **197**(1): p. 402-411.
145. Tsai, C.-F. and M.-Y. Chen, *Variable selection by association rules for customer churn prediction of multimedia on demand.* Expert Systems with Applications, 2010. **37**(3): p. 2006-2015.
146. He, B., Y. Shi, Q. Wan, and X. Zhao, *Prediction of customer attrition of commercial banks based on SVM model.* Procedia Computer Science, 2014. **31**: p. 423-430.
147. Xie, Y., X. Li, E. Ngai, and W. Ying, *Customer churn prediction using improved balanced random forests.* Expert Systems with Applications, 2009. **36**(3): p. 5445-5449.
148. Al-Jazira, *KSA Telecom Sector Report*. 2020: Saudi Arabia. p. 1-43.
149. STC, *PUSHING THE BOUNDARIES TOWARD THE FUTURE*. 2019.
150. Mobily. *Mobily*. 2020; Available from: https://www.mobily.com.sa.
151. Hakeem, M.A. *Zain to launch mobile phone services in Saudi Arabia today*. gulf news, 2008.
152. Pang, B., L. Lee, and S. Vaithyanathan. *Thumbs up?: sentiment classification using machine learning techniques*. in *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*. 2002. Association for Computational Linguistics.
153. Turney, P.D. *Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews*. in *Proceedings of the 40th annual meeting on association for computational linguistics*. 2002. Association for Computational Linguistics.
154. Go, A., R. Bhayani, and L. Huang, *Twitter sentiment classification using distant supervision.* CS224N Project Report, 2009. **1**(12): p. 2009.
155. Pak, A. and P. Paroubek. *Twitter as a corpus for sentiment analysis and opinion mining*. in *LREc*. 2010.
156. Davidov, D., O. Tsur, and A. Rappoport. *Enhanced sentiment learning using twitter hashtags and smileys*. in *Proceedings of the 23rd international conference on computational linguistics: posters*. 2010. Association for Computational Linguistics.
157. Paltoglou, G. and M. Thelwall, *Twitter, MySpace, Digg: Unsupervised sentiment analysis in social media.* ACM Transactions on Intelligent Systems Technology, 2012. **3**(4): p. 66.
158. Thelwall, M., K. Buckley, G. Paltoglou, D. Cai, and A. Kappas, *Sentiment strength detection in short informal text.* Journal of the American Society for Information Science Technology, 2011. **62**(2): p. 397-419.
159. Thelwall, M., K. Buckley, and G. Paltoglou, *Sentiment strength detection for the social web.* Journal of the American Society for Information Science Technology, 2012. **63**(1): p. 163-173.
160. Mohammad, S.M. and X. Zhu. *Sentiment Analysis of Social Media Texts*. in *Tutorial at the 2014 Conference on Empirical Methods on Natural Language Processing*. 2014.
161. Mohammad, S.M., S. Kiritchenko, and X. Zhu, *NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets.* arXiv preprint arXiv:. 2013.
162. Wilson, T., J. Wiebe, and P. Hoffmann. *Recognizing contextual polarity in phrase-level sentiment analysis*. in *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*. 2005.
163. Hu, M. and B. Liu. *Mining and summarizing customer reviews*. in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. 2004. ACM.

164. Mohammad, S.M. and P.D. Turney. *Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon*. in *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*. 2010. Association for Computational Linguistics.
165. Mohammad, S.M. and T.W. Yang. *Tracking sentiment in mail: How genders differ on emotional axes*. in *Proceedings of the 2nd workshop on computational approaches to subjectivity and sentiment analysis*. 2011. Association for Computational Linguistics.
166. Zhu, X., S. Kiritchenko, and S. Mohammad. *Nrc-canada-2014: Recent improvements in the sentiment analysis of tweets*. in *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*. 2014.
167. Rosenthal, S., P. Nakov, S. Kiritchenko, S. Mohammad, A. Ritter, and V. Stoyanov. *Semeval-2015 task 10: Sentiment analysis in twitter*. in *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*. 2015.
168. Severyn, A. and A. Moschitti. *Unitn: Training deep convolutional neural network for twitter sentiment classification*. in *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*. 2015.
169. Kiritchenko, S., S. Mohammad, and M. Salameh. *Semeval-2016 task 7: Determining sentiment intensity of english and arabic phrases*. in *Proceedings of the 10th international workshop on semantic evaluation (SEMEVAL-2016)*. 2016.
170. Rosenthal, S., N. Farra, and P. Nakov. *SemEval-2017 task 4: Sentiment analysis in Twitter*. in *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*. 2017.
171. Refaee, E. and V. Rieser. *iLab-Edinburgh at SemEval-2016 Task 7: A hybrid approach for determining sentiment intensity of Arabic Twitter phrases*. in *Proceedings of the 10th international workshop on semantic evaluation (SEMEVAL-2016)*. 2016.
172. Mohammad, S., F. Bravo-Marquez, M. Salameh, and S. Kiritchenko. *Semeval-2018 task 1: Affect in tweets*. in *Proceedings of the 12th international workshop on semantic evaluation*. 2018.
173. Duppada, V., R. Jain, and S. Hiray, *Seernet at semeval-2018 task 1: Domain adaptation for affect in tweets.* arXiv preprint arXiv:.06137, 2018.
174. Jabreel, M. and A. Moreno, *EiTAKA at SemEval-2018 Task 1: An ensemble of n-channels ConvNet and XGboost regressors for emotion analysis of tweets.* arXiv preprint arXiv:.09233, 2018.
175. Chatterjee, A., K.N. Narahari, M. Joshi, and P. Agrawal. *Semeval-2019 task 3: Emocontext contextual emotion detection in text*. in *Proceedings of the 13th International Workshop on Semantic Evaluation*. 2019.
176. Patwa, P., G. Aguilar, S. Kar, S. Pandey, S. PYKL, B. Gambäck, T. Chakraborty, A. Das, S. PYKL, and G. Aguilar, *Semeval-2020 task 9: Overview of sentiment analysis of code-mixed tweets. .* arXiv e-prints, arXiv-2008.2020 ,.
177. Joshi, N.S. and S.A. Itkat, *A survey on feature level sentiment analysis.* International Journal of Computer Science Information Technologies, 2014. **5**(4): p. 5422-5425.
178. Kaur, A. and V. Gupta, *A survey on sentiment analysis and opinion mining techniques.* Journal of Emerging Technologies in Web Intelligence, 2013. **5**(4): p. 367-371.
179. Dattu, B.S. and D.V. Gore, *A survey on sentiment analysis on twitter data using different techniques.* International Journal of Computer Science Information Technologies, 2015. **6**(6): p. 5358-5362.

180. Salloum, S.A., A.Q. AlHamad, M. Al-Emran, and K. Shaalan, *A survey of Arabic text mining*, in *Intelligent Natural Language Processing: Trends and Applications*. 2018, Springer. p. 417-431.
181. Liu, B., *Sentiment analysis: Mining opinions, sentiments, and emotions*. 2015, Cambridge: Cambridge University Press.
182. Assiri, A., A. Emam, and H. Aldossari, *Arabic sentiment analysis: a survey.* International Journal of Advanced Computer Science Applications, 2015. **6**(12): p. 75-85.
183. Liu, B., *Sentiment analysis and opinion mining.* Synthesis lectures on human language technologies, 2012. **5**(1): p. 1-167.
184. Kouloumpis, E., T. Wilson, and J. Moore. *Twitter sentiment analysis: The good the bad and the omg!* in *Fifth International AAAI conference on weblogs and social media*. 2011.
185. Kiritchenko, S., X. Zhu, and S.M. Mohammad, *Sentiment analysis of short informal texts.* Journal of Artificial Intelligence Research, 2014. **50**: p. 723-762.
186. Siddiqui, S., A.A. Monem, and K. Shaalan, *Evaluation and enrichment of Arabic sentiment analysis*, in *Intelligent Natural Language Processing: Trends and Applications*. 2018, Springer. p. 17-34.
187. Desai, M. and M.A. Mehta. *Techniques for sentiment analysis of Twitter data: A comprehensive survey*. in *2016 International Conference on Computing, Communication and Automation (ICCCA)*. 2016. IEEE.
188. Asghar, M.Z., S. Ahmad, A. Marwat, and F.M. Kundi, *Sentiment analysis on Youtube: a brief survey.* arXiv preprint arXiv:.09142, 2015.
189. Schouten, K. and F. Frasincar, *Survey on aspect-level sentiment analysis.* IEEE Transactions on Knowledge Data Engineering, 2015. **28**(3): p. 813-830.
190. Astya, P. *Sentiment analysis: approaches and open issues*. in *2017 International Conference on Computing, Communication and Automation (ICCCA)*. 2017. IEEE.
191. El-Masri, M., N. Altrabsheh, and H. Mansour, *Successes and challenges of Arabic sentiment analysis research: a literature review.* Social Network Analysis Mining, 2017. **7**(1): p. 54.
192. Banea, C., R. Mihalcea, and J. Wiebe, *Sense-level subjectivity in a multilingual setting.* Computer Speech Language, 2014. **28**(1): p. 7-19.
193. MartíN-Valdivia, M.-T., E. MartíNez-CáMara, J.-M. Perea-Ortega, and L.A. UreñA-LóPez, *Sentiment polarity detection in Spanish reviews combining supervised and unsupervised approaches.* Expert Systems with Applications, 2013. **40**(10): p. 3934-3942.
194. Abo, M.E.M., N. Ahmed, and V. Balakrishnan. *Arabic Sentiment Analysis: An Overview of the ML Algorithms*. in *Data Science Research Symposium 2018*. 2018.
195. Varghese, R. and M. Jayasree, *A survey on sentiment analysis and opinion mining.* International Journal of Research in Engineering Technology, 2013. **2**(11): p. 312-317.
196. Giachanou, A. and F. Crestani, *Like it or not: A survey of twitter sentiment analysis methods.* ACM Computing Surveys, 2016. **49**(2): p. 28.
197. Al-Twairesh, N., H. Al-Khalifa, and A. Al-Salman. *Subjectivity and sentiment analysis of Arabic: trends and challenges*. in *11th International Conference on Computer Systems and Applications (AICCSA)*. 2014. IEEE.
198. Al-Twairesh, N., H. Al-Khalifa, A. Alsalman, and Y. Al-Ohali, *Sentiment analysis of arabic tweets: Feature engineering and a hybrid approach.* arXiv preprint arXiv:.08533, 2018.
199. Abdul-Mageed, M., M. Diab, and S. Kübler, *SAMAR: Subjectivity and sentiment analysis for Arabic social media.* Computer Speech Language, 2014. **28**(1): p. 20-37.
200. El-Halees, A.M., *Arabic text classification using maximum entropy.* Arabic Text Classification Using Maximum Entropy, 2007. **15**(1).

201. Farra, N., E. Challita, R.A. Assi, and H. Hajj. *Sentence-level and document-level sentiment mining for arabic texts*. in *2010 IEEE international conference on data mining workshops*. 2010. IEEE.
202. Al-Ayyoub, M., A.A. Khamaiseh, Y. Jararweh, M. Al-Kabi, *A comprehensive survey of arabic sentiment analysis.* Information processing management, 2019. **56**(2): p. 320-342.
203. Abdul-Mageed, M. and M.T. Diab. *Subjectivity and sentiment annotation of modern standard arabic newswire*. in *Proceedings of the 5th linguistic annotation workshop*. 2011. Association for Computational Linguistics.
204. Elarnaoty, M., S. AbdelRahman, and A. Fahmy, *A machine learning approach for opinion holder extraction in Arabic language.* arXiv preprint 2012.
205. AlMurtadha, Y., *Mining Trending Hash Tags for Arabic Sentiment Analysis.* International Journal of Advanced Computer Science Applications, 2018.
206. Ismail, R., M. Omer, M. Tabir, N. Mahadi, and I. Amin. *Sentiment Analysis for Arabic Dialect Using Supervised Learning*. in *2018 International Conference on Computer, Control, Electrical, and Electronics Engineering (ICCCEEE)*. 2018. IEEE.
207. Balazs, J.A. and J.D. Velásquez, *Opinion mining and information fusion: a survey.* Information Fusion, 2016. **27**: p. 95-110.
208. Nanli, Z., Z. Ping, L. Weiguo, and C. Meng. *Sentiment analysis: A literature review*. in *2012 International Symposium on Management of Technology (ISMOT)*. 2012. IEEE.
209. Piryani, R., D. Madhavi, and V.K. Singh, *Analytical mapping of opinion mining and sentiment analysis research during 2000–2015.* Information Processing Management, 2017. **53**(1): p. 122-150.
210. Yadollahi, A., A.G. Shahraki, and O.R. Zaiane, *Current state of text sentiment analysis from opinion to emotion mining.* ACM Computing Surveys, 2017. **50**(2): p. 1-33.
211. Al-Moslmi, T., M. Albared, A. Al-Shabi, N. Omar, and S. Abdullah, *Arabic senti-lexicon: Constructing publicly available language resources for Arabic sentiment analysis.* Journal of Information Science, 2018. **44**(3): p. 345-362.
212. Al-Twairesh, N., H. Al-Khalifa, and A. Al-Salman. *Arasenti: Large-scale twitter-specific arabic sentiment lexicons*. in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* 2016.
213. Biltawi, M., W. Etaiwi, S. Tedmori, A. Hudaib, and A. Awajan. *Sentiment classification techniques for Arabic language: A survey*. in *2016 7th International Conference on Information and Communication Systems (ICICS)*. 2016. IEEE.
214. Boudad, N., R. Faizi, R.O.H. Thami, and R. Chiheb, *Sentiment analysis in Arabic: A review of the literature.* Ain Shams Engineering Journal, 2018. **9**(4): p. 2479-2490.
215. Dalila, B., A. Mohamed, and H. Bendjanna. *A review of recent aspect extraction techniques for opinion mining systems*. in *2018 2nd International Conference on Natural Language and Speech Processing (ICNLSP)*. 2018. IEEE.
216. Khan, K.S., R. Kunz, J. Kleijnen, and G. Antes, *Five steps to conducting a systematic review.* Journal of the royal society of medicine, 2003. **96**(3): p. 118-121.
217. Alayba, A.M., V. Palade, M. England, and R. Iqbal. *Arabic language sentiment analysis on health services*. in *2017 1st International Workshop on Arabic Script Analysis and Recognition (ASAR)*. 2017. IEEE.
218. Alayba, A.M., V. Palade, M. England, and R. Iqbal. *A combined CNN and LSTM model for arabic sentiment analysis*. in *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*. 2018. Springer.

219. Alayba, A.M., V. Palade, M. England, and R. Iqbal. *Improving sentiment analysis in Arabic using word representation*. in *2018 IEEE 2nd International Workshop on Arabic and Derived Script Analysis and Recognition (ASAR)*. 2018. IEEE.
220. Al-Ayyoub, M., A. Gigieh, A. Al-Qwaqenah, M.N. Al-Kabi, B. Talafhah, and I. Alsmadi, *Aspect-Based Sentiment Analysis of Arabic Laptop*, in *ACIT'2017, The International Arab Conference on Information Technology*. 2017: Yassmine Hammamet, Tunisia.
221. Hamdi, A., K. Shaban, and A. Zainal, *Clasenti: A class-specific sentiment analysis framework.* ACM Transactions on Asian Low-Resource Language Information Processing, 2018. **17**(4): p. 1-28.
222. Al-Rowaily, K., M. Abulaish, N.A.-H. Haldar, and M. Al-Rubaian, *BiSAL–A bilingual sentiment analysis lexicon to analyze Dark Web forums for cyber security.* Digital Investigation, 2015. **14**: p. 53-62.
223. Hathlian, N.F.B. and A.M. Hafezs. *Sentiment-subjective analysis framework for arabic social media posts*. in *2016 4th Saudi International Conference on Information Technology (Big Data Analysis)(KACSTIT)*. 2016. IEEE.
224. Elghazaly, T., A. Mahmoud, and H.A. Hefny. *Political sentiment analysis using twitter data*. in *Proceedings of the International Conference on Internet of things and Cloud Computing*. 2016.
225. Mahmoud, A. and T. Elghazaly, *Using twitter to monitor political sentiment for Arabic slang*, in *Intelligent Natural Language Processing: Trends and Applications*. 2018, Springer. p. 53-66.
226. Almas, Y. and K. Ahmad. *A note on extracting 'sentiments' in financial news in English, Arabic & Urdu*. in *The Second Workshop on Computational Approaches to Arabic Script-based Languages*. 2007.
227. Moraes, R., J.F. Valiati, and W.P.G. Neto, *Document-level sentiment classification: An empirical comparison between SVM and ANN.* Expert Systems with Applications, 2013. **40**(2): p. 621-633.
228. Wang, G., J. Sun, J. Ma, K. Xu, and J. Gu, *Sentiment classification: The contribution of ensemble learning.* Decision support systems, 2014. **57**: p. 77-93.
229. Abo, M.E.M., N.A.K. Shah, V. Balakrishnan, and A. Abdelaziz. *Sentiment analysis algorithms: evaluation performance of the Arabic and English language*. in *2018 International Conference on Computer, Control, Electrical, and Electronics Engineering (ICCCEEE)*. 2018. IEEE.
230. Aly, M. and A. Atiya. *Labr: A large scale arabic book reviews dataset*. in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 2013.
231. Al-Harbi, O., *Using objective words in the reviews to improve the colloquial arabic sentiment analysis.* arXiv preprint 2017. **arXiv:.08521**.
232. Gamal, D., M. Alfonse, E.-S.M. El-Horbaty, and A.-B.M. Salem, *Implementation of Machine Learning Algorithms in Arabic Sentiment Analysis Using N-Gram Features.* Procedia Computer Science, 2019. **154**: p. 332-340.
233. AL-Jumaili, A.S.A. and H.K. Tayyeh, *A Hybrid Method of Linguistic and Statistical Features for Arabic Sentiment Analysis.* Baghdad Science Journal, 2020. **17**(1 Supplement): p. 385-390.
234. Farha, I.A. and W. Magdy. *Mazajak: An online Arabic sentiment analyser*. in *Proceedings of the Fourth Arabic Natural Language Processing Workshop*. 2019.
235. Mubarak, H., A. Rashed, K. Darwish, Y. Samih, and A. Abdelali, *Arabic offensive language on twitter: Analysis and experiments.* arXiv preprint arXiv:.02192, 2020.

236. Bahassine, S., A. Madani, M. Al-Sarem, and M. Kissi, *Feature selection using an improved Chi-square for Arabic text classification.* Journal of King Saud University-Computer Information Sciences, 2020. **32**(2): p. 225-231.
237. Setiyaningrum, Y.D., A.F. Herdajanti, and C. Supriyanto. *Classification of Twitter Contents using Chi-Square and K-Nearest Neighbour Algorithm*. in *2019 International Seminar on Application for Technology of Information and Communication (iSemantic)*. 2019. IEEE.
238. Abdulla, N., *Twitter Data set for Arabic Sentiment Analysis Data Set.* Machine Learning Repository, 2019.
239. Al-Shalabi, R. and R. Obeidat. *Improving KNN Arabic text classification with n-grams based document indexing*. in *Proceedings of the Sixth International Conference on Informatics and Systems, Cairo, Egypt*. 2008. Citeseer.
240. Abdelwadood, M.d.A.M., *Chi square feature extraction based svms arabic language text categorization system.* Journal of Computer Science, 2007. **3**(6): p. 430-435.
241. Hassonah, M.A., R. Al-Sayyed, A. Rodan, A.-Z. Ala'M, I. Aljarah, and H. Faris, *An efficient hybrid filter and evolutionary wrapper approach for sentiment analysis of various topics on Twitter.* Knowledge-Based Systems, 2020. **192**: p. 105353.
242. Rushdi-Saleh, M., M.T. Martín-Valdivia, L.A. Ureña-López, and J.M. Perea-Ortega. *Bilingual experiments with an arabic-english corpus for opinion mining*. in *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*. 2011.
243. Witten, I.H., E. Frank, M.A. Hall, and C.J. Pal, *Data Mining: Practical machine learning tools and techniques*. 2016: Morgan Kaufmann.
244. Azmi, A. and S. Alzanin, *Aara'–a system for mining the polarity of Saudi public opinion through e-newspaper comments.* Journal of Information Science, 2014. **40**(3): p. 398-410.
245. Boiy, E., P. Hens, K. Deschacht, and M.-F. Moens. *Automatic Sentiment Analysis in On-line Text*. in *ELPUB*. 2007.
246. Abdul-Mageed, M. and M. Korayem, *Automatic identification of subjectivity in morphologically rich languages: the case of Arabic.* Computational approaches to subjectivity sentiment analysis, 2010. **2**: p. 2-6.
247. Pang, B. and L. Lee, *Opinion mining and sentiment analysis.* Foundations Trends® in Information Retrieval, 2008. **2**(1–2): p. 1-135.
248. Kotsiantis, S.B., I. Zaharakis, and P. Pintelas, *Supervised machine learning: A review of classification techniques.* Emerging artificial intelligence applications in computer engineering, 2007. **160**(1): p. 3-24.
249. Aldahawi, H., *Mining and analysing social network in the oil business: Twitter sentiment analysis and prediction approaches*. 2015, Cardiff University.
250. Shoukry, A. and A. Rafea. *A hybrid approach for sentiment classification of Egyptian dialect tweets*. in *2015 First International Conference on Arabic Computational Linguistics (ACLing)*. 2015. IEEE.
251. Alsaleem, S., *Automated Arabic Text Categorization Using SVM and NB.* Int. Arab. J. e Technol., 2011. **2**(2): p. 124-128.
252. Read, J. and J. Carroll. *Weakly supervised techniques for domain-independent sentiment classification*. in *Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion*. 2009. ACM.
253. Al-Thubaity, A., Q. Alqahtani, and A. Aljandal, *Sentiment lexicon for sentiment analysis of Saudi dialect tweets.* Procedia computer science, 2018. **142**: p. 301-307.

254. Abdulla, N.A., N.A. Ahmed, M.A. Shehab, M. Al-Ayyoub, M.N. Al-Kabi, and S. Al-rifai, *Towards improving the lexicon-based approach for arabic sentiment analysis.* International Journal of Information Technology Web Engineering, 2014. **9**(3): p. 55-71.
255. Itani, M.M., R.N. Zantout, L. Hamandi, and I. Elkabani. *Classifying sentiment in arabic social networks: Naive search versus naive bayes*. in *2012 2nd International Conference on Advances in Computational Tools for Engineering Applications (ACTEA)*. 2012. IEEE.
256. Abdulla, N., S. Mohammed, M. Al-Ayyoub, and M. Al-Kabi. *Automatic lexicon construction for arabic sentiment analysis*. in *International Conference on Future Internet of Things and Cloud*. 2014. IEEE.
257. Abdulla, N.A., N.A. Ahmed, M.A. Shehab, and M. Al-Ayyoub. *Arabic sentiment analysis: Lexicon-based and corpus-based*. in *IEEE Jordan conference on applied electrical engineering and computing technologies (AEECT)*. 2013. IEEE.
258. Abdul-Mageed, M. and M. Diab. *Sana: A large scale multi-genre, multi-dialect lexicon for arabic subjectivity and sentiment analysis*. in *LREC*. 2014.
259. Elhawary, M. and M. Elfeky. *Mining Arabic business reviews*. in *2010 ieee international conference on data mining workshops*. 2010. IEEE.
260. Ibrahim, H.S., S.M. Abdou, and M. Gheith, *Idioms-proverbs lexicon for modern standard Arabic and colloquial sentiment analysis.* arXiv preprint arXiv:.01906, 2015.
261. Areed, S., O. Alqaryouti, B. Siyam, and K. Shaalan, *Aspect-Based Sentiment Analysis for Arabic Government Reviews*, in *Recent Advances in NLP: The Case of Arabic Language*. 2020, Springer. p. 143-162.
262. Ihnaini, B. and M. Mahmuddin, *Valence Shifter Rules for Arabic Sentiment Analysis.* International Journal of Multidisciplinary Sciences and Advanced Technology, 2020. **1**(2): p. 167-184.
263. El-Beltagy, S.R., T. Khalil, A. Halaby, and M. Hammad. *Combining lexical features and a supervised learning approach for Arabic sentiment analysis*. in *International Conference on Intelligent Text Processing and Computational Linguistics*. 2016. Springer.
264. Mohammad, S., M. Salameh, and S. Kiritchenko. *Sentiment lexicons for Arabic social media*. in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. 2016.
265. El-Beltagy, S.R. and A. Ali. *Open issues in the sentiment analysis of Arabic social media: A case study*. in *2013 9th International Conference on Innovations in Information Technology (IIT)*. 2013. IEEE.
266. Turney, P.D. and M.L. Littman, *Measuring praise and criticism: Inference of semantic orientation from association.* ACM Transactions on Information Systems, 2003. **21**(4): p. 315-346.
267. Abdul-Mageed, M. and M. Diab. *Toward building a large-scale Arabic sentiment lexicon*. in *Proceedings of the 6th international global WordNet conference*. 2012.
268. Salameh, M., S. Mohammad, and S. Kiritchenko. *Sentiment after translation: A case-study on arabic social media posts*. in *Proceedings of the 2015 conference of the North American chapter of the association for computational linguistics: Human language technologies*. 2015.
269. Mobarz, H., M. Rashown, and I. Farag, *Using automated lexical resources in Arabic sentence subjectivity.* International Journal of Artificial Intelligence Applications, 2014. **5**(6): p. 1.
270. Pasha, A., M. Al-Badrashiny, M.T. Diab, A. El Kholy, R. Eskander, N. Habash, M. Pooleery, O. Rambow, and R. Roth. *Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of arabic*. in *In Proceedings of the Ninth*

International Conference on Language Resources and Evaluation (LREC'14). 2014. Reykjavik, Iceland: LREC.
271. Church, K. and P. Hanks, *Word association norms, mutual information, and lexicography.* Computational linguistics, 1990. **16**(1): p. 22-29.
272. Refaee, E. and V. Rieser. *An arabic twitter corpus for subjectivity and sentiment analysis*. in *LREC*. 2014.
273. Nabil, M., M. Aly, and A. Atiya. *Astd: Arabic sentiment tweets dataset*. in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 2015.
274. Black, W., S. Elkateb, H. Rodriguez, M. Alkhalifa, P. Vossen, A. Pease, and C. Fellbaum. *Introducing the Arabic wordnet project*. in *Proceedings of the third international WordNet conference*. 2006. Citeseer.
275. Alowaidi, S., M. Saleh, and O. Abulnaja, *Semantic sentiment analysis of Arabic texts.* International Journal of Advanced Computer Science Applications, 2017. **8**(2): p. 256-262.
276. Bayoudhi, A., H. Ghorbel, H. Koubaa, and L.H. Belguith, *Sentiment Classification at Discourse Segment Level: Experiments on multi-domain Arabic corpus.* J. Lang. Technol. Comput. Linguistics, 2015. **30**(1): p. 1-24.
277. Mahyoub, F.H., M.A. Siddiqui, and M. Dahab, *Building an Arabic sentiment lexicon using semi-supervised learning.* Journal of King Saud University-Computer and Information Sciences, 2014. **26**(4): p. 417-424.
278. Al-Twairesh, N., R. Al-Matham, N. Madi, N. Almugren, A.-H. Al-Aljmi, S. Alshalan, R. Alshalan, N. Alrumayyan, S. Al-Manea, and S. Bawazeer, *Suar: Towards building a corpus for the Saudi dialect.* Procedia computer science, 2018. **142**: p. 72-82.
279. Al-Thubaity, A., M. Alharbi, S. Alqahtani, and A. Aljandal. *A Saudi dialect Twitter Corpus for sentiment and emotion analysis*. in *2018 21st Saudi Computer Society National Computer Conference (NCC)*. 2018. IEEE.
280. Badaro, G., R. Baly, H. Hajj, N. Habash, and W. El-Hajj. *A large scale Arabic sentiment lexicon for Arabic opinion mining*. in *Proceedings of the EMNLP 2014 workshop on arabic natural language processing (ANLP)*. 2014.
281. Aldayel, H.K. and A.M. Azmi, *Arabic tweets sentiment analysis–a hybrid scheme.* Journal of Information Science, 2016. **42**(6): p. 782-797.
282. Esuli, A. and F. Sebastiani. *Sentiwordnet: A publicly available lexical resource for opinion mining*. in *LREC*. 2006. Citeseer.
283. Hasan, A.A. and A.C. Fong. *Sentiment analysis based fuzzy decision platform for the Saudi stock market*. in *2018 IEEE International Conference on Electro/Information Technology (EIT)*. 2018. IEEE.
284. Stone, P.J., D.C. Dunphy, and M.S. Smith, *The general inquirer: A computer approach to content analysis*. 1966.
285. Islam, M.R. and M.F. Zibran, *SentiStrength-SE: Exploiting domain specificity for improved sentiment analysis in software engineering text.* Journal of Systems Software, 2018. **145**: p. 125-146.
286. El-Halees, A.M., *Arabic opinion mining using combined classification approach.* Arabic opinion mining using combined classification approach, 2011.
287. Maamouri, M., A. Bies, T. Buckwalter, and W. Mekki. *The penn arabic treebank: Building a large-scale annotated arabic corpus*. in *NEMLAR conference on Arabic language resources and tools*. 2004. Cairo.
288. Baccianella, S., A. Esuli, and F. Sebastiani. *Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining*. in *Lrec*. 2010.

289. Herring, S.C., *A faceted classification scheme for computer-mediated discourse.* Language@ internet, 2007. **4**(1).
290. Al-Moslmi, T., M. Albared, A. Al-Shabi, N. Omar, and S. Abdullah, *Arabic senti-lexicon: Constructing publicly available language resources for Arabic sentiment analysis.* Journal of Information Science, 2018. **44**(3): p. 345-362.
291. ElSahar, H. and S.R. El-Beltagy. *A fully automated approach for arabic slang lexicon extraction from microblogs*. in *International conference on intelligent text processing and computational linguistics*. 2014. Springer.
292. El-Beltagy, S.R. *NileULex: a phrase and word level sentiment lexicon for Egyptian and modern standard Arabic*. in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. 2016.
293. Buckwalter, T., *Buckwalter arabic morphological analyzer version 2.0. linguistic data consortium, university of pennsylvania, 2002. ldc cat alog no*. 2004, Ldc2004l02. Technical report.
294. Eskander, R. and O. Rambow. *Slsa: A sentiment lexicon for standard arabic*. in *Proceedings of the 2015 conference on empirical methods in natural language processing*. 2015.
295. Graff, D., M. Maamouri, B. Bouziri, S. Krouna, S. Kulick, and T. Buckwalter, *Standard Arabic morphological analyzer (SAMA) version 3.1.* Linguistic Data Consortium LDCE73, 2009: p. 53-56.
296. Miller, G.A., R. Beckwith, C. Fellbaum, D. Gross, and K.J. Miller, *Introduction to WordNet: An on-line lexical database.* International journal of lexicography, 1990. **3**(4): p. 235-244.
297. ElSahar, H. and S.R. El-Beltagy. *Building large arabic multi-domain resources for sentiment analysis*. in *International Conference on Intelligent Text Processing and Computational Linguistics*. 2015. Springer.
298. Rabab'Ah, A.M., M. Al-Ayyoub, Y. Jararweh, and M.N. Al-Kabi. *Evaluating sentistrength for arabic sentiment analysis*. in *7th International Conference on Computer Science and Information Technology (CSIT)*. 2016. IEEE.
299. Park, S., W. Lee, and I.-C. Moon, *Efficient extraction of domain specific sentiment lexicon with active learning.* Pattern Recognition Letters, 2015. **56**: p. 38-44.
300. Saif, H., Y. He, M. Fernandez, and H. Alani, *Contextual semantics for sentiment analysis of Twitter.* Information Processing Management, 2016. **52**(1): p. 5-19.
301. Taboada, M., J. Brooke, M. Tofiloski, K. Voll, and M. Stede, *Lexicon-based methods for sentiment analysis.* Computational linguistics, 2011. **37**(2): p. 267-307.
302. Ohana, B. and B. Tierney. *Sentiment classification of reviews using SentiWordNet*. in *9th. IT & T conference*. 2009.
303. Kang, H., S.J. Yoo, and D. Han, *Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews.* Expert Systems with Applications, 2012. **39**(5): p. 6000-6010.
304. Bouchlaghem, R., A. Elkhelifi, and R. Faiz. *A machine learning approach for classifying sentiments in Arabic tweets*. in *Proceedings of the 6th international conference on web intelligence, mining and semantics*. 2016.
305. Alhumoud, S., T. Albuhairi, and W. Alohaideb. *Hybrid sentiment analyser for Arabic tweets using R*. in *2015 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K)*. 2015. IEEE.

306. Rennie, J.D., L. Shih, J. Teevan, and D.R. Karger. *Tackling the poor assumptions of naive bayes text classifiers*. in *Proceedings of the 20th international conference on machine learning (ICML-03)*. 2003.
307. Alhumoud, S., T. Albuhairi, and M. Altuwaijri. *Arabic sentiment analysis using WEKA a hybrid learning approach*. in *2015 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K)*. 2015. IEEE.
308. Biltawi, M., G. Al-Naymat, and S. Tedmori. *Arabic sentiment classification: A hybrid approach*. in *international conference on new trends in computing sciences (ICTCS)*. 2017. IEEE.
309. Rushdi-Saleh, M., M.T. Martín-Valdivia, L.A. Ureña-López, and J.M. Perea-Ortega, *OCA: Opinion corpus for Arabic.* Journal of the American Society for Information Science Technology, 2011. **62**(10): p. 2045-2054.
310. Najafabadi, M.M., F. Villanustre, T.M. Khoshgoftaar, N. Seliya, R. Wald, and E. Muharemagic, *Deep learning applications and challenges in big data analytics.* Journal of Big Data, 2015. **2**(1): p. 1.
311. Glorot, X., A. Bordes, and Y. Bengio. *Deep sparse rectifier neural networks*. in *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. 2011.
312. Rumelhart, D.E., G.E. Hinton, and R.J. Williams, *Learning representations by back-propagating errors.* nature, 1986. **323**(6088): p. 533-536.
313. Zhang, L., S. Wang, and B. Liu, *Deep learning for sentiment analysis: A survey.* Wiley Interdisciplinary Reviews: Data Mining Knowledge Discovery, 2018. **8**(4): p. e1253.
314. Sohangir, S., D. Wang, A. Pomeranets, and T.M. Khoshgoftaar, *Big Data: Deep Learning for financial sentiment analysis.* Journal of Big Data, 2018. **5**(1): p. 3.
315. Mohamed, A.-r., G.E. Dahl, and G. Hinton, *Acoustic modeling using deep belief networks.* IEEE transactions on audio, speech,language processing, 2011. **20**(1): p. 14-22.
316. Socher, R., E.H. Huang, J. Pennin, C.D. Manning, and A.Y. Ng. *Dynamic pooling and unfolding recursive autoencoders for paraphrase detection*. in *Advances in neural information processing systems*. 2011.
317. Collobert, R., J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, *Natural language processing (almost) from scratch.* Journal of machine learning research, 2011. **12**(Aug): p. 2493-2537.
318. Goldberg, Y., *A primer on neural network models for natural language processing.* Journal of Artificial Intelligence Research, 2016. **57**: p. 345-420.
319. LeCun, Y., Y. Bengio, and G. Hinton, *Deep learning.* nature, 2015. **521**(7553): p. 436-444.
320. Wei, X., H. Lin, Y. Yu, and L. Yang, *Low-resource cross-domain product review sentiment classification based on a CNN with an auxiliary large-scale Corpus.* Algorithms, 2017. **10**(3): p. 81.
321. Goodfellow, I., Y. Bengio, and A. Courville, *Deep learning*. 2016: MIT press.
322. Elman, J.L., *Finding structure in time.* Cognitive science, 1990. **14**(2): p. 179-211.
323. Bengio, Y., P. Simard, and P. Frasconi, *Learning long-term dependencies with gradient descent is difficult.* IEEE transactions on neural networks, 1994. **5**(2): p. 157-166.
324. Hochreiter, S. and J. Schmidhuber, *Long short-term memory.* Neural computation, 1997. **9**(8): p. 1735-1780.
325. Schuster, M. and K.K. Paliwal, *Bidirectional recurrent neural networks.* IEEE transactions on Signal Processing, 1997. **45**(11): p. 2673-2681.

326. Cho, K., B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, *Learning phrase representations using RNN encoder-decoder for statistical machine translation.* arXiv preprint, 2014.
327. Chung, J., C. Gulcehre, K. Cho, and Y. Bengio, *Empirical evaluation of gated recurrent neural networks on sequence modeling.* arXiv preprint, 2014.
328. Kim, Y., *Convolutional neural networks for sentence classification.* arXiv preprint arXiv:.01906, 2014.
329. Wang, X., Y. Liu, C.-J. Sun, B. Wang, and X. Wang. *Predicting polarities of tweets by composing word embeddings with long short-term memory*. in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 2015.
330. Yin, W. and H. Schütze, *Multichannel variable-size convolution for sentence classification.* arXiv preprint arXiv:.04513, 2016.
331. Shin, B., T. Lee, and J.D. Choi, *Lexicon integrated CNN models with attention for sentiment analysis.* arXiv preprint arXiv:.06272, 2016.
332. Oussous, A., F.-Z. Benjelloun, A.A. Lahcen, and S. Belfkih, *ASA: A framework for Arabic sentiment analysis.* Journal of Information Science, 2019: p. 16.
333. Alwehaibi, A. and K. Roy. *Comparison of pre-trained word vectors for Arabic text classification using deep learning approach*. in *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*. 2018. IEEE.
334. Alahmary, R.M., H.Z. Al-Dossari, and A.Z. Emam. *Sentiment analysis of Saudi dialect using deep learning techniques*. in *2019 International Conference on Electronics, Information, and Communication (ICEIC)*. 2019. IEEE.
335. Fouad, M.M., A. Mahany, N. Aljohani, R.A. Abbasi, and S.-U. Hassan, *ArWordVec: efficient word embedding models for Arabic tweets.* Soft Computing, 2019.
336. Almuqren, L.A., M. Moh'd Qasem, and A.I. Cristea, *Using deep learning networks to predict telecom company customer satisfaction based on Arabic tweets*, in *ISD*. 2019: Tolerance, France.
337. Al-Smadi, M., O. Qawasmeh, M. Al-Ayyoub, Y. Jararweh, and B. Gupta, *Deep Recurrent neural network vs. support vector machine for aspect-based sentiment analysis of Arabic hotels' reviews.* Journal of computational science, 2018. **27**: p. 386-393.
338. Al-Smadi, M., B. Talafha, M. Al-Ayyoub, Y. Jararweh, *Using long short-term memory deep neural networks for aspect-based sentiment analysis of Arabic reviews.* International Journal of Machine Learning and Cybernetics, 2018: p. 1-13.
339. Heikal, M., M. Torki, and N. El-Makky, *Sentiment analysis of Arabic Tweets using deep learning.* Procedia Computer Science, 2018. **142**: p. 114-122.
340. Abdullah, M., M. Hadzikadicy, and S. Shaikhz. *SEDAT: sentiment and emotion detection in Arabic text using CNN-LSTM deep learning*. in *17th IEEE International Conference on Machine Learning and Applications (ICMLA)*. 2018. IEEE.
341. Ombabi, A.H., W. Ouarda, and A.M. Alimi, *Deep learning CNN–LSTM framework for Arabic sentiment analysis using textual information shared in social networks.* Social Network Analysis Mining, 2020. **10**(1): p. 1-13.
342. Almani, N. and L.H. Tang. *Deep Attention-Based Review Level Sentiment Analysis for Arabic Reviews*. in *2020 6th Conference on Data Science and Machine Learning Applications (CDMA)*. 2020. IEEE.
343. Elnagar, A., R. Al-Debsi, and O. Einea, *Arabic text classification using deep learning models.* Information Processing Management, 2020. **57**(1): p. 102121.

344. Beseiso, M. and H. Elmousalami, *Subword Attentive Model for Arabic Sentiment Analysis: A Deep Learning Approach.* ACM Transactions on Asian Low-Resource Language Information Processing, 2020. **19**(2): p. 1-17.
345. Elzayady, H., K.M. Badran, and G.I. Salama, *Arabic Opinion Mining Using Combined CNN-LSTM Models.* International Journal of Intelligent System Applications, 2020. **12**(4).
346. Albayati, A.Q., A.S. Al-Araji, and S.H. Ameen, *Arabic Sentiment Analysis (ASA) Using Deep Learning Approach.* Journal of Engineering, 2020. **26**(6): p. 85-93.
347. Fouadi, H., H. El Moubtahij, H. Lamtougui, K. SATORI, and A. Yahyaouy. *Applications of Deep Learning in Arabic Sentiment Analysis: Research Perspective*. in *2020 1st International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET)*. 2020. IEEE.
348. Baker, Q.B., F. Shatnawi, S. Rawashdeh, M. Al-Smadi, and Y. Jararweh, *Detecting Epidemic Diseases Using Sentiment Analysis of Arabic Tweets.* J. UCS, 2020. **26**(1): p. 50-70.
349. Yang, X., C. Macdonald, and I. Ounis, *Using word embeddings in twitter election classification.* Information Retrieval Journal, 2018. **21**(2-3): p. 183-207.
350. Al Sallab, A., H. Hajj, G. Badaro, R. Baly, W. El-Hajj, and K. Shaban. *Deep learning models for sentiment analysis in Arabic*. in *Proceedings of the second workshop on Arabic natural language processing*. 2015.
351. Song, S., H. Huang, and T. Ruan, *Abstractive text summarization using LSTM-CNN based deep learning.* Journal of building construction planning Research, 2019. **78**(1): p. 857-875.
352. Dey, R. and F.M. Salemt. *Gate-variants of gated recurrent unit (GRU) neural networks*. in *2017 IEEE 60th international midwest symposium on circuits and systems (MWSCAS)*. 2017. IEEE.
353. Athiwaratkun, B. and J.W. Stokes. *Malware classification with LSTM and GRU language models and a character-level CNN*. in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2017. IEEE.
354. Gers, F.A., D. Eck, and J. Schmidhuber, *Applying LSTM to time series predictable through time-window approaches*, in *Neural Nets WIRN Vietri-01*. 2002, Springer. p. 193-200.
355. Ruder, S., P. Ghaffari, and J.G. Breslin, *Insight-1 at semeval-2016 task 5: Deep learning for multilingual aspect-based sentiment analysis.* arXiv preprint arXiv:.02748, 2016.
356. Atassi, A., I. El Azami, and A. Sadiq. *The new deep learning architecture based on GRU and word2vec*. in *2018 International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS)*. 2018. IEEE.
357. Al-Smadi, M., B. Talafha, M. Al-Ayyoub, and Y. Jararweh, *Using long short-term memory deep neural networks for aspect-based sentiment analysis of Arabic reviews.* International Journal of Machine Learning and Cybernetics, 2019. **10**(8): p. 2163-2175.
358. ElJundi, O., W. Antoun, N. El Droubi, H. Hajj, W. El-Hajj, and K. Shaban. *hULMonA: The Universal Language Model in Arabic*. in *Proceedings of the Fourth Arabic Natural Language Processing Workshop*. 2019.
359. Howard, J. and S. Ruder, *Universal language model fine-tuning for text classification.* arXiv preprint arXiv:.06146, 2018.
360. Liu, R., Y. Shi, C. Ji, and M. Jia, *A survey of sentiment analysis based on transfer learning.* IEEE Access, 2019. **7**: p. 85401-85412.

361. Mikolov, T., K. Chen, G. Corrado, and J. Dean, *Efficient estimation of word representations in vector space.* arXiv preprint arXiv:. 2013.
362. Pennington, J., R. Socher, and C.D. Manning. *Glove: Global vectors for word representation*. in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014.
363. Joulin, A., E. Grave, P. Bojanowski, M. Douze, H. Jégou, and T. Mikolov, *Fasttext. zip: Compressing text classification models.* arXiv preprint arXiv:.03651, 2016.
364. Bojanowski, P., E. Grave, A. Joulin, and T. Mikolov, *Enriching word vectors with subword information.* Transactions of the Association for Computational Linguistics, 2017. **5**: p. 135-146.
365. Grave, E., P. Bojanowski, P. Gupta, A. Joulin, and T. Mikolov, *Learning word vectors for 157 languages.* arXiv preprint arXiv:.06893, 2018.
366. Altowayan, A.A. and L. Tao. *Word embeddings for Arabic sentiment analysis*. in *2016 IEEE International Conference on Big Data (Big Data)*. 2016. IEEE.
367. Soliman, A.B., K. Eissa, and S.R. El-Beltagy, *Aravec: A set of arabic word embedding models for use in arabic nlp.* Procedia Computer Science, 2017. **117**: p. 256-265.
368. Mikolov, T., I. Sutskever, K. Chen, G.S. Corrado, and J. Dean. *Distributed representations of words and phrases and their compositionality*. in *Advances in neural information processing systems*. 2013.
369. Radford, A., J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, *Language models are unsupervised multitask learners.* OpenAI Blog, 2019. **1**(8).
370. Wang, A., A. Singh, J. Michael, F. Hill, O. Levy, and S.R. Bowman, *Glue: A multi-task benchmark and analysis platform for natural language understanding.* arXiv preprint arXiv:.07461, 2018.
371. Merity, S., N.S. Keskar, and R. Socher, *Regularizing and optimizing LSTM language models.* arXiv preprint arXiv:.02182, 2017.
372. Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova, *Bert: Pre-training of deep bidirectional transformers for language understanding.* arXiv preprint arXiv:.04805, 2018.
373. Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, and I. Polosukhin. *Attention is all you need*. in *Advances in neural information processing systems*. 2017.
374. Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, *Roberta: A robustly optimized bert pretraining approach.* arXiv preprint arXiv:.11692, 2019.
375. Al-Twairesh, N. and H. Al-Negheimish, *Surface and Deep Features Ensemble for Sentiment Analysis of Arabic Tweets.* IEEE Access, 2019. **7**: p. 84122-84131.
376. Baly, F. and H. Hajj. *AraBERT: Transformer-based Model for Arabic Language Understanding*. in *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*. 2020.
377. Zahran, M.A., A. Magooda, A.Y. Mahgoub, H. Raafat, M. Rashwan, and A. Atyia. *Word representations in vector space and their applications for arabic*. in *International Conference on Intelligent Text Processing and Computational Linguistics*. 2015. Springer.
378. Abdelali, A., K. Darwish, N. Durrani, and H. Mubarak. *Farasa: A fast and furious segmenter for arabic*. in *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Demonstrations*. 2016.

379. Rushdi-Saleh, M., M.T. Martín-Valdivia, L.A. Ureña-López, J.M. Perea-Ortega. *OCA: Opinion corpus for Arabic.* Journal of the American Society for Information Science Technology, 2011. **62**(10): p. 2045-2054.
380. Atia, S. and K. Shaalan. *Increasing the accuracy of opinion mining in Arabic*. in *2015 First International Conference on Arabic Computational Linguistics (ACLing)*. 2015. IEEE.
381. Bouamor, H., N. Habash, M. Salameh, W. Zaghouani, O. Rambow, D. Abdulrahim, O. Obeid, S. Khalifa, F. Eryani, and A. Erdmann. *The madar arabic dialect corpus and lexicon*. in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. 2018.
382. Takezawa, T., G. Kikui, M. Mizushima, and E. Sumita. *Multilingual spoken language corpus development for communication research*. in *International Journal of Computational Linguistics & Chinese Language Processing, Volume 12, Number 3, September 2007: Special Issue on Invited Papers from ISCSLP 2006*. 2007.
383. Habash, N., O. Rambow, and R. Roth. *MADA+ TOKAN: A toolkit for Arabic tokenization, diacritization, morphological disambiguation, POS tagging, stemming and lemmatization*. in *Proceedings of the 2nd international conference on Arabic language resources and tools (MEDAR), Cairo, Egypt*. 2009.
384. Khalifa, S., N. Zalmout, and N. Habash. *Yamama: Yet another multi-dialect arabic morphological analyzer*. in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*. 2016.
385. Inoue, G., H. Shindo, and Y. Matsumoto. *Joint prediction of morphosyntactic categories for fine-grained arabic part-of-speech tagging exploiting tag dictionary information*. in *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*. 2017.
386. Zribi, I., M. Ellouze, L.H. Belguith, and P. Blache, *Spoken Tunisian Arabic corpus "STAC": transcription and annotation.* Research in computing science, 2015. **90**: p. 123-135.
387. Smaıli, K., M. Abbas, K. Meftouh, and S. Harrat. *Building resources for Algerian Arabic dialects*. in *15th Annual Conference of the International Communication Association Interspeech*. 2014.
388. Khalifa, S., N. Habash, D. Abdulrahim, and S. Hassan, *A large scale corpus of Gulf Arabic.* arXiv preprint arXiv:.02960, 2016.
389. Khalifa, S., N. Habash, F. Eryani, O. Obeid, D. Abdulrahim, and M. Al Kaabi. *A morphologically annotated corpus of Emirati Arabic*. in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. 2018.
390. Ibrahim, H.S., S.M. Abdou, and M. Gheith. *MIKA: A tagged corpus for modern standard Arabic and colloquial sentiment analysis*. in *2015 IEEE 2nd International Conference on Recent Trends in Information Systems (ReTIS)*. 2015. IEEE.
391. Refaee, E. and V. Rieser. *An arabic twitter corpus for subjectivity and sentiment analysis*. in *In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. 2014. Reykjavik, Iceland: LREC.
392. De Roeck, A., *ELRA's Al-Hayat Dataset: Text Resources in Arabic Language Engineering.* ELRA Newsletter, 2002. **7**(1).
393. Eckart, T., F. Alshargi, U. Quasthoff, and D. Goldhahn. *Large Arabic Web Corpora of high quality: the dimensions time and origin*. in *Workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools Workshop Program*. 2014.
394. Abdul-Mageed, M. and M. Diab. *AWATIF: A Multi-Genre Corpus for Modern Standard Arabic Subjectivity and Sentiment Analysis*. in *LREC*. 2012.

395. Alqarafi, A., A. Adeel, A. Hawalah, K. Swingler, and A. Hussain. *A Semi-supervised Corpus Annotation for Saudi Sentiment Analysis Using Twitter*. in *International Conference on Brain Inspired Cognitive Systems*. 2018. Springer.
396. Assiri, A., A. Emam, and H. Al-Dossari, *Saudi twitter corpus for sentiment analysis.* World Academy of Science, Engineering Technology, International Journal of Computer, Electrical, Automation, Control nformation Engineering, 2016. **10**(2): p. 272-275.
397. Al-Harbi, W.A. and A. Emam, *Effect of Saudi dialect preprocessing on Arabic sentiment analysis.* International Journal of Advanced Computer Technology, 2015. **4**(6): p. 91-99.
398. Al-Rubaiee, H., R. Qiu, and D. Li. *Identifying Mubasher software products through sentiment analysis of Arabic tweets*. in *2016 International Conference on Industrial Informatics and Computer Systems (CIICS)*. 2016. IEEE.
399. Buckwalter, T., *Buckwalter arabic morphological analyzer version 2.0.*, in *linguistic data consortium*. 2004, university of pennsylvania.
400. Diab, M., K. Hacioglu, and D. Jurafsky, *Automated methods for processing arabic text: from tokenization to base phrase chunking.* Arabic Computational Morphology: Knowledge-based Empirical Methods., 2007.
401. Maamouri, M., D. Graff, B. Bouziri, S. Krouna, A. Bies, and S. Kulick, *Standard Arabic morphological analyzer (SAMA) version 3.1.* Linguistic Data Consortium, Catalog No.: LDCL01, 2010.
402. Khoja, S. and R. Garside, *Stemming arabic text*. 1999: Lancaster, UK, Computing Department, Lancaster University..
403. Benajiba, Y., M. Diab, and P. Rosso. *Arabic named entity recognition using optimized feature sets*. in *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. 2008.
404. Benajiba, Y., M. Diab, and P. Rosso, *Arabic named entity recognition: A feature-driven study.* IEEE Transactions on Audio, Speech, Language Processing, 2009. **17**(5): p. 926-934.
405. Murray, D.G., *Tableau your data!: fast and easy visual analysis with tableau software*. 2013: John Wiley & Sons.
406. Al-Kabi, M., I. Alsmadi, R.T. Khasawneh, and H. Wahsheh, *Evaluating social context in arabic opinion mining.* Int. Arab J. Inf. Technol., 2018. **15**(6): p. 974-982.
407. Eskander, R. *SentiArabic: A Sentiment Analyzer for Standard Arabic*. in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*. 2018.
408. Al-Subaihin, A.S. and H.S. Al-Khalifa, *A system for sentiment analysis of colloquial Arabic using human computation.* The Scientific World Journal, 2014. **2014**.
409. Hamouda, A.E.-D.A. and F.E.-z. El-taher, *Sentiment analyzer for arabic comments system.* Int. J. Adv. Comput. Sci. Appl, 2013. **4**(3).
410. Elawady, R.M., S. Barakat, and M.E. Nora, *Sentiment analyzer for arabic comments.* International Journal of Information Science Intelligent System, 2014. **3**(4): p. 73-86.
411. El-Beltagy, S.R., M.E. Kalamawy, and A.B. Soliman, *Niletmrg at semeval-2017 task 4: Arabic sentiment analysis.* arXiv preprint arXiv:.08458, 2017.
412. Elmadany, A., H. Mubarak, and W. Magdy, *Arsas: An arabic speech-act and sentiment corpus of tweets.* OSACT, 2018. **3**: p. 20.
413. Al-Kabi, M., A. Gigieh, I. Alsmadi, H. Wahsheh, and M. Haidar. *An opinion analysis tool for colloquial and standard Arabic*. in *The Fourth International Conference on Information and Communication Systems (ICICS 2013)*. 2013.

414. Jarrar, M., N. Habash, F. Alrimawi, D. Akra, and N. Zalmout, *Curras: an annotated corpus for the Palestinian Arabic dialect.* Language Resources Evaluation, 2017. **51**(3): p. 745-775.
415. El-Khair, I.A., *1.5 billion words arabic corpus.* arXiv preprint arXiv:.04033, 2016.
416. Saudi InformationTechnology Commission, *Communications and information technology sector performance indicators*. 2017: Saudi Arabia.
417. Roberts, C. and D. Torgerson, *Randomisation methods in controlled trials.* Bmj, 1998. **317**(7168): p. 1301-1310.
418. Marshall, M.N., *Sampling for qualitative research.* Family practice, 1996. **13**(6): p. 522-526.
419. Kim, H., S.M. Jang, S.-H. Kim, and A. Wan, *Evaluating sampling methods for content analysis of Twitter data.* Social Media+ Society, 2018. **4**(2): p. 1-10.
420. Gerlitz, C. and B. Rieder, *Mining one percent of Twitter: Collections, baselines, sampling.* M/C Journal, 2013. **16**(2).
421. Smith, M.A., B. Shneiderman, N. Milic-Frayling, E. Mendes Rodrigues, V. Barash, C. Dunne, T. Capone, A. Perer, and E. Gleave. *Analyzing (social media) networks with NodeXL*. in *Proceedings of the fourth international conference on Communities and technologies*. 2009.
422. Barbosa, L. and J. Feng. *Robust sentiment detection on twitter from biased and noisy data*. in *In Proceedings of the International Conference on Computational Linguistics (COLING-2010)*. 2010. Beijing.
423. Sun, S., C. Luo, and J. Chen, *A review of natural language processing techniques for opinion mining systems.* Information fusion, 2017. **36**: p. 10-25.
424. Soler-Company, J. and L. Wanner, *On the role of syntactic dependencies and discourse relations for author and gender identification.* Pattern Recognition Letters, 2018. **105**: p. 87-95.
425. Refaee, E. *Sentiment analysis for micro-blogging platforms in Arabic*. in *International Conference on Social Computing and Social Media*. 2017. Springer.
426. Hinze, A., R. Heese, M. Luczak-Rösch, and A. Paschke. *Semantic enrichment by non-experts: usability of manual annotation tools*. in *International Semantic Web Conference*. 2012. Berlin, Heidelberg: Springer.
427. Wissler, L., M. Almashraee, D.M. Díaz, and A. Paschke. *The Gold Standard in Corpus Annotation*. in *In IEEE Germany Student Conference.* . 2014. Germany: University of Passau
428. Abbasi, A., H. Chen, and A. Salem, *Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums.* ACM Transactions on Information Systems, 2008. **26**(3): p. 1-34.
429. Fang, X. and J. Zhan, *Sentiment analysis using product review data.* Journal of Big Data, 2015. **2**(1): p. 5.
430. Mourad, A. and K. Darwish. *Subjectivity and sentiment analysis of modern standard Arabic and Arabic microblogs*. in *Proceedings of the 4th workshop on computational approaches to subjectivity, sentiment and social media analysis*. 2013.
431. Pustejovsky, J. and A. Stubbs, *Natural Language Annotation for Machine Learning: A guide to corpus-building for applications*. 2012, Sebastopol, CA: O'Reilly Media, Inc.
432. Cambria, E., D. Das, S. Bandyopadhyay, and A. Feraco, *A practical guide to sentiment analysis*. 2017, Cham, Switzerland: Springer International Publishing.

433. Rajadesingan, A., R. Zafarani, and H. Liu. *Sarcasm detection on twitter: A behavioral modeling approach*. in *Proceedings of the eighth ACM international conference on web search and data mining*. 2015.
434. Davies, M. and J.L. Fleiss, *Measuring agreement for multinomial data.* Biometrics, 1982. **38**(4): p. 1047-1051.
435. Fleiss, J.L., *Measuring nominal scale agreement among many raters.* Psychological bulletin, 1971. **76**(5): p. 378-382.
436. James, G., D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning*. Vol. 112. 2013, New York: Springer.
437. Forman, G., *An extensive empirical study of feature selection metrics for text classification.* Journal of machine learning research, 2003. **3**(Mar): p. 1289-1305.
438. Commission, S.I., *ICT Indicators.* 2019.
439. Varki, S. and M. Colgate, *The role of price perceptions in an integrated model of behavioral intentions.* Journal of Service Research, 2001. **3**(3): p. 232-240.
440. Woo, K.-S. and H.K. Fock, *Customer satisfaction in the Hong Kong mobile phone industry.* Service Industries Journal, 1999. **19**(3): p. 162-174.
441. Athanassopoulos, A.D. and A. Iliakopoulos, *Modeling customer satisfaction in telecommunications: Assessing the effects of multiple transaction points on the perceived overall performance of the provider.* Production Operations Management, 2003. **12**(2): p. 224-245.
442. Nguyen, E.H.X., *Customer Churn Prediction for the Icelandic Mobile Telephony Market*. PhD diss, 2011. University of Iceland, Reykjavik
443. Eria, K. and B.P. Marikannan, *Systematic Review of Customer Churn Prediction in the Telecom Sector.* Journal of Applied Technology Innovation, 2018. **2**(1).
444. Óskarsdóttir, M., C. Bravo, W. Verbeke, C. Sarraute, B. Baesens, and J. Vanthienen. *A comparative study of social network classifiers for predicting churn in the telecommunication industry*. in *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. 2016. IEEE Press.
445. Snyder, H., *Literature review as a research methodology: An overview and guidelines.* Journal of Business Research, 2019. **104**: p. 333-339.
446. Webster, J. and R.T. Watson, *Analyzing the past to prepare for the future: Writing a literature review.* MIS quarterly, 2002: p. xiii-xxiii.
447. Cohen, L., L. Manion, and K. Morrison, *Research methods in education*. 2011: Routledge.
448. Creswell, J.W., *Qualitative inquiry and research design: Choosing among five approaches*. 2012: Sage.
449. Kahn, S.R., D.L. Lamping, T. Ducruet, L. Arsenault, M.J. Miron, A. Roussin, S. Desmarais, F. Joyal, J. Kassis, and S. Solymoss, *VEINES-QOL/Sym questionnaire was a reliable and valid disease-specific quality of life measure for deep venous thrombosis.* Journal of clinical epidemiology, 2006. **59**(10): p. 1056. e1-1056. e4.
450. Kline, S.J., *Similitude and approximation theory*. 2012: Springer Science & Business Media.
451. Cohen, L., L. Manion, and K. Morrison, *Observation.* Research methods in education, 2007. **6**: p. 396-412.
452. Saunders, M.N. and P. Lewis, *Doing research in business & management: An essential guide to planning your project*. 2012: Pearson.

453. Wright, K.B., *Researching Internet-based populations: Advantages and disadvantages of online survey research, online questionnaire authoring software packages, and web survey services.* Journal of health care marketing, 2005. **10**(3): p. JCMC1034.
454. Yun, G. and C. Trumbo, *Comparative response to a survey executed by post, e-mail, and web form. 2000.* Journal of Computer-Mediated Communication, 2006. **6**(1).
455. Cronbach, L., *Coefficient alpha and the internal structure of tests.* psychometrika, 1951. **16**(3): p. 297-334.
456. Hariri, A.A., *Adoption of learning innovations within UK universities: validating an extended and modified UTAUT model*. 2014, University of Warwick.
457. Boone, H.N. and D.A. Boone, *Analyzing likert data.* Journal of extension, 2012. **50**(2): p. 1-5.
458. Chakravarti, I.M., R.G. Laha, and J. Roy, *Handbook of methods of applied statistics.* Wiley Series in Probability Mathematical Statistics eng, 1967.
459. Guest, G., K.M. MacQueen, and E.E. Namey, *Applied thematic analysis*. 2011: sage publications.
460. Popping, R., *Analyzing open-ended questions by means of text analysis procedures.* Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique, 2015. **128**(1): p. 23-39.
461. Somiah, M., G. Osei-Poku, and I. Aidoo, *Relative importance analysis of factors influencing unauthorized siting of residential buildings in the Sekondi-Takoradi Metropolis of Ghana.* Journal of building construction planning Research, 2015. **3**(03): p. 117.
462. Mimmack, G., G. Manas, and D. Meyer, *Introductory Statistics for Business*. 2001: Pearson South Africa.
463. Akmal, M., *Factor Causing Customer Churn: A Qualitative Explanation of Customer Churns In Pakistan Telecom Industry*. 2017, MS Thesis.
464. Singh, R. and A.A. Tiwari, *Churn Analysis of Indian Telecom Customers.* IMI Konnect, 2019. **8**(1).
465. Keramati, A., R. Jafari-Marandi, M. Aliannejadi, I. Ahmadian, M. Mozaffari, and U. Abbasi, *Improved churn prediction in telecommunication industry using data mining techniques.* Applied Soft Computing ,2014. **24**: p. 994-1012.
466. Saraswat, S. and A. Tiwari, *A New Approach for Customer Churn Prediction in Telecom Industry.* International Journal of Computer Applications, 2018. **181**(11): p. 40-46.
467. Tang, J., S. Alelyani, and H. Liu, *Feature selection for classification: A review.* Data classification: Algorithms applications, 2014: p. 37.
468. Hastie, T., R. Tibshirani, J. Friedman, and J. Franklin, *The elements of statistical learning: data mining, inference and prediction.* The Mathematical Intelligencer, 2005. **27**(2): p. 83-85.
469. Martineau, J.C. and T. Finin. *Delta tfidf: An improved feature space for sentiment analysis*. in *Third international AAAI conference on weblogs and social media*. 2009.
470. Al-Ayyoub, M., A. Nuseir, G. Kanaan, and R. Al-Shalabi, *Hierarchical classifiers for multi-way sentiment analysis of arabic reviews.* International Journal of Advanced Computer Science Applications, 2016. **7**(2): p. 531-539.
471. Chicco, D. and G. Jurman, *The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation.* BMC genomics, 2020. **21**(1): p. 6.
472. Miner, G., R. Nisbet, and I. Elder, *Handbook of statistical analysis and data mining applications*. 2009: Academic Press.

473. Powers, D.M., *Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation.* arXiv preprint arXiv:2010, 2020. **16061**
474. Sokolova, M. and G. Lapalme, *A systematic analysis of performance measures for classification tasks.* Information processing management, 2009. **45**(4): p. 427-437.
475. Abbasi, A. and H. Chen, *Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums.* ACM Transactions on Information Systems, 2008.
476. Athanassopoulos, A.D. and A. Iliakopoulos, *Modeling customer satisfaction in telecommunications: assessing the effects of multiple transaction points on the perceived overall performance of the provider.* Production and Operations Management, 2003. **12**: p. 224-245.
477. Vapnik, V.N., *An overview of statistical learning theory.* IEEE transactions on neural networks, 1999. **10**(5): p. 988-999.
478. Chang, C.-C. and C.-J. Lin, *LIBSVM: A library for support vector machines.* ACM transactions on intelligent systems technology, 2011. **2**(3): p. 1-27.
479. Hsu, C.-W., C.-C. Chang, and C.-J. Lin, *A practical guide to support vector classification*. 2003, Taipei.
480. Aggarwal, C.C. and C. Zhai, *Mining text data*. 2012: Springer Science & Business Media.
481. Ahmed, S., M. Pasquier, and G. Qadah. *Key issues in conducting sentiment analysis on Arabic social media text*. in *2013 9th International Conference on Innovations in Information Technology (IIT)*. 2013. IEEE.
482. Yang, Y. and J.O. Pedersen. *A comparative study on feature selection in text categorization*. in *Icml*. 1997. Nashville, TN, USA.
483. Chollet, F., *Keras: The python deep learning library.* ascl, 2018: p. ascl: 1806.022.
484. Abadi, M., A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G.S. Corrado, A. Davis, J. Dean, and M. Devin, *Tensorflow: Large-scale machine learning on heterogeneous distributed systems.* arXiv preprint 2016. **arXiv:.04467**.
485. Xiong, J., K. Zhang, and H. Zhang. *A vibrating mechanism to prevent neural networks from overfitting*. in *2019 15th International Wireless Communications & Mobile Computing Conference (IWCMC)*. 2019. IEEE.
486. Kingma, D.P. and J. Ba, *Adam: A method for stochastic optimization.* arXiv preprint arXiv:.01906, 2014.
487. Liu, G. and G. Jiabao, *Bidirectional LSTM with attention mechanism and convolutional layer for text classification.* Neurocomputing 2019: p. 325-338.
488. Lachenbruch, P.A. and C.J. Lynch., *Assessing screening tests: extensions of McNemar's test.* Statistics in medicine, 1998. **17**(19): p. 2207-2217..
489. Alrajhi, L., K. Alharbi, and A.I. Cristea, *A multidimensional deep learner model of urgent instructor intervention need in MOOC Forum Posts.*, in *International Conference on Intelligent Tutoring* 2020, Springer, Cham.. p. 226-236.
490. Gulordava, K., L. Aina, and G. Boleda. *How to represent a word and predict it, too: Improving tied architectures for language modelling*. in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing; 2018 Oct 31-Nov 4; Brussels, Belgium. Stroudsburg: Association for Computational Linguistics; 2018. p. 2936–41.* 2018. ACL (Association for Computational Linguistics).
491. Radford, A., R. Jozefowicz, and I. Sutskever, *Learning to generate reviews and discovering sentiment.* arXiv preprint arXiv:.01444, 2017.
492. Google. *Google Colab*. [cited 2019 10/1]; Available from: https://colab.research.google.com.

493. Sharma, S., S. Srivastava, A. Kumar, and A. Dangi. *Multi-Class Sentiment Analysis Comparison Using Support Vector Machine (SVM) and BAGGING Technique-An Ensemble Method*. in *2018 International Conference on Smart Computing and Electronic Enterprise (ICSCEE)*. 2018. IEEE.
494. Al Shboul, B., M. Al-Ayyoub, and Y. Jararweh. *Multi-way sentiment classification of arabic reviews*. in *2015 6th International Conference on Information and Communication Systems (ICICS)*. 2015. IEEE.
495. Stein, R.A., P.A. Jaques, and J.F. Valiati, *An analysis of hierarchical text classification using word embeddings.* Information Sciences, 2019. **471**: p. 216-232.
496. Bickerstaffe, A. and I. Zukerman. *A hierarchical classifier applied to multi-way sentiment detection*. in *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*. 2010.
497. Rosenthal, S., N. Farra, and P. Nakov, *SemEval-2017 Task 4: Sentiment Analysis in Twitter.* Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017), 2017: p. 502-518.
498. Mohammad, S., F. Bravo-Marquez, M. Salameh, and s. Kiritchenko. *Semeval-2018 task 1: Affect in tweets*. in *In Proceedings of the 12th international workshop on semantic evaluation*. 2018.
499. Baly, R., G. Badaro, A. Hamdi, R. Moukalled, R. Aoun, G. El-Khoury, A. Al Sallab, H. Hajj, N. Habash, and K. Shaban. *Omam at semeval-2017 task 4: Evaluation of english state-of-the-art sentiment analysis models for arabic and a new topic-based model*. in *Proceedings of the 11th international workshop on semantic evaluation (SEMEVAL-2017)*. 2017.
500. Elnagar, A., Y.S. Khalifa, and A. Einea, *Hotel Arabic-reviews dataset construction for sentiment analysis applications*, in *Intelligent Natural Language Processing: Trends and Applications*. 2018, Springer. p. 35-52.
501. Azmi, A.M. and S. Alzanin, *Aara'–a system for mining the polarity of Saudi public opinion through e-newspaper comments.* Journal of Information Science 2014. **40**(3): p. 398-410.
502. Nuseir, A., M. Al-Ayyoub, M. Al-Kabi, G. Kanaan, and R. Al-Shalabi, *Improved Hierarchical Classifiers for Multi-Way Sentiment Analysis.* International Arab Journal of Information Technology, 2017. **14**.
503. Silla, C.N. and A.A. Freitas, *A survey of hierarchical classification across different application domains.* Data Mining Knowledge Discovery, 2011. **22**(1-2): p. 31-72.
504. ADDI, H.A., R. EZZAHIR, and A. MAHMOUDI. *Three-level binary tree structure for sentiment classification in Arabic text*. in *In proceeding of the 3rd International Conference on Networking, Information Systems & Security*. 2020.
505. Hao, P.-Y., J.-H. Chiang, and Y.-K. Tu, *Hierarchically SVM classification based on support vector clustering method and its application to document categorization.* Expert Systems with applications, 2007. **33**(3): p. 627-635.
506. ARMANO, G., F. MASCIA, and E. VARGIU, *Using taxonomic domain knowledge in text categorization tasks.* INTERNATIONAL JOURNAL OF INTELLIGENT CONTROL AND SYSTEMS 2007.
507. Angiani, G., S. Cagnoni, N. Chuzhikova, P. Fornacciari, M. Mordonini, and M. Tomaiuolo. *Flat and hierarchical classifiers for detecting emotion in tweets*. in *Conference of the Italian Association for Artificial Intelligence*. 2016. Springer.

508. Kumar, S., J. Ghosh, and M.M. Crawford, *Hierarchical fusion of multiple classifiers for hyperspectral data analysis.* Pattern Analysis Applications, 2002. **5**(2): p. 210-220.
509. Allen, D.M., *Mean square error of prediction as a criterion for selecting variables.* Technometrics, 1971. **13**(3): p. 469-475.
510. Sun, A., E.P. Lim, and W.K. Ng, *Performance measurement framework for hierarchical text classification.* Journal of the American Society for Information Science Technology, 2003. **54**(11): p. 1014-1028.
511. Haenlein, M., *Social interactions in customer churn decisions: The impact of relationship directionality.* International Journal of Research in Marketing, 2013. **30**(3): p. 236-248.
512. Hudaib, A., R. Dannoun, O. Harfoushi, R. Obiedat, and H. Faris, *Hybrid data mining models for predicting customer churn.* International Journal of Communications, Network System Sciences, 2015. **8**(05): p. 91.
513. Sonia, S.E., S.B. Rajakumar, and C. Nalini, *Churn Prediction using MAPREDUCE.* International Journal of Scientific Engineering Technology 2014. **3**(5): p. 597-600.
514. Bin, L., S. Peiji, and L. Juan. *Customer churn prediction based on the decision tree in personal handyphone system service*. in *2007 International Conference on Service Systems and Service Management*. 2007. IEEE.
515. Bakır, B., I. Batmaz, F. Güntürkün, İ. İpekçi, G. Köksal, and N. Özdemirel, *Defect cause modeling with decision tree and regression analysis.* World Acad Sci Eng Technoly, 2006. **24**: p. 1-4.
516. Hassouna, M. and M. Arzoky. *Agent based modelling and simulation: Toward a new model of customer retention in the mobile market*. in *Proceedings of the 2011 summer computer simulation conference*. 2011.p.33-35.
517. Stovel, M. and N. Bontis, *Voluntary turnover: knowledge management–friend or foe?* Journal of intellectual Capital, 2002.
518. Saudi InformationTechnology Commission. *indicator_mobtelservices_byyear*. 2019 [cited 11 Aug 2019; Available from: https://ictind.citc.gov.sa/extensions/ICTPublicReports/Ar/indicator_mobtelservices_byyear_ar.html.
519. Wu, X., X. Zhu, G.-Q. Wu, and W. Ding, *Data mining with big data.* IEEE transactions on knowledge and data engineering, 2013. **26**(1): p. 97-107.
520. Brachman, R.J. and T. Anand, *The process of knowledge discovery in databases*, in *Advances in knowledge discovery and data mining*. 1996. p. 37-57.
521. Frawley, W.J., G. Piatetsky-Shapiro, and C.J. Matheus, *Knowledge discovery in databases: An overview.* AI magazine, 1992. **13**(3): p. 57-57.
522. Shafique, U. and H. Qaiser, *A comparative study of data mining process models (KDD, CRISP-DM and SEMMA).* International Journal of Innovation Scientific Research, 2014. **12**(1): p. 217-222.
523. Rudin, C. and K.L. Wagstaff, *Machine learning for science and society*. 2014, Springer.
524. Mariscal, G., O. Marban, and C. Fernandez, *A survey of data mining and knowledge discovery process models and methodologies.* The Knowledge Engineering Review, 2010. **25**(2): p. 137-166.
525. Li, H., D. Yang, L. Yang, and X. Lin. *Supervised massive data analysis for telecommunication customer churn prediction*. in *2016 IEEE International Conferences on Big Data and Cloud Computing (BDCloud), Social Computing and Networking (SocialCom), Sustainable Computing and Communications (SustainCom)(BDCloud-SocialCom-SustainCom)*. 2016. IEEE.

526. Karahoca, A., D. Karahoca, and N. Aydin. *GSM churn management using an adaptive neuro-fuzzy inference system*. in *The 2007 International Conference on Intelligent Pervasive Computing (IPC 2007)*. 2007. IEEE.

Appendices

Appendix A: Annotation Guideline

Annotation Guidelines ارشادات التوسيم

هذه الدراسة تهدف الى قياس رضاء المستخدمين اتجاه شركات الاتصال عن طريق تحليل أراء العملاء في تويتر وتصنيفها حسب الجدول الموضح بالأسفل.

The aim of this study is to measure customer satisfaction toward telecommunication companies through analysing the customer tweets on Twitter according to the table shown below.

Label الوسم	Definition التعريف	Example مثال
إيجابي **Positive**	إذا كانت التغريدة تحتوي على دائل واضحة على رأي إيجابي نحو الشركة ، حتى لو لم يكن الرأي قوي جدا . If there is a clear indicator that the opinion is positive even if it is not strong.	انترنت Mobily ممتاز . Internet of Mobily is good **تصنيف التغريده:** أيجابي. **Tweet label :**Positive
سلبي **Negative**	إذا كانت التغريدة تحتوي على دائل واضحة على رأي سلبي ، حتى لو لم يكن الرأي قوي جدا . If there is a clear indicator that the opinion is a negative, even if it is not strong.	خدمة الانترنت في Mobily ضعيفة . Internet of Mobily is weak. **تصنيف التغريده:** سلبي. **Tweet label:** Negative.
ساخر **Sarcasm**	تعتبر التغريده ساخره أذا كان في ظاهرها أيجابي ولكنها ضمنيا تحمل رأي سلبي او بالعكس. The tweet considers a sarcasm if a tweet says something positive while it really means something negative or vice versa.	موبايلي المستشار الخبير الرائع فشل في حل المشكلة . Mobily is the wonderfully expert advisor failed to resolve the problem. **التصنيف:** ساخر. **Label :**Sarcasm.
لايمكن تحديده **Indeterminate**	عند عدم وضوح الرأي نرجو عدم التخمين ولكن اختيار(لا يمكن تحديده). When the sentiment doesn't clear choose (indeterminate).	**موبايلي شركة جيدة ولكن من الاسوأ.** Mobily is a good company, but it is one of the worst companies here.

اذا كانت التغريدة تحتوي على خدمة من خدمات شركات الاتصالات المحددة مسبقا فالرجاء تحديد الخدمة والمشاعر اتجاهها.

According to the table below, choose the predefined telecommunication services from the list mentioned in the tweet and sentiment toward it.

Label الوسم	Definition التعريف	Example مثال
إيجابي **Positive**	إذا كانت التغريدة تحتوي على دائل واضحة على رأي إيجابي نحو خدمة من خدمات الشركة ، حتى لو لم يكن الرأي قوي جدا . If there is a clear indicator that the sentiment is positive toward the service even if it is not strong.	انترنت Mobily ممتاز Internet of Mobily is good **تصنيف التغريده:** أيجابي **Service label :**Positive **الكلمة المؤثرة :** ممتازة **Sentiment-bearing word :**good **تصنيف الكلمة المؤثرة:** ايجابي **Label of sentiment-bearing word**: Positive **الخدمة:** انترنت **Service:** Internet
ايجابي بشدة **Strongly Positive**	تعتبر المشاعر اتجاه الخدمة ايجابيه بشدة اذا تطابقت مع احد الحالات التالية: If the service's sentiment is one of the following cases that indicates the feeling is strongly positive . 1)إذا كانت التغريدة تحتوي على دائل واضحة على رأي إيجابي مع كلمة توكيدية مثل (جدا, كثيرا, مره, دايما, أكيد..) أو دعاء. The tweet has a positive sentiment with intensifier such as very, or extremely 2) تعددت الكلمات الإيجابية في التغريده الواحده كلمتين وأكثر. The tweet has more than two positive words toward the service. 3) رأي إيجابي مع تكرارالحرف اكثر من ثلاث مرات. The tweet has a positive sentiment-bearing word with repeated letters. 4) أذا كانت التغريده تحتوي على كلمة ايجابيه بشكل مقارنة مثل أفضل, أحسن, أسرع. The tweet has a positive sentiment-bearing word in superlative form.	1) الانترنت سريع جدا Internet of Mobily is very fast. **تصنيف الخدمة:** أيجابي بشدة **service label:** Strongly positive **الكلمة المؤثرة:** سريع **Sentiment-bearing word:** fast **تصنيف الكلمة المؤثرة:** أيجابي **Label of sentiment-bearing word:** positive 2)الانترنت في موبايلي ممتاز وسريع وقوي Internet of Mobily is excellent and fast. 3) الانترنت في موبايلي سرييييع Internet of Mobily is fassst. 4) الانترنت في موبايلي الاسرع Internet of Mobily is fastest.

Label الوسم	Definition التعريف	Example مثال
سلبي **Negative**	إذا كانت التغريدة تحتوي على دائل واضحة على رأي سلبي ، حتى لو لم يكن الرأي قوي جدا . If there is a clear indicator that the sentiment is a negative toward the service, it is not strong. even if	خدمة الانترنت في Mobily ضعيفة Internet of Mobily is weak. **تصنيف الخدمة- Service label** سلبي Negative **الكلمة المؤثرة :**ضعيف **Sentiment-bearing word:** Weak **تصنيف الكلمة الموثرة:** سلبي **Label of sentiment-bearing word:** Negative **الخدمة** : الانترنت **Service** Internet
سلبي بشدة **Strongly Negative**	تعتبر التغريدة سلبيه بشدة اذا تطابقت مع احد الحالات التالية If the service's sentiment is one of the following cases that indicates the feeling is strongly negative. 1) إذا كانت التغريدة تحتوي على دائل واضحة على رأي سلبي مع كلمة توكيدية مثل (جدا, كثيرا, مره,دايما, كمان ...) أو دعاء. The tweet has a negative sentiment with intensifier such as very, or extremely 2)اذا تعددت الكلمات السلبيه في التغريده الواحده كلمتين وأكثر. The tweet has more than two negative words toward the service.	1) انترنت موبايلي بطئ جدا Internet of Mobily is very slowly. **تصنيف الخدمة:** سلبي بشدة **Service label** :Strongly Negative **الكلمة المؤثرة :** بطيء **Sentiment-bearing word :** slowly **تصنيف الكلمة الموثرة:** سلبي **Label of sentiment-bearing word:** Negative **الخدمة:** الانترنت **Service :**Internet 2) ا نترنت موبايلي بطئ وضعيف Internet of Mobily is slow and weak.
	3)أذا كانت التغريده تحتوي على كلمة مقارنة مثل أفشل, أسوأ, أبطء The tweet has a negative sentiment-bearing word in superlative form.	3)انترنت موبايلي أبطأ نت Internet of Mobily is the slowest one.

4) انترنت موبايلي بطيييء Internet of Mobily is slooooow.	4) رأي سلبي مع تكرارالحرف اكثر من ثلاث مرات The tweet has a negative sentiment-bearing word with repeated letters.	
احب شركة موبايلي. I love Mobily company.	إذا لم يكون هناك رأي معبر عنه في التغريدة اتجاه احد الخدمات. There is no sentiment in the tweet toward any predefined service.	**محايد** **Neural**
هل تشتغل خدمة G4 ؟ Is the 4G service working today?		
الكلمة المؤثرة: لا يوجد **Sentiment-bearing word**: Empty		
الخدمة : لا يوجد **Service:** Empty		

Appendix B: Questionnaire

Exploratory Survey about customer satisfaction

This Survey will help with better understanding the parameters that lead to customer satisfaction toward Telecom companies in Saudi Arabia. This survey is done as part of research undertaken in Princess Nourah bint Abdulrahman University, Saudi Arabia. No personal information is collected in the survey and the results of the survey will only be used for research purposes.
It will take no more than 10 minutes of your time. Your participation is much appreciated.

Please answer all these questions according to their types:
○ a single choice, □ a multiple-choice, and ___ a text-field

Thank you.

A. Personal Background

Age group:
○ 18–24
○ 25–34
○ 35–44
○ 45–54
○ 55–64
○ 65+

Gender:
○ Female
○ Male

Twitter account (optional): _____

What Telecommunication Company are you using for your mobile phone or Internet access?

○ STC (Saudi Telecommunication Company)
○ Mobily
○ Zain
○ Others: _____

Part B: Behaviours and Characteristics of Participants who Changed their Telecommunication Company.

1. **Did you change your telecommunication company before?**

○ Yes
○ No

If you have changed your telecommunication company in the last ten years, answer 2, 3,4,5 and 6.

2. What was the previous telecommunication company that you used as a cell phone network?

- ○ STC (Saudi Telecommunication Company)
- ○ Mobily
- ○ Zain
- ○ Others: _____

3. For how long did you use the previous telecommunication company?

- ○ Less than one year
- ○ 1-5 years
- ○ 5-10 years
- ○ More than 10 years

4. Before you left the previous telecom company, did you have overdue payments?

- ○ Yes
- ○ No
- ○ I dont remember.

5. Has one of your family members ever used your previous telecommunication company as their cell phone network?

- ○ Yes
- ○ No
- ○ I dont know.

6. Why did you change your previous telecommunication company?

_____

Part C: Communication Methods

1. Do you use the web or social media platforms to communicate with the telecommunication company (for example, for complaints or suggestions)?

- ○ Yes
- ○ No

2. What type of methods do you currently use to communicate with your telecom network (for example, for complaints, requests or suggestions)?

- ☐ Telecommunication Company Twitter account
- ☐ Telecommunication Company IOS/android application
- ☐ Telecommunication Company website

☐ By Telephone

3. Do you think that service quality has been enhanced because of your communication with your telecom company through Twitter (for example for complaints, requests or suggestions)?

○ Yes
○ No

Part D: Customer satisfaction metrics towards the telecom companies

please choose one choice that describes the importance of the customer satisfaction metric toward the services that are provided by the Telecommunication Company that you are currently using for your mobile phone or Internet access, from your perspective:

Services	**importance**	**Very important**	**important**	**Neither important nor unimportant**	**Unimportant**	**Very unimportant**
Good Network Coverage		**5**	**4**	**3**	**2**	**1**
Good Quality of Voice Transmission		**5**	**4**	**3**	**2**	**1**
Quick Response Provided from Customer Service		**5**	**4**	**3**	**2**	**1**
Number of Successful Calls		**5**	**4**	**3**	**2**	**1**
Billing Price		**5**	**4**	**3**	**2**	**1**
High Internet Speed		**5**	**4**	**3**	**2**	**1**
Reasonable Fees When Calling Someone Who Uses Another Telecom Company		**5**	**4**	**3**	**2**	**1**
Good Offers		**5**	**4**	**3**	**2**	**1**

Appendix C: Ethical Approval from the Institutional Review Board (IRB), PNU

Kingdom of Saudi Arabia
Ministry of Education
Princess Nourah bint
Abdulrahman University
(048)
Graduate Studies and Scientific
Research Vice- Rectorate

المملكة العربية السعودية
وزارة التعليم
جامعة الأميرة
نورة بنت عبدالرحمن
(٠٤٨)
وكالة الجامعة للدراسات العليا والبحث العلمي

IRB Registration Number with KACST, KSA: H-01-R-059

June 21, 2020
IRB Log Number: 20-0236
Project Title: A Twitter Analysis to Predict the Satisfaction of Telecom Company Customers
Category of Approval: EXEMPT

Dear Latifah Almuqren,

Thank you for submitting your proposal to the PNU Institutional Review Board. Your proposal was evaluated considering the national regulations that govern the protection of human subjects. The IRB has determined that your proposed project poses no more than minimal risk to the participants. Therefore, your proposal has been deemed ***EXEMPT*** from IRB review. Please note that this approval is from the research ethics perspective only. You will still need to get permission from the head of the department in PNU or an external institution to commence data collection.

Please note that the research must be conducted according to the proposal submitted to the PNU IRB. If changes to the approved protocol occur, a revised protocol must be reviewed and approved by the IRB before implementation. For ***any*** proposed changes in your research protocol, please submit a Request for Modification form to the PNU IRB. Please be aware that changes to the research protocol may prevent the research from qualifying for exempt review and require submission of a new IRB application or other materials to the PNU IRB. In addition, if an unexpected situation or adverse event happens during your investigation, please notify the PNU IRB as soon as possible. If notified, we will ask for a complete explanation of the event and your response.

Please be advised that regulations require that you submit a progress report on your research every 6 months. Please refer to the protocol number denoted above in all communication or correspondence related to your application and this approval. You are also required to submit any manuscript resulting from this research for approval by IRB before submission to journals for publication.

For statistical services you are advised to contact the Data Clinic at the Health Sciences Research Center (hsrc-DC@pnu.edu.sa) or the Scientific Research Center at the Deanship of Scientific Research (dsr-rsc@pnu.edu.sa) extension 30711.

We wish you well as you proceed with the study. Should you have additional questions or require clarification of the contents of this letter, please contact me.

Sincerely Yours,

جامعة الأميرة نورة بنت عبدالرحمن
مجلس المراجعة المؤسسي
Institutional Review
Board (IRB) 21 JUN 2020

Prof. Omar H. Kasule Sr.
Chairman, Institutional Review Board (IRB)
Princess Nourah bin Abdulrahman University, Riyadh, KSA
Tel: +966 548867916
E-mail: irb@pnu.edu.sa

الرقم: ..... التاريخ: ...../...../١٤..هـ المشفوعات: .....

Appendix D: Evaluation questionnaire

Evaluation Questionnaire about customer satisfaction

This questionnaire will evaluate the proposed model that measures customer satisfaction toward Telecom companies in Saudi Arabia. This questionnaire is done as part of research undertaken in Warwick University, UK, and Princess Nourah bint Abdulrahman University, Saudi Arabia. All information provided by you will be confidential and anonymous and used for academic research purposes only. Your participation in this questionnaire is voluntary, and you are entirely free to withdraw at any time you wish.
It will take no more than 2 minutes of your time. Your participation is much appreciated.
Thank you.

What Telecommunication Company are you using for your mobile phone or Internet access in 2017?

- ○ STC (Saudi Telecommunication Company)
- ○ Mobily
- ○ Zain
- ○ Others: _____

What was your satisfaction level toward your Telecommunication Company?

- ○ Satisfied
- ○ Unsatisfied