Sudan University of Science and Technology

College of graduated studies

Sentiment Analysis of Arabic Tweets Written by Sudanese Dialect

تحليل المشاعر للتغريدات العربية المكتوبة بالعامية السودانية

A Thesis Submitted in Fulfillment of the Requirements for

PhD Degree in Computer Science

By

Huda Jamal Abdelhameed Musa

Under the supervision of

Prof. Susana Munoz - Hernandez

September 2019

DEDICATION

To my family

To my teachers

To my husband

To my best friends

To those who gave me all kinds of support,

To all, I dedicate this work

ACKNOLEDGMENT

Praise is to ***Allah***, the Almighty for having guided me at every stage of my life.

I would also like to express my sincere thanks to my research supervisor **prof. Susana Munoz Hernandez** for providing me their precious advices and suggestions. This thesis wouldn't have been success without her comments and suggestions.

Next, I would like to express my brothers and my sisters without their support I would never had dreamt of pursuing higher studies.

Also, I would like to express my husband **Mustafa Mohamed Ahmed** for their unconditional love and support in every part of my life.

Special thanks to **Sudan University of Science and Technology** for providing me such a graceful opportunity to become a part of its family.

Lastly, I would like to thanks all persons those are related to the thesis directly or indirectly.

ABSTRACT

Social networks have become one of the important daily activities in our life. A huge volume of comments is daily generated in social networks. Colloquial Arabic comments have become more widely used between the people's in social networks. Therefore, sentiment analysis of colloquial Arabic comments has become very interesting. There are recognized challenges in this field; some of which are inherited from the nature of the Arabic language itself such as using word "جميل" to express the name of a person and the same word may express feeling. While other problems are derived from the scarcity of tools and sources. This thesis considered sentiment analysis of Arabic tweets which are written in Most Standard Arabic or Sudanese dialectical Arabic. A new lexicon of Sudanese dialect was built which consists of 2500 sentiments. Machine learning techniques which are Support Vector Machine, Naive Bayes, K-Nearest Neighbor and Decision Tree were applied to detect the polarity of the tweets. The results of the first experiment show that, SVM achieved the best Accuracy, Recall and F-measure and it equals 95.1%, 76.5% and 84.4% respectively. While Naïve Bayes achieved best Precision and it equals to 85.1%. The results of the second experiment show that, SVM achieved the best Accuracy and F-Measure and it equals 75.2%, 83.9% respectively. While Naive Bayes achieved best Precision and it equals 88.41%. Also, the best Recall was achieved by Decision Tree and it equals 99.9%. In addition, the percentages of positive and negative opinions toward the Sudanese government services was calculated. 9.4% represents positive opinions related the government services, while 90.6% represents negative opinion.

المستخلص

أصبحت الشبكات الاجتماعية واحدة من الأنشطة اليومية الهامة في حياتنا. نسبة لأنه يتم إنشاء حجم ضخم من التعليقات يوميًا في الشبكات الاجتماعية. كما أصبحت التعليقات العامية العربية أكثر انتشارًا بين الأشخاص في الشبكات الاجتماعية. لذلك أصبح تحليل الآراء للتعليقات العربية العامية مثيراً للاهتمام. هناك تحديات معترف بها في هذا المجال بعضها موروثة من طبيعة اللغة العربية نفسها كاستخدام كلمة "جميل" لتعبر عن اسم شخص وقد تعبر نفس الكلمة عن مشاعر إيجابية. في حين أن البعض الآخر من المشاكل مستمد من ندرة الأدوات والمصادر. ركزت هذه الرسالة على تحليل المشاعر للتغريدات العربية التي كتبت باللغة العربية الفصحى أو العربية العامية السودانية. تم بناء معجم جديد من اللهجة السودانية والذي يتكون من 2500 شعور. تم تطبيق تقنيات التعلم الآلي وهي خوارزمية الدعم الالي، خوارزمية بيز الساذجة، خوارزمية أقرب جار وخوارزمية شجرة القرار للكشف عن قطبية التغريدات. أظهرت نتائج التجربة الأولى أن خوارزمية الدعم الالي حققت أفضل دقة واسترجاع وقياس دقة الاختبار وهي تساوي 76.5٪ ،95.1٪ ،84.4٪ على التوالي. بينما حققت خوارزمية بيز الساذجة أفضل دقة وتساوي 85.1 ٪. كما اظهرت نتائج التجربة الثانية أن خوارزمية الدعم الالي حققت أفضل دقة وقياس دقة الاختبار وهي تساوي 75.2٪ و83.9٪ على التوالي. بينما حققت خوارزمية بيز الساذجة أفضل دقة وتساوي 88.41٪. ايضأ تم تحقيق أفضل استرجاع بواسطة خوارزمية شجرة القرار وتساوي 99.9 ٪. بالإضافة إلى ذلك، تم حساب النسب المئوية للآراء الإيجابية والسلبية تجاه الخدمات الحكومية السودانية. وجد ان 9.4٪ تمثل آراء إيجابية تجاه الخدمات الحكومية، بينما 90.6٪ تمثل آراء سلبية.

TABLE OF CONTENTS

DEDICATION……..II

ACKNOLEDGMENT III

ABSTRACT IV

المستخلص V

Table of Contents VI

List of Tables IX

List of Figures X

List of Equations XI

CHAPTER ONE INTRODUCTION

1.1 Overview 1

1.2 Motivation 2

1.3 Problem statement and significance 3

1.4 Related Works and Open Issues 4

1.4.1 Related Works 4

1.4.2 Open Issuse 5

1.5 Research Hypotheses 6

1.6 Research Objectives 6

1.7 Research Methodology 7

1.8 Research Scope 8

1.9 Research Contribution 8

1.10 Thesis Structurer 9

CHAPTER TWO BACKGROUND

2.1 Introduction 10

2.2 Sentiment Analysis and Opinion Mining 10

2.3 Challenges of Sentiment Analysis 11

2.4 Application of Sentiment Analysis 12

2.5 Social Networks Analysis 13

2.6 Arabic Language 15

2.7 Methodologies Used for Sentiment Analysis 7

2.8 Sentiment Analysis Approaches 17

2.9 Classification Techniques 19
2.9.1 Unsupervised Approach 19
2.9.2 Supervised approach 19
2.9.2.1 Support Vector Machines 20
2.9.2.2 Naive Bayes 23
2.9.2.3 Maximum Entropy 25
2.9.2.4 K-Nearest Neighbor 26
2.9.2.5 Decision Tree 28
2.10 Models of the vector representation of text data 29
2.10.1 Bag-of-Words 29
2.10.2 Bag-of-N-Grams 30
2.11 Cross Validation (CV) 31
2.12 Twitter API 31
CHAPTER THREE RELATED WORKS
3.1 Introduction 33
3.2 Related Studies 33
CHAPTER FOUR METHODOLOGY
4.1 Introduction 51
4.2 Data Collection 52
4.3 Data Preprocessing 55
4.3.1 Data Cleaning: 55
4.3.2 Removing Duplicated Characters: 56
4.3.3 Tokenization: 56
4.3.4 Normalization: 56
4.3.5 Handling Negation: 56
4.3.6 Removing Stopwords: 59
4.3.7 Stemming: 59
4.3.8 Term Weight: 60
4.3.8.1 Boolean Model: 60
4.3.8.2 Term Frequency 60
4.3.8.3 Inverse Document Frequency 61
4.3.8.4 Term Frequency-Inverse Document Frequency 61
4.4 Sentiment Classification 62

4.5 Evaluation 63
4.6 RapidMiner Tool..... 65
CHAPTER FIVE RESULTS AND DISCUSSION
5.1 Introduction..... 67
5.2 Results..... 67
5.2.1 Results of Expermient 1..... 67
5.2.2 Results of Expermient 2..... 71
CHAPTER SIX CONCLUSION AND FUTURE WORK
6.1 Conclusion 78
6.2 Future Works 79
Biblography ……...81
APPENDIX A: SAMPLE OF DATASET……...89
APPENDIX B: QUSETIONERS USING GOOGLE FORM…….…….…….…...95
APPENDIX C: PUBLISHED PAPERS…….……...…...…...96

LIST OF TABLES

Table 1.1: Analysis of related works……..……..4
Table 3.1: Structured of the collected reviews…….36
Table 3.2: Description of Corpus used in the experiments……...37
Table 4.1: Questioner details…….53
Table 4.2: The collected datasets……...54
Table 4.3: Size of the dataset, training and testing dataset details…….54
Table 4.4 Arabic negation words……...58
Table 4.5: Some derivations of the root "لعب" ……..59
Table 4.6: Example of preprocessing a tweet……59
Table 4.7: Sample of classified tweets……..62
Table 4.8: Confusion matrix for two classes positive and negative……...64
Table 5.1: True Positive and True Negative for the SVM…….68
Table 5.2: True Positive and True Negative for the Naive Bayes…….68
Table 5.3: True Positive and True Negative for the K-Nearest Neighbor…….68
Table 5.4: Precision, Recall, Accuracy and F-Measure for the classifiers……....69
Table 5.5: Measures of SVM with K-folds Cross-validation……70
Table 5.6 Measures of NB with K-folds Cross-validation……71
Table 5.7: Measures of DT with K-folds Cross-validation……..……72
Table 5.8: Precision, Recall, Accuracy and F-Measure for the classifiers……73

LIST OF FIGURES

Figure 1.1: Methodology steps…….……7
Figure 2.1: Sentiment Analysis in Social Networks……11
Figure 2.2: The number of monthly active user accounts of social media (in Millions) …14
Figure 2.3: Twitter monthly active users from 2017-2019 (in Millions) ……...15
Figure 2.4: An example of different people's sentiment……..17
Figure 2.5: Sentiment Analysis methodologies……..……..18
Figure 2.6: Sentiment classification approaches……...…19
Figure 2.7: Supervised classification techniques……..20
Figure 2.8: Example of SVM Schema……..21
Figure 2.9: Example of Linear SVM……23
Figure 2.10: Naïve Bayes classifier……..23
Figure 2.11: Example of K-NN Classification…….27
Figure 2.12: Decision tree structure……..28
Figure 3.1: Results of the Kanakaraj and Guddeti study……...48
Figure 4.2: Sources of building the Sudanese Dialectical Lexicon……...54
Figure 4.3: Data preprocessing Steps…….55
Figure 5.1: Confusion Metrics For SVM……...69
Figure 5.2: Confusion Metrics For NB……..69
Figure 5.3: Confusion Metrics For KNN……...70
Figure 5.4: Evaluation measures for SVM, NB, K-NN…….71
Figure 5.5: Comparison between the results of SVM based on K-folds cross validation…….73
Figure 5.6: Comparison between the results of NB based on K-folds cross validation……74
Figure 5.7: Comparison between the results of KNN based on K-folds cross validation…….75
Figure 5.8: Accuracy, Precision, Recall, F-measure for SVM, NB, KNN classifiers…..…….76

LIST OF EQUATIONS

Equation 2.1: Linear Model of SVM Classifier……..22

Equation 2.2: Linear Classification…….22

Equation 2.3: Probability of Naïve bayes ……...25

Equation 2.4: Probability of Naïve bayes when P (d) is constant……25

Equation 2.5: The binary bag-of-words Model……29

Equation 2.6: The bag-of-words Vector……...29

Equation 4.1: Term frequency (TF) …….60

Equation 4.2: Inverse Document Frequency (IDF)……...61

Equation 4.3: Term Frequency and Inverse Document Frequency (TF-IDF) …….62

Equation 4.4: Accuracy Measure……..64

Equation 4.5: Precision Measure……..65

Equation 4.6: Recall Measure……...65

Equation 4.7: F-score Measure…….65

CHAPTER ONE

INTRODUCTON

Chapter One
Introduction

1.1 Overview

In recent years, the use of Internet has become one of the daily activates in our life. Social media constitutes a major component of the Web 2.0 and includes social networks, blogs, forum discussions, micro-blogs. Users of social media generate a huge volume of comments on a daily basis. These comments reflect their opinions about different issues such as products, news, entertainments, or sports. Therefore, different organizations may be interesting on analyzing these comments (Al-Kabi *et al.*, 2014). In general, the analysis of social media has attracted a great deal of attention recently and the motivation is not only related marketing reasons, but also security and privacy reasons.

Opinion mining or sentiment analysis is the field of science that is interested in extracting opinions embedded in customer's comments (Montoyo *et al.*, 2012). It has been extensively studied in the literature for the English language (R M Duwairi, Ahmed and Al-Rifai, 2015). By comparison, relatively few works have targeted sentiment analysis in Arabic text.

There are several granularities for sentiment analysis. A popular work is to determine whether a text is subjective or objective. Another common work is to determine whether a text is written to express a positive or negative opinion (Hedar and Doss, 2013) Sentiment analysis deal with extracting the polarity of the text (positive, negative or neutral). A third category deals with finding the strength of an emotional state in text. Such as "happy", "sad" and "angry" (Pang and Lee, 2004). There are two approaches for

detecting sentiment in text. The first one relies on linguistic resources such as dictionaries and lexicons. The second one is based on machine learning (Melville, Gryc and Lawrence, 2009). Some researchers have combined the previous two approaches. The lexicons are very hard to build manually and they are depending on the domain.

Sentiment analysis is hard to detect for many reasons; one reason is that people use different writing styles to express their opinions. A second reason is that sentiment is context dependent (Wilson, Wiebe and Hoffmann, 2009).

1.2 Motivation

In today's connected world, users can send messages in any time. However, social media is not only used as a casual tool for messaging and sharing private things and thoughts; it is also used by journalists, politicians and public figures, series of companies and universities who want to be more open to the public, share their thoughts and take an interest in opinion of persons. The active growth of the audience of social media on the Internet led to the formation of these resources as a new source of the people's mood and opinion (Al-Kabi, Abdulla and Al-Ayyoub, 2013).

Researchers note that the billions of publications left by people monthly, cannot be processed manually by holding public opinion polls. This fact highlights the need for automated methods of intellectual analysis of text information, what allows in a short time to process large amounts of data and to understand the meaning of user messages. This understanding of the meaning of messages is the most important and complex element of the automated processing. Use of modern technologies and methods of big data, using artificial intelligence, has already been helping researchers to automate the process of content analysis, in particular to collect data, to prepare, to manage and to

visualize data. These innovations give the opportunity to conduct large-scale research and to monitor social media in real-time.

Existing sentiment analysis techniques occur from the fields of natural language processing, computational linguistics, text mining, and a range from machine learning methods to rule-based methods. Machine learning methods involve training of models on specific collections of documents. Recently, many researchers deal with the determination of sentiment of people in various data collected from social media. They have used well-known machine learning techniques for classification and clustering data. However, this thesis focuses on sentiment classification and compare of existing machine learning techniques applied to sentiment analysis of data collected from the social network Twitter.

1.3 Problem Statement and its Significance

Arabic is a morphologically complex language that has a high inflectional and derivational nature. There are many challenges faced Arabic language for example using word "سليم" to express the name of a person and the same word may express opinion about specific thing. In another hand, most of the communications within the social media context is carried out using colloquial Arabic rather than the Most Standard Arabic (Harrat, Meftouh and Smaili, 2017). However, the major problem that faced sentiment analysis of Arabic data is unavailability of sentiment lexicons. There are no publicly available Colloquial Arabic sentiment lexicons in Sudanese Dialect. That's why we need to generate lexicons by our self. In addition, the accuracy of available tools in Arabic is still not comparable to accuracies obtained for other languages. Many Arabic sentiment analysis tools require the use of other tools to accomplish the tasks.

The significance of this work is helping the marketing department in organizations to analysis the customer's comments about their products and knowing about their satisfaction.

1.4 Related Works and Open Issues

A lot of related studies has been carried out to address the sentiment analysis of Arabic and other languages.

1.4.1 Related Works

Table 1.1 below summarized some related works focused on sentiment analysis using several techniques.

Table 1.1: Analysis of Some Related Works

Author	Techniques	Dataset	Language
(Jain and Jain, 2019)	SVM, NB and KNN	1265 text	English
(Al-Kabi *et al.*, 2018)	SVM, NB and KNN	3015 opinion	MSA
(Soliman and Ali, 2013)	SVM, DT, CNB, and KNN	-	MSA
(Al-Kabi, Abdulla and Al-Ayyoub, 2013)	SVM, NB	4625 comment	MSA
(El-halees, 2011)	KNN	8793 statements	MSA
(Mohammed *et al.*, 2014)	Naive Bayes	1080 Reviews	MSA and colloquial
(Taysir H. A. Soliman *et al.* 2014)	SVM	1846 comment	MSA

Author	Techniques	Dataset	Language
(Mohammed R. et al. 2011)	SVM Naïve Bayes	500 reviews	MSA
(Mohamed Elhawary et al., 2010)	SVM	1600 words	MSA
(Afnan A. Al-Subaihin et al. 2011)	SVM, Naïve Bayes and Maximum Entropy	-	Colloquial Arabic
(Rehab M. and Nizar A. Ahmed 2015)	SVM	4400 tweets	English
(Moghaddam and Popowich, 2002)	Naïve Bayes	30 adjectives	English
(Phua and Yee Ling, 2013)	Naïve Bayes classifier	5000	Colloquial Singapore English
(Alexander Pak and Patrick, 2010)	Naïve Bayes and SVM	216 sentiments	English

1.4.2 Open Issues:

The main difficulty of performing sentiment analysis in Arabic social media lies in the fact that communication within the social media context is carried out using “spoken” or colloquial Arabic rather than the more formal Most Standard Arabic (MSA). Not only is the vocabulary of colloquial Arabic different than that of MSA, the structure of the sentences is much more random which is why parsing this text poses a major challenge. Some of the problems that are bound to face anyone working on colloquial Arabic sentiment analysis such as unavailability of colloquial Arabic parsers, unavailability of

sentiment lexicons, the need for person name recognition and handling the compound texts.

1.5 Research Hypotheses

The main hypotheses that are claimed by this thesis are presented below.

Hypothesis 1: there is no Arabic lexicons are provided for the sentiment analysis in the Sudanese dialect.

Hypothesis 2: the Arabic language needs a more variety of features and representation, such as handling the duplicated characters and classifying the texts which includes emotions and smiley.

Hypothesis 3: different machine learning techniques such as Support vector machine, Naïve bayes and Decision Tree are doing well and gave best result with Arabic language.

Hypothesis 4: applying Bigram Model representation in the Arabic sentiment classification would have a big impact on the accuracy of the classifier and outperform 14% of Unigram and Trigram.

Hypothesis 5: having an awareness of the negation while analyzing the sentiment in the Arabic language leads to the best accuracy.

1.6 Research Objectives

This thesis aims to build a sentiment analysis approach based on text mining techniques. This approach will analyze and classify tweets which written in Most Standard Arabic (MSA) or Sudanese dialect in the social networks to determine positive and negative sentiments. The approach will be achieved with the following sub-objectives:

- Building a lexicon for Sudanese dialect.
- Classifying the sentiments (in the lexicon) manually into positive and negative classes.
- Several specific sub-objectives while preprocessing tweets such as removal of duplicated character from the tweets, handling negation and handling tweets which included emotions and smiley (happy and sad).
- Determining the polarity of the tweets which are written in MSA or Sudanese dialect.

1.7 Research Methodology

To solve the problem which discussed above, there are many phases were done. The first phase is collecting data**.** In this phase, Arabic and Sudanese words from different sources were collected**.** The second phase is preprocessing data. Before determining the polarity of the collected tweets, preprocessing of the collected tweets are necessary to get the cleaned data. preprocessed tweets will be applied as input of the model. This phase also includes tokenization, stopword filtering, steaming and normalization. The third phase is sentiment classification. In this phase SVM, NB, KNN and DT were used as classification techniques to classifying tweets based on new generated lexicon into one of the two categories as positive or negative. The last phase is Evaluation**.** In this phase, the classification performance, accuracy, precision, recall and f-measure were calculated. Figure 4.1 below show these phases.

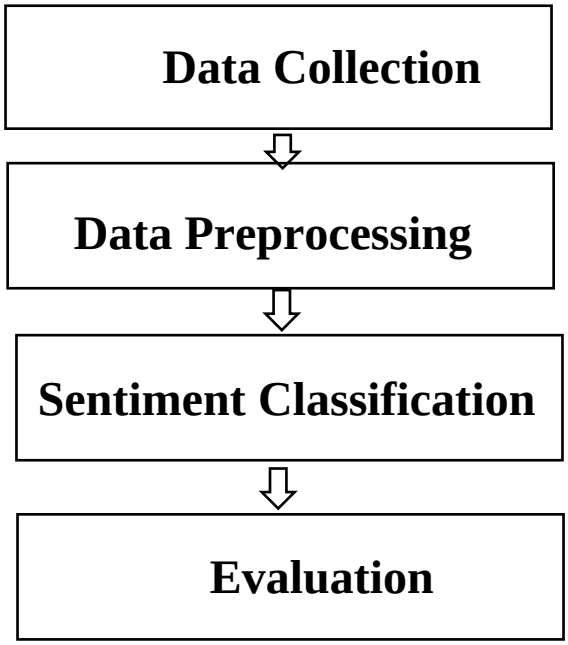


Figure 1.1: Methodology Steps

1.8 Research Scope

This thesis focuses on sentiment classification of tweets which written in (MSA or SDA) to positive and negative classes.

1.9 Research Contribution

In this thesis there are main contributions as follows:

- A new lexicon for Sudanese dialect was built.
- Tweet related Sudanese revolution was collected and classified based on its polarity to positive and negative opinions and the percentages of each class was calculated.
- Handling duplicated characters by remove the repeated characters in the text and eliminated it to two character, for example the word "شدييييييد " after removing the duplicated "ي" became "شديد "
- Handling negation in tweets. There are negated words in Arabic such as (لا، لن،

ليس، ما) these words change the text polarity, for example the text “الفلم دا حلو ” is positive opinion related the movie but the text “الفلم دا ما حلو ” is negative opinion. So, the negated word “ما ” changed the polarity of the text from positive to negative.

- Handling emotions (happy and sad) which included in the tweet.

1.10 Thesis Structure

This thesis consists of six chapters as follow.

Chapter Two: background, talking about sentiment analysis and opinion mining, Arabic language challenges, text mining, concept of sentiment classification, machine learning tools, classification techniques and comparison of different studies related to sentiment classification in Arabic language.

Chapter Three: related works, discuss related studies in the sentiment analysis and what was done before to solve the same problem.

Chapter Four: methodology, details of data collection, data preprocessing, sentiment classification techniques which have used to classify the dataset and RapidMiner tool.

Chapter Five: experimental results, discuss the results of sentiment classification.

Chapter Six: conclusion, and future work explained in this chapter.

CHAPTER TWO

BACKGROUND

Chapter Two
Background

2.1 Introduction

Since sentiment analysis in social media requires good knowledge of sentiment analysis, and methods, in this chapter a proper overview of all of these related concepts is provided. The first part of this chapter deals with opinion mining and sentiment analysis. In the second part social network analysis were discussed. In the third part deferent algorithms for sentiment analysis and classification algorithms were discussed which becomes the basis of the experiments in Chapters 4. After that, the evaluation measures are represented.

2.2 Sentiment Analysis and Opinion Mining

Opinion Mining is the process of a set of search results for a given item, generating a list of attributes (quality, features, etc.) and aggregating opinions about each of them (poor, mixed, good). Much of the subsequent research self-identified as opinion mining fits this description in its emphasis on extracting and analyzing judgments on various aspects of given items. However, the term has recently been interpreted more broadly to include many different types of analysis of evaluative text (Lotte *et al.*, 2007).

Also, in (Saleh *et al.*, 2011) defined opinion mining as a field of the Text Mining (TM) that has been designated by different terms like subjectivity analysis, sentiment analysis or sentiment orientation. There are lots of definitions for each one. (Pang and Lee, 2008) captured different definitions about these terms based on applications done in this field. For example, Subjectivity Analysis is defined as the recognition of opinion-oriented language in order to distinguish it from objective language.

Figure 2.1 below show the processes related to sentiment analysis in social networks.

Figure 2.1: Sentiment Analysis in Social Networks (Lin *et al.*, 2014)

2.3 Challenges of Sentiment Analysis

Generally, sentiment analysis or classification is considered a special case of text classification in a natural language processing. Although the number of classes in sentiment analysis are small, the process of sentiment classification is more difficult than the traditional topic text classification (Liu, 2009). In topic text classification, classification relies on using keywords, but this does not generally work well in the case of sentiment analysis (Turney, 2002).

The other difficulties in sentiment analysis come from the nature of this problem. Sometimes, the negative sentiment might be expressed in a sentence without using any obvious negative words. Moreover, there is a fine line between whether a sentence should be labeled objective or subjective. Determining the opinion holder -the one who expresses the sentiment in the text- is one of the most difficult tasks in sentiment analysis. The sentiment analysis highly depends on the domain of the data. The words sometimes

have positive sentiment in a specific domain, whereas they have another polarity sentiment in a different domain. Finally, some other writing styles such as irony, sarcasm, or negated sentences could bring more challenges to sentiment analysis (Liu, 2009).

2.4 Application of Sentiment Analysis

In a marketplace, businesses realize the importance of the internet in gathering user's opinions and reviews about their products and services. Time is more valuable to businesses than to normal users. Normal users often spend some time surfing the internet in order to establish the opinions of other users, while businesses generally need an automated system that can help them ascertain the sentiments and opinions of users of their products and services. A tool that can obtain and analyze user reviews in order to understand the final sentiment is more valuable to businesses. This tool may provide them with the feelings of customers and ideas that help them to improve their products and services.

The World Wide Web provides a great place that the people gain knowledge from the information. There is no need to ask a friend when you are wanting to buy a product, going on a vacation, or needing some services. The only thing that you need is the internet to surf through this unstructured information. Therefore, sentiment analysis should be able to surf this information and bring it in structured format to the end users.

Nowadays, people tend to use the internet to broadcast their thoughts and ideas about topics or issues by using forums or other social networks. Some of these ideas are positive, while others are more violent in manner and content.

Therefore, sentiment analysis has the potential to be more valuable in these cases in monitoring the sentiment of groups over the internet. This helps the government to discover any violence at an early stage and to begin to deal with it before it expands (Alotaibi, 2015).

2.5 Social Networks Analysis

Social media can be referred to as the "group of internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of user-generated content", as defined by (Vargas *et al.*, 2016). In recent years in addition to the leaders of the World Wide Web such as Facebook, Google+, LinkedIn and Twitter, there are new services for different groups of users: social network for students, the network for specific groups of professionals, communities of ethnic minorities, and even a special network for all the world's drinkers. This extends the scope to very deferent kinds of research from consumer preferences to psychological characteristics. As follows from the Figure 2.1, in early 2015 Facebook retained the first place among social platforms, and also Twitter was in the top ten. According to the same study by (Vargas *et al.*, 2016), more than 2 billion people worldwide are active users of social networks and blogs.

Facebook dominates the global social media landscape, claiming 2.3 billion actives users in March 2019. Meanwhile, instant messenger services and chat apps continue to grow, with WhatsApp, WeChat, Facebook Messenger and Viber all reporting more than 100 million new monthly active users over the past 2014. Instant messenger services and chat apps now account for 3 of the top 5 global social platforms, and 8 instant messenger

brands now claim more than 100 million monthly active users. In Twitter, the number of monthly active users is 330 million in 2019.

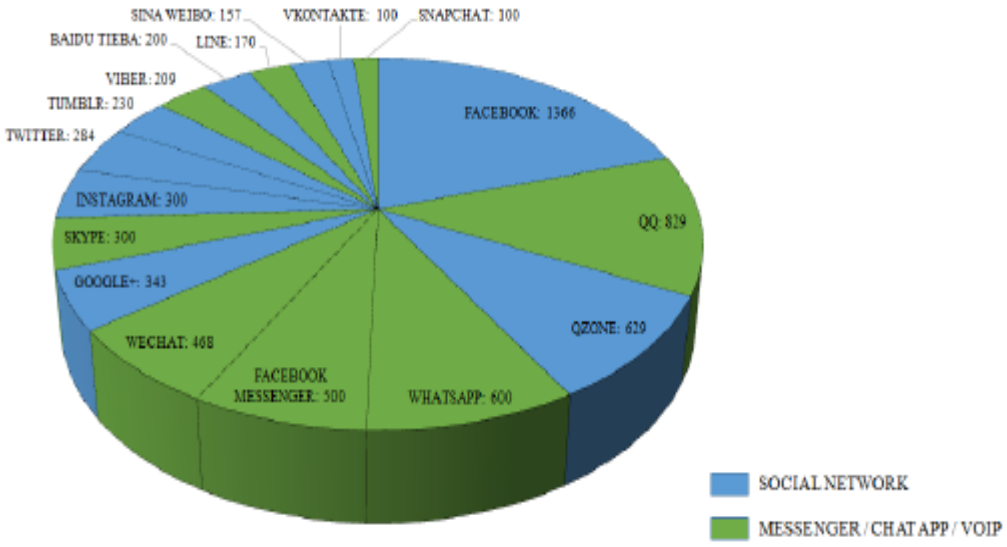


Figure 2.2: The number of monthly active user accounts of social media (Vargas *et al.*, 2016)

A social network analysis (SNA) examines the structure of social relationships in a group to uncover the informal connections between people. In a consulting setting, these relationships are often ones of communication, awareness, trust, and decision-making. As an approach for looking at these relationships, SNA has been used a long time ago. It assumes that people are all interdependent teach other. This assumption is radically different from traditional research approaches which assume that what people do, think, and feel is independent of who they know (Ehrlich and Carboni, 2005). Figure below show the number of monthly active user accounts of twitter till 2019.

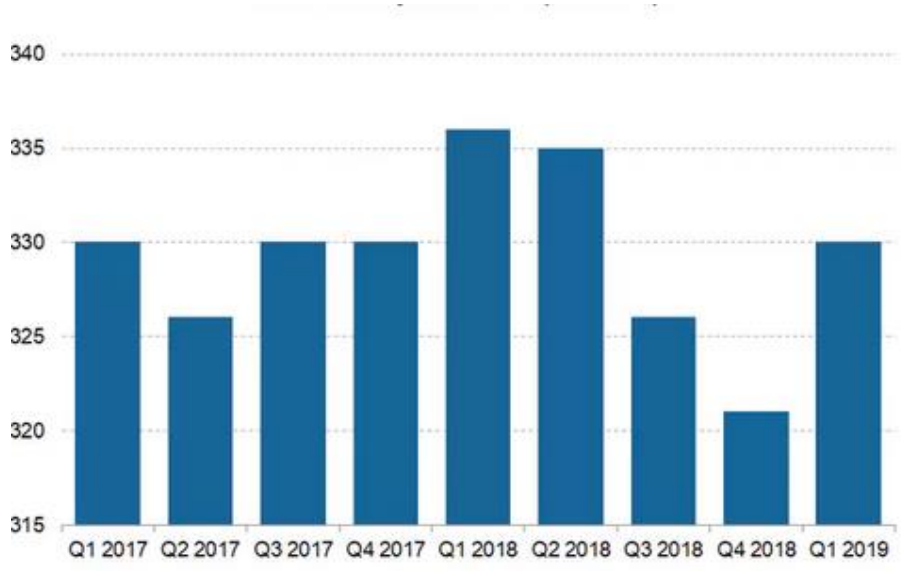


Figure 2.3: Twitter Monthly Active Users From 2017-2019 (Vargas *et al.*, 2016)

2.6 Arabic Language

Arabic Language is one of the widely used languages in the world. Arabic language is a Semitic language that has a complex and much morphology than English; it is a highly inflected language and that due to this complex morphology (Al-Harbi *et al.*, 2008).

Arabic Language consists of 28 alphabet characters: ا ب ت ث ج ح خ د ذ ز ر ش س ش ص ض ع ط ظ ع غ ف ق ك ل و ي. In addition to the hamza (ء) which is considered as a letter by some Arabic linguistics. Arabic is written from right to left. Arabic letters have different styles when appearing in a word depending on the letter position at beginning, middle or end of a word and on whether the letter can be connected to its neighbor letters or not (Alsaleem, 2011).

There are two types of Arabic sentences, nominal and verbal, these are determined by the part-of- speech of the first word in a sentence. A nominal sentence has no verb. It is formed of a subject and a predicate. These vary from very simple forms to more complicated sentences (M. K. Saad, 2010). The simple nominal sentence consists only

of nouns and adjectives, whereas the subject is composed of two words, and the predicate is another sentence within a complicated one. Arabic words may work with three types of affixes: prefixes, infixes, and suffixes. Affixes may be one letter long or a combination of multiple letters. In addition to their complex nature, the level of ambiguity of Arabic morphemes is notable. Determining whether a letter is an affix or part of the stem is not an easy task, especially when there is an absence of short vowels. These characteristics affect the NLP tools that deal with Arabic, such as the part-of-the speech tagger, morphology analyzer, name entity recognition and syntactical parsing (Ryding, 2005).

One of the problems in Arabic is using a noun with positive polarities as a person names such as the word (سليم Saleem); which means Right in English. سليمas an adjective indicates positive sentiment but as a person name it is neutral (i.e. it has no sentiment).

In general Arabic is divided into three types: Classical Arabic, Modern Arabic, and Colloquial Arabic. As the official language of 22 countries, there are 49 million Arab users of Facebook (Rushdi-Saleh *et al.*, 2011). Arabic language is a high complex language, which embeds five critical challenges for Natural Language Processing (NLP) tasks. First, Arabic is not a case-sensitive language; it has no capital letters. Second, Arabic is a high inflectional language; often a single word has more than one affix, such that it may be expressed as a combination of prefix(s), lemma, and suffix(s) (Soliman *et al.*, 2014). Third, Arabic has some variants in spelling and typographic forms. Fourth, Arabic texts have different meanings. For example, “ رجبRagab” in Arabic may be used as a person name, or month. Fifth, Arabic resources, such as corpora, gazetteers, and NLP tools, are not free (Helmy and Daud, 2010).

2.7 Sentiment Analysis Methodologies

A large range of approaches are used to investigate the problem of sentiment analysis. Most of these approaches are built to deal with the English language as it is the dominant language of science. However, this should not stop researchers from building techniques that work with other languages, such as Chinese, Korean, Japanese and Arabic.

There are two main approaches that are found in the literature to analyzing sentiment as shown in the figure below.

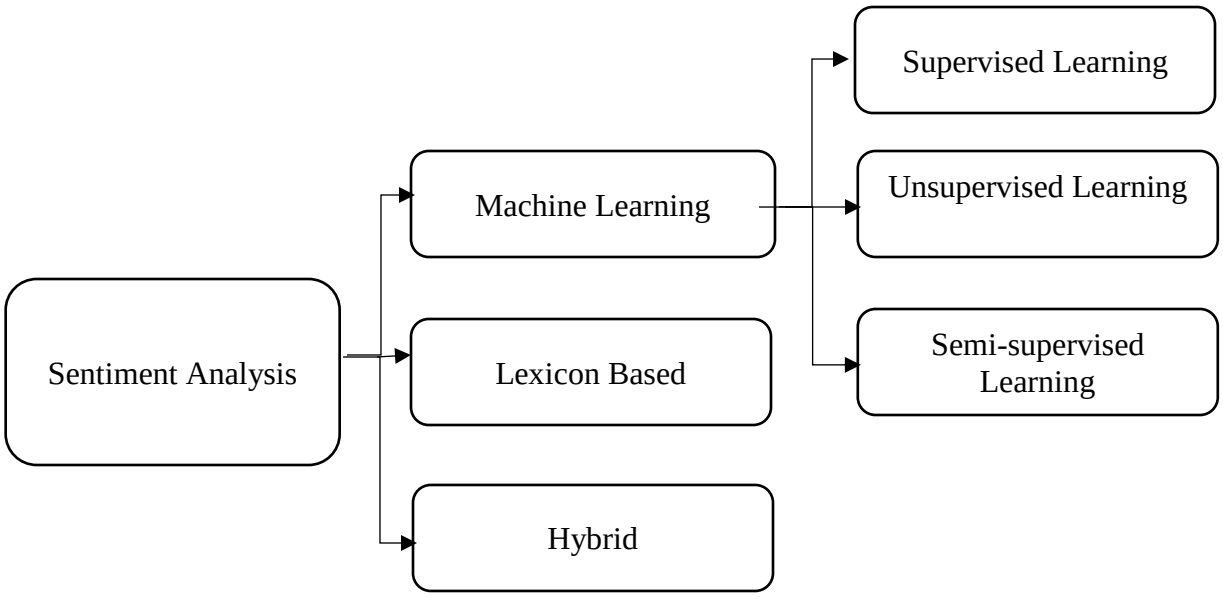


Figure 2.4: Sentiment Analysis methodologies (Neethu and Rajasree, 2013)

2.8 Sentiment Analysis Approaches

The term “sentiment” is used in reference to the automatic analysis of evaluative text and tracking of the predictive judgments. Figure 2.5 Below show an example of different people’s sentiments.

Figure 2.5: An example of different people's sentiment (Fisher *et al.*, 2012)

There are many Sentiment Analysis (SA) algorithms, however, most of them can be grouped into one of two main approaches: corpus-based and lexicon-based. the corpus-based approach (supervised approach) starts with a dataset (or corpus) of labeled examples and extracts discriminative features of the examples of each class. These features are fed into a classification algorithm such as Naive Bayes (NB) or Support Vector Machine (SVM). Lexicon-based approach (unsupervised approach) uses a lexicon (dictionary) composed of words of each sentiment. For example, a typical lexicon may contain a list of positive words or and a list of negative words. A word that does not appear in either list is considered neutral. Once the lexicon is ready, the sentiment of any new example is simply determined based on the words or phrases it contains and to which list they belong (Al-Kabi, Abdulla and Al-Ayyoub, 2013). Figure 2.5 below show the sentiment classification approaches.

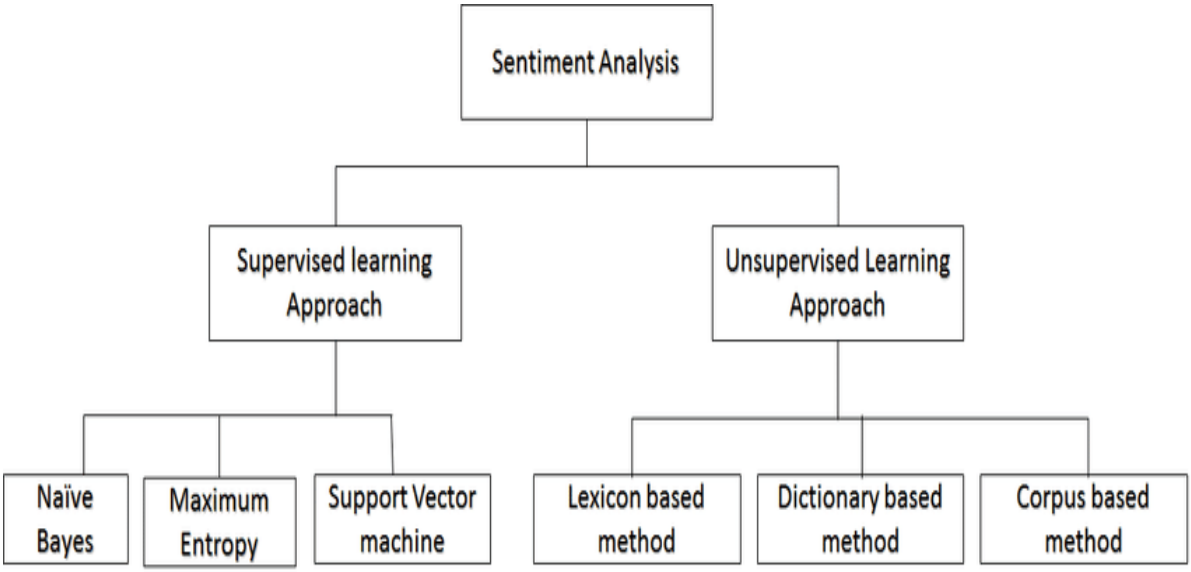


Figure 2.6: Sentiment Classification Approaches (Jagdale, Shirsat and Deshmukh, 2016)

2.9 Classification Techniques

The task of classification techniques is to determining whether a sentence or comment has either positive or negative. The approaches for this task can be decomposed into two approaches: the unsupervised approach and the supervised approach.

2.9.1 Unsupervised Approach

In his model, sentiment orientation SO of a phrase is estimated analyzing.

The set of pre-defined positive words such as “excellent, good”, and the set of pre-defined negative words such as “poor, bad” (Go, Bhayani and Huang, 2009).

2.9.2 Supervised approach

This approach is based on the supervised machine learning-based method. The learning process is driven by the knowledge of the categories (positive/negative, in this comment) and of the training instances that belong to them. (Pang et al. 2002) examined with three machine learning methods: Support Vector Machines Classification, Naive Bayes

Classification, and Maximum Entropy (Go, Bhayani and Huang, 2009). Figure 2.6 below show supervised classification techniques.

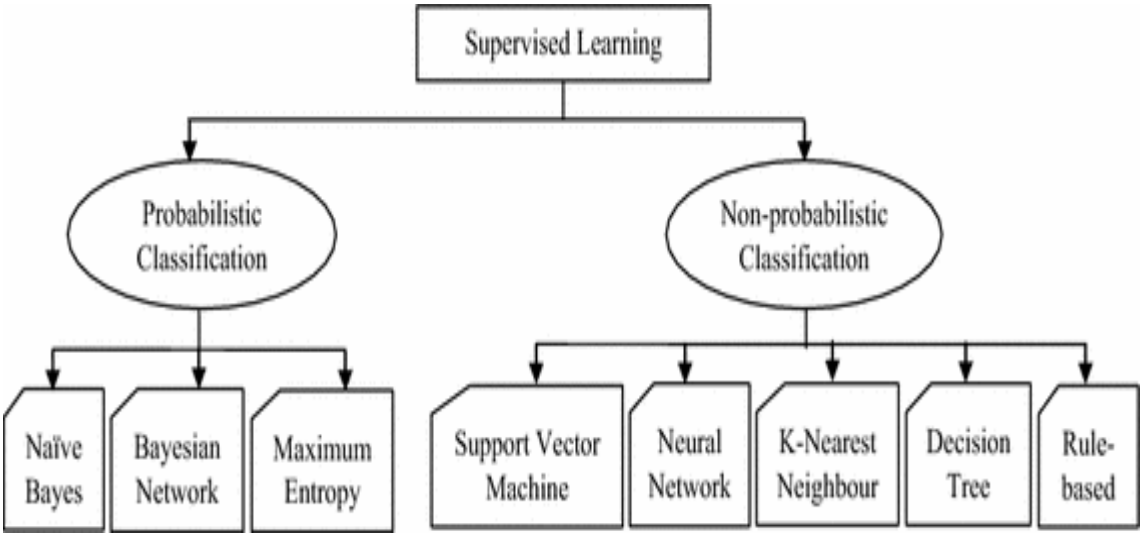


Figure 2.7: Supervised Classification Techniques (Jagdale, Shirsat and Deshmukh, 2016)

2.9.2.1 Support Vector Machines

Support Vector Machines (SVM) is one of the discriminative classification approaches which is commonly recognized to be more accurate. SVM classification approach is based on Structural Risk Minimization (SRM) principle from statistical learning theory. SRM is an inductive principle for model selection used for learning from finite training data and it provides a method for controlling the generalization ability of learning machines that uses a small size training data. The idea of this principle is to find a hypothesis to guarantee the lowest true error. In addition to this, the derivation of SVM is mathematically rigorous and very open to theoretical understanding and analysis.

SVM needs both positive and negative training datasets which are uncommon for other classification methods. It is outstanding from the others with its better classification performance and its ability in handling documents with high-dimensional input space

and culls out most of the irrelevant features. The good generalization characteristic of SVM is due to the implementation of SRM which entails finding an optimal hyper-plane, thus guaranteeing the lowest classification error. Besides, a capacity which is independent of the dimensionality of the feature space makes SVM a highly accurate classifier in most applications. However, the major drawback of SVM is its relatively complex training and categorizing algorithms and also the high time and memory consumptions during the training stage and classifying stage due to its convoluted training and categorizing algorithms. Besides, confusions occur during the classification tasks because the documents could be annotated to several categories because of similarities are typically calculated individually for each category (Slamet *et al.*, 2018). Support Vector Machines are based on the concept of decision planes that define decision boundaries. A decision plane is one that separates between a set of objects having different class memberships. A schematic example is shown in the illustration below. In this example, the objects belong either to class GREEN or RED. The separating line defines a boundary on the right side of which all objects are GREEN and to the left of which all objects are RED. Any new object (white circle) falling to the right is labeled, i.e., classified, as GREEN (or classified as RED should it fall to the left of the separating line).

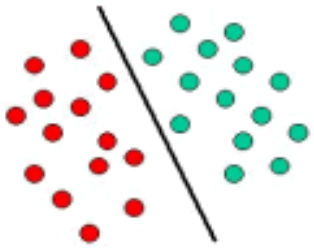


Figure 2.8: Example of SVM Schema (Xu, Ding and Wang, 2007)

An SVM also uses a discriminant hyperplane to identify classes. However, concerning SVM, the selected hyperplane is the one that maximizes the margins, i.e., the distance from the nearest training points. Maximizing the margins is known to increase the generalization capabilities. As LDA, an SVM uses a regularization parameter that enables accommodation to outliers and allows errors on the training set (Pang and Lee, 2008). SVM is used to find a linear model of the following form:

$$y(x) = w^T x + b$$

Equation 2.1: Linear Model of SVM Classifier

Where x is input vector, w and b are parameters which can be adjusted for a certain model and estimated in an empirical way. In simple linear classification the task is to minimize a regularized error function given by Equation below.

$$C\sum_{n=1}^{N}\xi_n + \frac{1}{2}\|w\|^2$$

Equation 2.2: Linear Classification

Figure 2.8 below illustrates an example of a linear SVM that has been trained on examples from two classes. Here the SVM constructs a separating hyperplane and then tries to maximize the" margin" between the two classes. To calculate the margin, the SVM con-structs two parallel hyperplanes, one on each side of the initial one. These hyperplanes are then "pushed" perpendicularly away from each other until they come in contact with the closest examples from either class. These examples are known as the support vectors and are illustrated in bold in Figure 2.8 below.

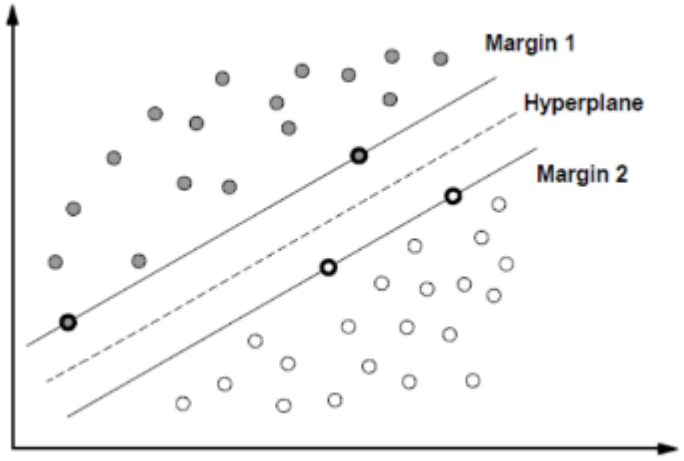


Figure 2.9: Example of Linear SVM (Ryding, 2005)

2.9.2.2 Naive Bayes

The Naive Bayes model is an old method for classification and predictor selection that is enjoying a renaissance because of its simplicity and stability.

Also, (Zhang, 2004) defined Naive Bayes as one of the most efficient and effective inductive learning algorithms for machine learning and data mining. Its competitive performance in classification is surprising, because the conditional independence assumption on which it is based, is rarely true in real world applications. Figure 2.9 below show the naïve bayes classifier.

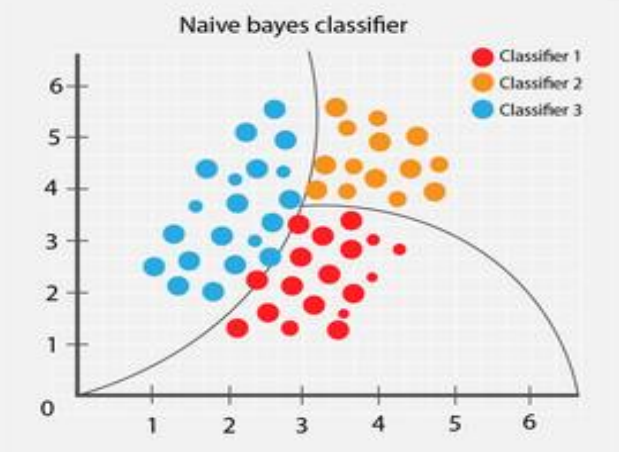


Figure 2.10: Naïve Bayes Classifier (Fisher *et al.*, 2012)

Naïve Bayes based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. Even if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability. Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods (El-Beltagy and Ali, 2013). Naïve Bayes having strong independence assumption (naïve). Previous studies on algorithms of classification have proven that Naïve Bayes is one of the best algorithms in comparison with the other ones such of Decision Tree, Naïve Bayes, and KNN. It has also been found that accuracy and speed are the most supporting and helpful features of the algorithm in classifying data (Rajeswari, Juliet and Aradhana, 2017). The advantage of Bayesian classifier is that it requires small training data set for classification. It is easier for implementation, fast to classify and more efficient. It is non sensitive to irrelevant features. It is used in personal email sorting, document categorization, email spam detection and sentiment detection (Rajeswari, Juliet and Aradhana, 2017). In Naïve bayes algorithm the probability that a document (d) belongs to class (c) is calculated as follows.

Where P (d | c) is the probability of generating instance d given class c, P (c) is the probability of the occurrence of class c, and P (d) is the probability of instance d occurring. P (d | c) is difficult to estimate due to the number of possible vectors; d is too high. By using the naïve assumption, the difficulty can be overcome so that any two coordinates of the document are statistically independent (Aghila, 2010).

$$P(c \mid d) = \frac{P(d \mid c)\, P(c)}{P(d)}$$

Equation 2.3: Probability of Naïve bayes

Since P (d) is constant for all classes, we only need to calculate P (d | c) * P (c)

$$P(c \mid d) = P(d \mid c)\, P(c)$$

Equation 2.4: Probability of Naïve bayes when P (d) is constant

If we have two classes, c1 and c2, and want to compute if document d belongs to c1 or c2, let us calculate P (c1 |d) and P (c2 |d). When we compare the two results, the higher result means document d belongs to it.

The NB classifier is fast, simple, and computationally efficient; it provides good classification performance. It can be used for both binary and multiclass classification problems. However, it requires a very large number of records to obtain good results.

2.9.2.3 Maximum Entropy

The maximum entropy algorithm estimates probabilities based on the principle of making as few assumptions as possible, other than the constraints imposed. Such constraints are derived from training data, expressing some relationship between features and outcome (Kobayashi, Inui and Matsumoto, 2007).

. In some studies, ME worked better than Naïve Bayes and Nearest Neighbor classification for their classification. Unlike the Naïve Bayes machine learning, Maximum Entropy makes no independence assumptions about the occurrence of words. The Maximum Entropy modeling technique provides a probability distribution that is as close to the uniform as possible given that the distribution satisfies certain

constraints. We provide only a terse overview of Maximum entropy. It requires a set of features, which define a category. For example, in case of documents features could be the words that belong to the documents in that category (Slamet *et al.*, 2018).

2.9.2.4 K-Nearest Neighbor

K-Nearest Neighbor (KNN) is an instant-based learning algorithm that categorized objects based on closest feature space in the training set. The training data is mapped into multi-dimensional feature space. The feature space is partitioned into regions based on the category of the training set. A point in the feature space is assigned to a particular category if it is the most frequent category among the k nearest training data. During the classifying stage, KNN classification approach finds the k closest labeled training samples for an unlabeled input sample and assigns the input sample to the category that appears most frequently within the k subset. As KNN outperforms the other classification approaches by its simplicity, it only requires a small training set with small number of training samples, an integer which specifies the variable of k and a metric to measure closeness (Rushdi-Saleh *et al.*, 2011).

KNN algorithm is a mature theoretical tool and is easily implemented. It is often used to solve nonlinear problems, such as credit ratings and bank customer rankings, in which the collected data do not always follow the theoretical linear assumption, thus it should be one of the first choices when there is little or no prior knowledge about the distribution data. In addition, it can successfully reduce the influences of the variables on the experimental processes. It has higher forecasting accuracy and has no assumptions for the collected data, and particularly, it is not sensitive to the outliers. It has been widely applied in real-world problems, such as analyzing the structure of the stock market, fault

detection and diagnosis for photovoltaic systems, and social images recognition in social networks. In addition, several improved KNN algorithms have also been explored (Fan *et al.*, 2019). Figure 2.10 below show an example of KNN classification used for the automatic classification or categorization of text documents.

KNN classifier is based on the measure of Euclidean distance or measure of similarity between documents and k training data. This Classifier emphasizes on the measure of similarity for identifying neighbors of particular document. KNN is easy to implement, it is effective and non-parametric. The drawback of KNN is its long time taken for classification. KNN method has widely been used in the applications of data mining and machine learning due to its simple implementation and distinguished performance. However, setting all test data with the same k value in the previous KNN methods has been proven to make these methods impractical in real applications (Slamet *et al.*, 2018).

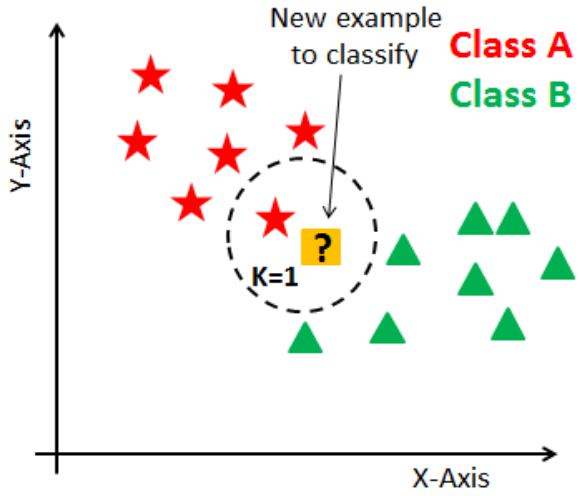


Figure 2.11: Example of KNN Classification (Fisher *et al.*, 2012)

2.9.2.5 Decision Tree

Decision trees can be adapted to almost any type of data, therefore it is one of the most widely used in machine learning algorithms. They are a supervised machine learning

algorithm that divides its training data into smaller and smaller parts in order to identify patterns that can be used for classification. The data is then presented in the form of logical structure similar to as Figure 2.11 that can be easily understood without any statistical knowledge. The algorithm is particularly well suited to cases where many hierarchical categorical distinctions can be made.

Decision tree was built using a heuristic called recursive partitioning. This is generally known as the divide and conquer approach because it uses feature values to split the data into smaller and smaller subsets of similar classes. The structure of a decision tree consists of a root node which represents the entire dataset, decision nodes which perform the computation and leaf nodes which produce the classification. In the training phase the algorithm learns what decisions have to be made in order to split the labelled training data into its classes.

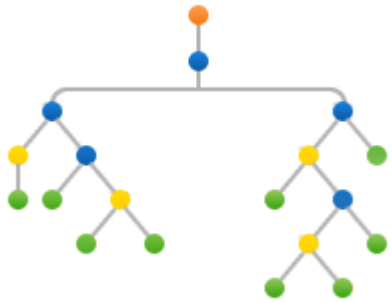


Figure 2.12: Decision Tree Structure (Ryding, 2005)

2.10 Models of text data vector representation

A vector representation of text data (word representation) lies at the core of machine learning methods, assigning to each word of the text collection a mathematical object, is often a vector of real numbers (Aisopos, Papadakis and Varvarigou, 2011). Approaches to represent the text as vectors are tested and compared by researchers to

identify the capabilities of different models to solve specific problems related to the text processing. All instances from the text collection (the training set and the test set) are n-dimensional feature vectors. The choice of features directly affects the quality of the trained model and thus the classifier performance.

2.10.1 Bag-of-Words

A simple and popular approach for representing texts is to assume that word order does not matter. A document d_i was interpreted as a set of its words $w \in d_i$ and ignore the order in which they occurred. This approach is called as the bag-of-words model, since the process can consider as taking all words from the text and throwing them in a bag, losing sequence information in the process. The binary bag-of-words model can obtain through the following feature function.

$$f_i(X) = \begin{cases} 1 & \text{if } d_i \text{ contains word } w_i, \\ 0 & \text{else.} \end{cases}$$

Equation 2.5: The binary bag-of-words Model

The bag-of-words representation assumes that it is enough to use individual words as indicators. Thus, the sentence is represented as vector.

$$d_i = (w_{i1}, w_{i2}, \ldots, w_{ij}, \ldots, w_{in})$$

Equation 2.6: The bag-of-words Vector

where w_{ij} is the weight of token w_i in the sentence d_i, n is number of all tokens in the collection |D| (the Lexicon).

2.10.2 Bag-of-N-grams

In natural language processing (NLP), the n contiguous sequence of items in the text are together called a n-gram. The items can be phonemes, syllables, letters, words or base pairs according to the application. The n-grams typically are collected from a text corpus.

For n = 1, the n-gram is called "unigram"; for n = 2, the n-gram is called "bigram", for n = 3, the n-gram is called "trigram", for n > 3, we simply replace the letter n by its numerical value, such as 4-gram, 5-gram, etc. A vector of unigrams is often called the Bag-of-Words model.

Consider the sentence "Jane likes coffee and tea". That can be represented as a vector of unigram [Jane; likes; coffee; and; tea]. Besides, this sentence can be represented as a vector of bigram [[Jane likes]; [likes coffee]; [coffee and]; [and tea]].

The character n-gram (or bag of character n-grams) is n consecutive characters of text. For example, consider a word "word", the character bigrams will be as follows: [_w, wo, or, rd, d_]. The character trigrams of the same words will be: [_wo, wor, ord, rd_].

Such a vector model used in the studies (Kanaris *et al.*, 2007) , (Kanakaraj and Guddeti, 2015) and shows good results.

Methods of defining the weight of a term in Bag-of-N-grams feature vector are similar to methods of defining the weight of a term the Bag-of-Words feature vector: binary frequency, Term Frequency (TF), and Inverse Document Frequency (TF-IDF).

2.11 Cross Validation (CV)

Cross-validation (CV) is a technique that estimates the performance of a model. It splits data into the training data set and testing data and evaluates the risk of the algorithm. The training data set is used for training the algorithm, and the testing data set is used for estimating the risk of the algorithm. The training sample is independent from the testing sample, so CV avoids over fitting. The aim in CV is to ensure that every example from the original dataset has the same chance of appearing in the training and testing set. The common types of CV are K-fold cross-validation, repeated random sub-sampling validation, and Leave-one-out cross-validation (Arlot and Celisse, 2010).

In CV the data set is divided into k subsets, and the holdout method is repeated k times. Each time, one of the k subsets is used as the test set, and the other k-1 subsets are put together to form a training set. Then the average error across all k trials is computed. In DM and machine learning 10-fold cross-validation (when k = 10) is the most common. The advantage of k-fold cross validation is that all the examples in the dataset are eventually used for both training and testing.

2.12 Twitter Application Program Interface (API)

Twitter is a microblogging social networking tool that allows users to write short messages (tweets). No more than 140 characters can be in a tweet, including links, Web pages, images, and videos. Following a user in Twitter means can seeing what people write in feed. Unfollowing someone means will stop seeing the tweets of the people that following. Retweeting is sharing a tweet with followers. The hashtag (#) is used to categorize tweets into different topics. When click on a hashtag, all tweets written on that topic will appear. "API" stands for "Application Programming Interface." In Twitter,

programmers use API to make applications, websites, widgets, and other projects that interact with Twitter. Users employ http protocol to interact with Web pages. Twitter API version 1.1 is the update of the Twitter API. Changes in the new version are: JSON support only, authentication is required, improved rate limiting, and changes to the developer rules of the road. In every request to the API, authentication is required on all endpoints. Changes of the rules include that display guidelines will be display requirements, requiring pre-installed client applications to be certified by Twitter, and requiring developers to work with twitter directly if needing a large amount of user tokens (Mollett, Moran and Dunleavy, 2011).

CHAPTER THREE

RELATED WORKS

Chapter Three
Related Works

3.1 Introduction

In this section, studies related to the sentiment analysis of Arabic language are presented with focus on social networks especially in Middle East countries.

3.2 Related Studies

(Jain and Jain, 2019) disccussed about the sentimenal analysis process which is a classification problem. To study the emotions of people about alternate energy sources, they carried out sentiment analysis on Twitter data. They carried out sentiment analysis and classification task of tweets belonging to #RenewableEnergy. They applied five different machine learning algorithms for the classification of tweets into three categories without feature selection technique and with feature selection techniques. They have used CfsSubsetEvaluation and Information Gain feature selection methods to reduce the number of features from the dataset. Their result obtained through the techniques followed in thier study, shows that the accuracy of sentiment classification is better with feature selection methods. The best accuracy (92.96%) is achieved with Support Vector Machine (Using PUK Kernel) and CfsSubsetEval feature selection method.

(Al-Kabi *et al.*, 2018) This study is based on a benchmark corpora consisting of 3,015 textual Arabic opinions collected from Facebook. These collected Arabic opinions are distributed equally among three domains (Food, Sport, and Weather), to create a balanced benchmark corpus. To accomplish this study ten Arabic lexicons were constructed manually, and a new tool called Arabic Opinions Polarity Identification (AOPI) is

designed and implemented to identify the polarity of the collected Arabic opinions using the constructed lexicons. Furthermore, this study includes a comparison between the constructed tool and two free online sentiment analysis tools (SocialMention and SentiStrength) that support the Arabic language. The effect of stemming on the accuracy of these tools is tested in this study. The evaluation results using machine learning classifiers show that AOPI is more effective than the other two free online sentiment analysis tools using a stemmed dataset.

(Hedar and Doss, 2013) have used classification techniques to detect crimes and identify their nature of different classification algorithms. Their experiments evaluated different algorithms, such as SVM, DT, CNB, and KNN, in terms of accuracy and speed in the crime domain. Also, different feature extraction techniques are evaluated, including root-based stemming, light stemming, n-gram. Their experiments revealed the superiority of n-gram over other techniques. Specifically, the results indicate the superiority of SVM with tri-gram over other classifiers, with a 91.55% accuracy.

(Pang and Lee, 2004) discussed about the sentimental analysis process. They carried out sentiment analysis and classification task of tweets belonging to #Renewable Energy. They applied five different machine learning algorithms for the classification of tweets into three categories. They have carried classification without feature selection technique and with feature selection techniques. They have used CfsSubsetEvaluation and Information Gain feature selection methods to reduce the number of features from the dataset. Their results show that the accuracy of sentiment classification is better with feature selection methods. The best accuracy (92.96%) is achieved with Support Vector Machine (Using PUK Kernel) and CfsSubsetEval feature selection method.

(Mustafa, A. and Sohail, 2017) Presented sentiment analysis of tweets written in English, belonging to different telecommunication companies in Saudi Arabia. They apply different machine learning algorithms such as a k nearest neighbor algorithm, Artificial Neural Networks (ANN), Naïve Bayesian etc. They classified the tweets into positive, negative and neutral classes based on Euclidean distance as well as cosine similarity. Moreover, they also learned similarity matrices for KNN classification. fsSubsetEvaluation as well as Information Gain was used for feature selection. Their results of CfsSubsetEvaluation were better than the ones obtained with Information Gain. Moreover, in their study, KNN performed better than the other algorithms and gave 75.4%, 76.6% and 75.6% for Precision, Recall and F-measure, respectively. Furthermore, interesting trends wrt days, months etc. Was also discovered.

(Tagwa, 2016) presented some of the previous works in sentiment analysis by using two techniques: a lexicon-based technique and a Corpus-based technique. They addressed some experiments and studies that deal with sentiment analysis in Arabic. Their study aims to use sentiment classification for Arabic tweets around Khartoum. They used different techniques for Arabic sentiment analysis applied in Arabic tweets around Khartoum and decide if the sentiment is happiness (positive), sadness (negative) or neutral. They were created a corpus of Arabic tweets around Khartoum. Then build a lexicon for Arabic words. This lexicon contains a total of words divided in two groups; the words indicating happiness (positive) and sadness words (negative) with experts in language. They used two types of classification techniques, SVM and naive Bayes.

(Al-Kabi, Abdulla and Al-Ayyoub, 2013) Arabic comments were collected and analyzed from a social network (Yahoo!-Maktoob). They detailed analysis of different information such as the reviews' length, numbers of likes/dislikes, polarity distribution

and the languages used. They used dataset to test popular classifiers commonly used for Standard Arabic (SA). Table 1 below show the structure of their database used to store the collected data from Yahoo!-Maktoob.

Table 3.1: Structured of The Collected Reviews

Column Name	Type
Review-Number	Auto- Number
Topic-ID	Number
URL	Hyperlink
Topic-Title	Text
Topic	Memo
Review	Memo
Review-ID	Text
Number of Likes	Number
Number of Dislikes	Number
Polarity	Text
Gender	Text
date of Collection	Date/time
Class	Text
Sub-Class	Text
Dialect	Text

They applied two classifiers (SVM and Naïve Bayes) on these datasets and compared between them. The total number of the Arabic reviews and comments used in this study is 4625, contains the topic, comments, manual polarity, gender of the users. which leads to unbalanced classes 2812, 1230, and 583 for negative, positive and neutral classes, respectively. The result of their study illustrate that the best accuracy achieved is 68.2% using the SVM.

(El-Halees, 2011) have proposed a combined approach which aims to mining opinions from Arabic documents. They found that using only one method on Arabic opinioned documents produce a poor performance. So, they used a combined approach that consists of three methods. They collected documents related to opinions expressed in Arabic from three different domains: "education", "politics" and "sports". As depicted in table 2 below, he used total of 1143 posts contain 8793 Arabic statements with average of 7.7 statements in each post.

Table 3.2: Description of Corpus Used in The Experiment

	Number of Files		**Number of Statements**	
Domain	**Positive**	**Negative**	**Positive**	**Negative**
Education	**204**	**170**	**1166**	**990**
Politics	**205**	**200**	**182**	**2193**
Sports	**226**	**138**	**1380**	**935**
Total	**635**	**508**	**2728**	**4118**

At the beginning, a manually built lexicon is used to classify the opinions. The classified opinions are used as training set for maximum entropy method which subsequently classifies some other documents. In the final stage, k-nearest Neighbor (*KNN*) method uses the classified documents as training set and classifies the rest of the documents. Their experiments showed that in average, the accuracy moved from 50% when using only lexicon-based method to 60% when used lexicon-based method and maximum entropy together, to 80% when using the three combined methods.

(Al-Kabi *et al.*, 2014) developed an opinion mining and analysis tool for Arabic language (Standard or MSA, and colloquial). The tool accepts comments and opinions as input. And it is capable to identify the polarity, subjectivity, and strength of each comment.

They build 18 lexicons manually. Two general purpose lexicons were built to identify polarity, and 16 domain-specific lexicons were built to identify the polarity with eight different domains: Technology, Books, Education, Movies, Places, Politics, Products and Society.

In their study they used A Naive Bayes Algorithm to classify the domain of the comments. The total number of the collected Arabic reviews (either colloquial Arabic or MSA, or both) was 1080. They used Egyptian, Iraqi, Jordanian, Lebanese, Saudi, and Syrian dialects. Their experiments showed that the proposed tool yields more accurate results when it is applied on domain-based Arabic comments relative to general-based Arabic comments. As they present the tool yield 93.9% accuracy to classify the comments into their proper domains, a 90% accuracy to identify the real polarity, and a 96.9% accuracy to identify the strength of the comments with a 10% error rate. They identify some of the reasons that may show limitations in the tool by the use of spam comments, spelling mistakes, short comment length (One word) and s the polarity of some of the phrases depends mainly on the domain they were used into. For example, the Arabic word (high, " عالية") within the comment “This is a high cost product.” leads to consider the polarity of the comment as negative, while using the same Arabic word (high, " عالية") within the comment “High-quality service” leads to consider the polarity of the comment as positive. This study used a small dataset, and the proposed tool is incapable to deal with emoticons and chat language.

(Soliman *et al.*, 2014) built a sentiment analysis approach for Slang Sentimental Words and Idioms Lexicon (SSWIL) of opinions. In this study they proposed a Gaussian kernel SVM classifier for Arabic slang language to classify Arabic news’s comments on

Facebook. They collected 1846 comments from news websites like: Aljazera1, BBCarabic2, Alyoum Alsabe3 and Alarabia4 and Constitution Facebook Page.

Support Vector Machines is a classification technique that has been used in their study. In addition, they applied three type of classification. The first classification type using Classical Lexicon (SVM) without SSWIL, the second type using Classic Lexicon and SSWIL, and the third type using SSWIL only. They show that the extraction techniques fail to extract the opinion words at the first classification type but it performs well at the second type after adding the SSWIL. The first classification type (using classic lexicon) produce 75.35% accuracy rate, while the second classification type (using SSWIL with classic lexicon) produce 86.86% and applying the system using SSWIL only, it gives 43.02% as a percent of comments classification and 56.98% not classified. As we see the results are enhanced in the second type after applying SSWIL lists.

(Rushdi-Saleh *et al.*, 2011) collected a collection of Arabic reviews about movies. Then they translated the opinion corpus for Arabic (OCA) into English, and generated the EVOCA corpus, which is the English version of the OCA corpus using an automatic machine translation tool. Both corpora include a total of 500 reviews, 250 positives and 250 negatives. They used SVM and Naïve Bayes algorithms to classify the polarity of reviews. They discover that the translation process reduces the effectiveness of determining the polarity of each comment.

(Elhawary and Elfeky, 2010) presents a system for mining Arabic business reviews to identify their polarity (positive, negative or neutral). Their system helps users to provide the information needed about the local businesses, therefore provide a better search experience or the Middle East region, which mostly speaks Arabic. Also, they show the

general opinion of the Arab public about different products and services. They collect 1600 words (600 positive, 900 negatives, and 100 neutral). Their system comprises two main components: a reviews classifier that classifies any webpage whether it contains reviews or not, and a sentiment analyzer that identifies the review text itself and identifies the individual sentences that actually contain a sentiment (positive, negative, neutral or mixed). The output of this work was a lexicon that was one of the main components of the developed sentiment analyzer. One of the limitations in this study, they do not show what is the algorithm used to classify the comments.

(Elhawary and Elfeky, 2010) proposed a lexicon-based sentiment analysis tool for colloquial Arabic text used in chatting, daily conversation and within social media. They have an independent component in their work which is game-based lexicons, that are based on human expertise. Support Vector Machines, Naïve Bayes Classifiers and Maximum Entropy approaches are used in their study. However, they have proven that the method that has higher accuracy was Support Vector Machines. Their tool should rely partially texts based on human judgment to overcome the problem arise from using non-standardized colloquial Arabic text.

(R. M. Duwairi, Ahmed and Al-Rifai, 2015) introduced a framework for sentiment analysis of Arabic tweets. The core of their framework is a sentiment lexicon which consists of 2376 entries: 1777 negative entries and 600 positive entries. Its built by translating the terms from English to Arabic to determine the polarity of tweets. They recognized the overall sentiment of a tweet by calculating the summation of its respective term's weights. The dataset was collected is 4400 tweets. These tweets were manually annotated with their sentiment: positive or negative. They have done two experiments.

In the first experiment, they classified tweets using the unsupervised sentiment detection framework. But the words of the tweets were not stemmed. A second experiment on the same set of tweets was carried out stemmer to stem the tokens of the tweets before they were fed to the sentiment detection framework. The idea of the second experiment is to determine the effect of stemming on sentiment analysis. In their results, they show that stemming enhances the performance and improved the overall accuracy.

(R. M. Duwairi, Ahmed and Al-Rifai, 2015) propose a technique for identifying polarity of reviews by identifying the polarity of the adjectives that appear in them. Their algorithm first collects all of the adjectives from the review and then computes the frequency of each of them. In the next step, it predicts the polarity of each adjective using a learned classifier. Then by aggregating the polarity of the opinion's adjectives (based on their frequencies), the polarity of the opinion is identified. For determining the polarity of adjectives, they used a naïve Bayes classifier. They used 30 adjectives, 10 of them are tagged as positive, 10 as negative and 10 as neutral adjectives. They randomly select 15 adjectives for training the classifier and use the remaining as test set. Their approach increases the accuracy by 10% higher than pure machine learning techniques. Their approach has good results, but they have a very simple training set.

(Maynard, Bontcheva and Rout, 2012) connect measures of public opinion measured from polls with sentiment measured from text. They analyzed several surveys on consumer confidence and political opinion over the 2008 to 2009 period, and find they correlate to sentiment word frequencies in contemporaneous Twitter messages. They derived day-to-day sentiment scores by counting positive and negative messages. Positive and negative words are defined by the subjectivity lexicon that containing about

1,600 positive words and 1,200 negative words. They classify the messages as positive if it contains any positive word, and negative if it contains any negative word. Their results show a correlation as high as 80%. Also, they suggest more advanced NLP techniques to improve opinion estimation.

They are some other studies done in other languages. (O'Connor *et al.*, 2010) analyzed various standard methods used in supervised sentiment and supervised topic detection on social media for Colloquial Singapore English. For supervised topic detection, they created a naïve Bayes classifier that performed classification on 5000 Facebook posts. They compared the result of their classifier against open source classifiers such as Support Vector Machine (SVM), Maximum Entropy and Labeled Latent Dirichlet Allocation (LDA). For supervised sentiment analysis, they classified the polarity of 425 Facebook posts. They used a naïve Bayes classifier in their work. They gave best result of accuracy of 89% for supervised topic. But they gave 35.5% of accuracy for supervised sentiment analysis with negative polarity class achieving a high precision of 94.3%.

Another study worked with English by (Pak and Paroubek, 2010), they proposed a method for automatically collect a corpus for sentiment analysis and opinion mining purposes. They perform linguistic analysis of the collected corpus and explain discovered phenomena, then build a sentiment classifier, that is able to determine positive, negative and neutral sentiments for a document. This study focused on using Twitter, the most popular microblogging platform, for the task of sentiment analysis. They build a sentiment classifier using the multinomial Naïve Bayes and SVM classifiers. Their dataset consists 216 sentiments. 108 sentiments as positive, 75 sentiments as negative, and 33 sentiments as neutral. Their Experimental evaluations

show that the proposed method is efficient and performs better than previously proposed methods.

(Lin and He, 2009) proposed a novel probabilistic modeling framework based on Latent Dirichlet Allocation (LDA), called joint sentiment/topic model (JST), which detects sentiment and topic simultaneously from text. Unlike other machine learning approaches to sentiment classification which often require labeled corpora for classifier training, their proposed JST model is fully unsupervised. Their model has been evaluated on the movie review dataset to classify the review sentiment polarity. They used two categories of free format movie review texts, with their overall sentiment polarity labeled either positive or negative. The results of this study demonstrated that their model is able to give competitive performance in document level sentiment classification compared with the results generated by other existing supervised approaches. One of the limitations of their model as they said is that, it represents each document as a bag of words and thus ignores the word ordering. It will probably predict the sentiment of "not good movie" being positive and the sentiment of "not bad movie" being negative.

(Xu, Ding and Wang, 2007) proposed a method for classify the Chinese news and reviews. They study how to apply machine learning techniques to solve sentiment classification problems. Naive Bayes and Maximum Entropy classification was used for the news and reviews sentiment classification. Their experimental results show that the accuracy of classification can achieve about 90%. Moreover, they find that selecting the words with polarity as features, negation tagging and representing test documents as feature presence vectors, can improve the performance of sentiment classification.

(Nadali, Murad and Kadir, 2010) proposed a fuzzy logic model to perform classification

and determine the strength of opinion orientation (very weak, weak, moderate, very strong and strong). The proposed method is based on the combinations of adjective, adverb and verb of opinions around each product feature in a review sentence. The main aim of their work is to increase the accuracy of lexicon approach. As they said Fuzzy logic, unlike statistical data mining techniques, not only allows using non-numerical values also introduces the notion of linguistic variables. They expect the accuracy of classification will be increased by combining opinion words. In this work the details about which dataset used is not discussed.

(Soni and Sharaff, 2015) focused on the problem of sentiment analysis of customer's online reviews about the product. They proposed a technique for developing Hidden Markov Model based sentiment analyzer which will help in analyzing online customer reviews to know whether the comment is positive or negative. Their work is divided in two phases. First, they propose a Hidden Markov Model and second, they test and reveal the comment for analyzing consumer opinions about the products. The objective of this study is to provide a Sentiment-based result for a large number of customer reviews of a products sold online. Their experiment is implemented on MATLAB software package. The dataset in their work was collected from Amazon.com. It consists of review comments on various popular products and is in Part of Speech (POS) tagged format. Their experimental results indicate that, the proposed technique is very promising in performing its tasks, and they have achieved maximum possible Precision and Accuracy. One of the limitations of this study that, the author’s dose not discuss clearly a detail about the dataset which used in their work.

Another study in chines language done by (Liu, Yang and Chen, 2012), they proposed a

method to construct an ambiguous sentiment confined library without hard work. Followed by preprocessing and extracting features based on their ambiguous sentiment confined library. Maximum Entropy classifier is used in this study to classify Chinese review sentiment. They collected dataset from riders' car reviews on Sina auto forum, each of which shows reviewer's positive or negative attitude. They labeling 20,000 reviews, 75% linguistic data are randomly chosen for training set and 25% for testing set. The Results show that, feature selection based on ambiguous sentiment confined library can improve the performance of Chinese review sentiment classification compared to feature selection based on National Taiwan University Sentiment Dictionary (NTUSD), and reach a higher accuracy, from 76.9% to 84.3%. A lack of objective standard for sentiment knowledge because of various understandings, is a basic limitation of this work.

(Zhai *et al.*, 2009) applied topic sentiment analysis (which is text analysis method that estimates the polarity of sentiments across units of text within large text corpora) to public opinion as expressed in social media by comparing reactions to the Trayvon Martin controversy in spring 2012 by commenters on the partisan news websites the Huffington Post and Daily Caller. They predict that high-profile commentators will be more polarizing than other news personalities and topics. Text data from the Daily Caller and Huffington Post was scraped with Helium Scraper. The analysis of their study includes three basic steps which are: a topic discovery step (Latent Semantic Analysis (LSA)), a sentiment analysis step, and a correspondence analysis step. In the First step they used vector space modeling to create a matrix of 2,072 terms (rows) and 1,600 comments or documents (column), and then performed a truncated singular value decomposition. In the second step, they extracted sentiments for each topic for both samples based on sentiment lexicons of positive and negative words. They used Liu's

lexicons of 6,800 positive and negative sentiment terms.

In the last step, they used multivariate statistical descriptive technique for categorical variables to graphically display data in low-dimensional space. The Results support their previous prediction, and it show that, for the Huffington Post commenters, specific topic is close to the negative pole, where for Daily Caller commenters the same topic is closer to the positive pole.

(Ceron *et al.*, 2014) presented the development and evaluation of a semantic analysis task in Twitter that lies at the intersection of sentiment analysis and natural language processing of social media text. They gathered tweets that express sentiment about popular topics, and attracted the highest number of participating teams at SemEval in 2013 and 2014, and created a large contextual and message-level polarity corpus that consist of tweets, SMS messages and Live Journal Messages. For this purpose, they extracted named entities using a Twitter-tuned NER system. SVM classifier was used in this study to classify the training dataset.

(Itani *et al.*, 2012) presented an application of two different approaches to classify Arabic Facebook posts. The first one depends on syntactic features, using common patterns used in different Arabic dialects to express opinions. These patterns achieved high accuracy in determining the polarity of informal Arabic sentiment. Second approach is an ordinary probabilistic model, based on Naïve-Bayes classifier, that assumes the independence of features in determining the class. They create a database to help them in the classification process which contain five different sets: Negative phrases (negative opinions), Positive Phrases (positive opinions), Spam (advertisement), negative emoticons (negative feelings) and positive emoticons (positive feelings). The highest coverage and accuracy

achieved were 49.5% and 83.4 % respectively in the first approach, and 91.2% and 85% respectively in the second one, when Naïve search used to classify the posts as objective or subjective.

(Kanakaraj and Guddeti, 2015) proposed a method for analyzing the mood of the society on a particular news from Twitter posts. They decided to include natural language processing techniques (NLP) especially semantics and word sense disambiguation to increase the accuracy of classification. They used Ensemble methods" in machine learning to solve the classification problems. They combined the effect of multiple machine learning algorithms to obtain a better predictive power than its constituent algorithms by separately. Also, they analyzed the performance of Decision Tree, Random Forest, Extremely Randomized Trees and Decision Tree regression with Ada Boost Classifiers on Twitter sentiment analysis. Experiments were conducted to compare the performance of Ensemble method against other machine learning algorithms like SVM, Baseline, MaxEntropy and Naive Bayes. Common results of their study represented on Figure 3.1 below.

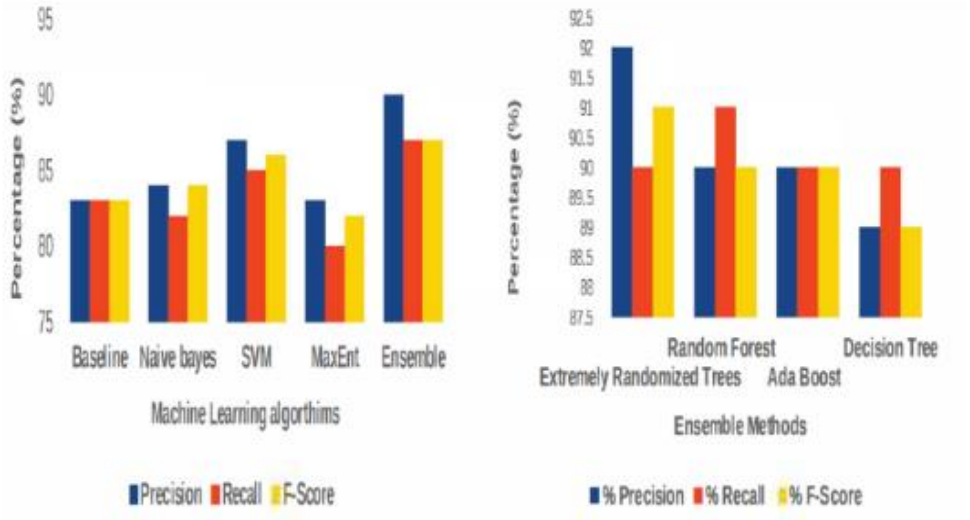


Figure 3.1: Results of the Kanakaraj and Guddeti study (Kanakaraj and Guddeti, 2015)

(Pak and Paroubek, 2010) worked on the sentiment analysis. They collected training dataset using the assumption of the emoticons contained in text represented the overall sentiment in this text. By using this assumption, a large quantity of training data was automatically collected. This study used an ensemble of two different Naive Bayes classifiers; one trained using the presence of unigrams while the second used part of speech tagging. They combined two classifiers and they achieved 74%. of accuracy.

(Pak and Paroubek, 2016) used a single Naive Bayes classifier on a movie review corpus to achieve similar results as the previous study. Multiple Naive Bayes models were trained using different features such as part of speech tagging, unigrams, and bigrams. They achieved a classification accuracy of 77.3% which was considered a high performance of the Naive Bayes classifier on that domain.

(Ritterman, Osborne and Klein, 2009) used Twitter data to ascertain public sentiment and to inform prediction model markets. Their approach also implements an SVM-based algorithm used to analyze microblog messages about a particular topic in order to forecast public sentiment. The method was applied to microblog messages about an influenza pandemic and the results were compared with prediction market data from an independent source. Their work suggests that social media data can be used as a" proxy" for public opinion.

(Java, 2008) have developed an application called BlogVox, to retrieve opinions from the blogosphere about a given topic. After pre-processing to remove spam and superfluous information, BlogVox uses an SVM to determine whether or not a blog post expresses an opinion. This differs from topic detection in that the data miner is interested in how people feel about a particular topic versus the topic itself.

Table 3.3 below shows the analysis and compression between the previous studies.

Table 3.3: Analysis of Some Related Works

Author	Techniques	Dataset	Language
(Jain and Jain, 2019)	SVM, NB and KNN	1265 text	English
(Al-Kabi *et al.*, 2018)	SVM, NB and KNN	3015 opinion	MSA
(Soliman and Ali, 2013)	SVM, DT, CNB, and KNN	-	MSA
(Al-Kabi, Abdulla and Al-Ayyoub, 2013)	SVM, NB	4625 comment	MSA
(El-halees, 2011)	KNN	8793 statements	MSA
(Mohammed *et al.*, 2014)	Naive Bayes	1080 Reviews	MSA and colloquial
(Taysir H. A. Soliman *et al.* 2014)	SVM	1846 comment	MSA
(Mohammed R. et al. 2011)	SVM Naïve Bayes	500 reviews	MSA
(Mohamed Elhawary et al., 2010)	SVM	1600 words	MSA
(Afnan A. Al-Subaihin et al. 2011)	SVM, Naïve Bayes and Maximum Entropy	-	Colloquial Arabic
(Rehab M. and Nizar A. Ahmed 2015)	SVM	4400 tweets	English
(Moghaddam and Popowich, 2002)	Naïve Bayes	30 adjectives	English

Author	Techniques	Dataset	Language
(Phua and Yee Ling, 2013)	Naïve Bayes classifier	5000	Colloquial Singapore English
(Alexander Pak and Patrick, 2010)	Naïve Bayes and SVM	216 sentiments	English

CHAPTER FOUR

METHODOLOGY

Chapter Four

Methodology

4.1 Introduction

In previous chapters a necessary background regarding this work on sentiment analysis of Arabic data from social networks was described. In this chapter our datasets, preprocessing of dataset, sentiment classification, evaluation and RapidMiner tool were discussed.

The first phase is collecting data**.** In this phase, Arabic and Sudanese dialectical words from different sources were collected**.** The second phase is Preprocessing data. Before determining the polarity of the collected tweets, preprocessing of the collected tweets are necessary to get the cleaned data. Pre-processed tweets will be applied as input of the model. This phase also includes tokenization, stopword filtering, steaming and normalization. The third phase is Sentiment classification. In this phase SVM, NB, KNN and DT were used as classification techniques to classifying tweets based on new generated lexicon into one of the two categories as positive or negative. The last phase is Evaluation**.** In this phase, the classification performance, accuracy, precision, recall and f-Measure was calculated. Figure 4.1 below show these phases.

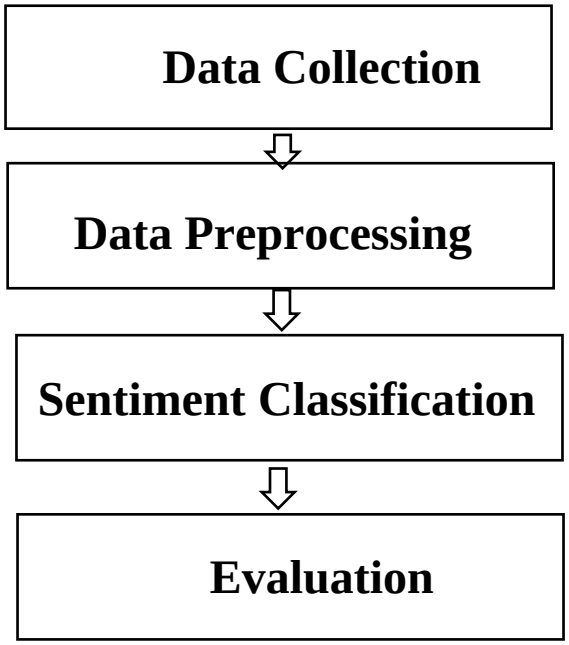


Figure 4.1: Methodology Steps

4.2 Data Collection

One of difficulties for Arabic language is the lack of publicly available Arabic lexicons in comparison with English. In general, for sentiment analysis, it is better to collect a large amount of data to be used for training the classifier. Because increasing the amount of training data in the dataset, it is always improving the accuracy of the classification.

In our work, the first step was started by building our own lexicon which contains 1854 words from Sudanese dialect. Two approaches were used to build the lexicon. In the first approach, a Sudanese word was manually collected from Twitter and different Arabic websites. In addition, a java program was implemented to collect a large number of tweets, because using the twitter API limit the amount of the collected tweets per day. In the second approach, questioner was created using google form. The questioner has seven sections as follows:

Section one: Words or sentence express the positives of a person.

Section Two: Words or sentence express the Negatives of a person.

Section Three: Words or sentence express the positives of a product.

Section Four: Words or sentence express the negatives of a product.

Section Five: Words or sentence express the positives of a movie.

Section Six: Words or sentence express the negatives of a movie.

Section Seven: Words or sentence express the positives and negatives of a food.

Section eight: general Words or sentence Sudanese dialectical Arabic.

The replies from the google form was 71 in average. Table 4.1 below show the details.

Table 4.1: Questioner details

#	Section	Data size
1	Words or sentence express the positives of a person	63
2	Words or sentence express the Negatives of a person	63
3	Words or sentence express the positives of a product	61
4	Words or sentence express the negatives of a product	61
5	Words or sentence express the positives of a movie	63
6	Words or sentence express the negatives of a movie	62
7	Words or sentence express the positives and negatives of a food	62
8	Words or sentence express the positives and negatives in general	56

Twitter's API was also used to collect Arabic tweets. The collected data has different sizes and different categories used for training and testing. Figure 4.2 below show the sources of collecting dataset.

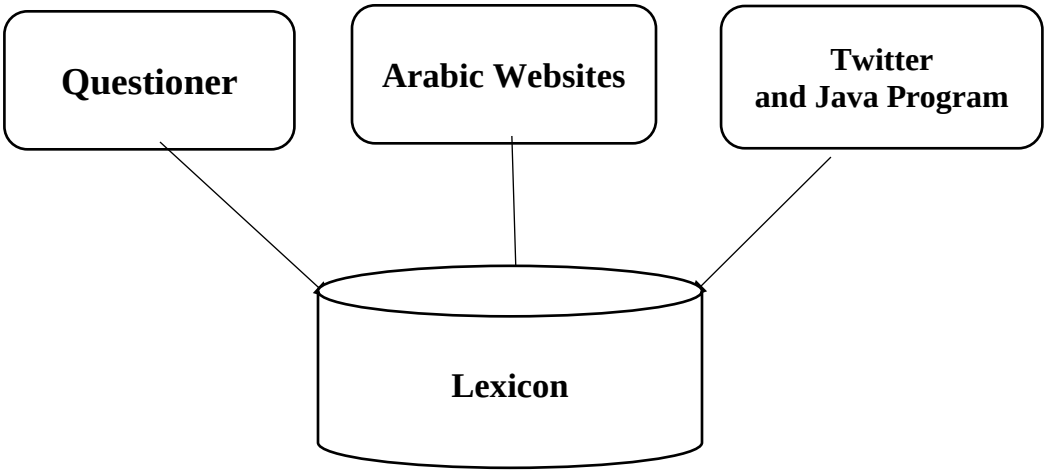


Figure 4.2: Sources of building the Lexicon of Sudanese Dialect

In this research two different dataset was collected. The first dataset is general (not specific), while the second is related the Sudanese revolution. Table 4.1 below show the details of the dataset.

Table 4.2: The collected datasets

Datasets	Number of Arabic Text
General Dataset	2500
Sudanese Revolution Dataset	6268

In the second step, the dataset was divided into training and testing dataset (see table 4.2). Then, its manually classified to positive and negative classes.

Table 4.3: Datasets Size, Training and Testing Dataset details

	General Dataset	Sudanese Revolution Dataset
Dataset Size	2500	6268
Training Dataset	1854	3882
Testing Dataset	646	2386

4.3 Data Preprocessing:

This phase includes many subphases. The process starts by data cleaning, tokenizing string to words, after the stop word removed, after that normalizing words, and finally applying stemming algorithm (stemming, light stemming). Figure 4.3 below show these steps.

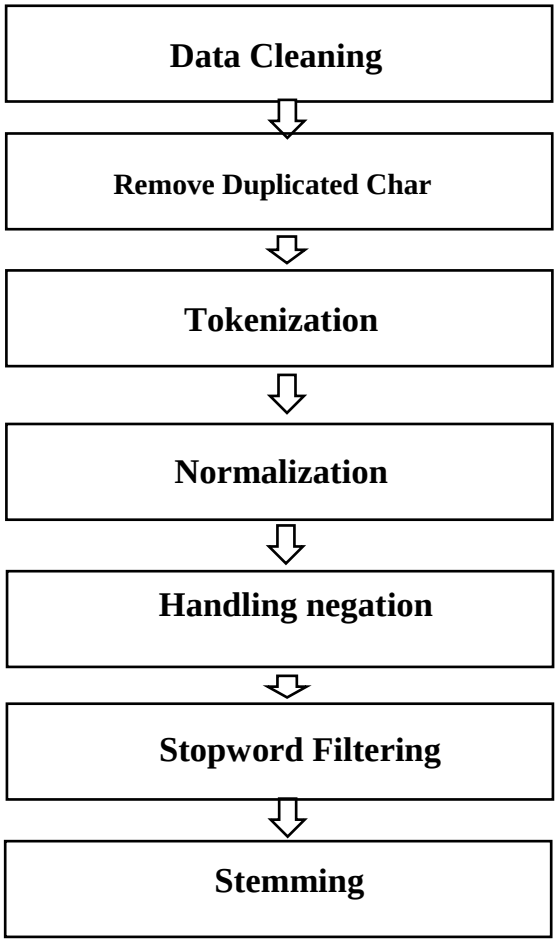


Figure 4.3: Data Preprocessing Steps

4.3.1 Data Cleaning:

This step includes removing irrelevant information, such as URLs and special characters, for example @, &.

4.3.2 Removing Duplicated Characters:

This is a common practice in tweets and other social media in Arabic, where one of the letters is repeated many times, for example "رهيييييب" which mean nice in English. We reduced any repeated characters in to two character.

Because this feature is not available at RapidMiner tool. In our work, it has been developed as exception by using java program and imported it in RapidMiner.

4.3.3 Tokenization:

Tokenization is very important in natural language processing. It can be seen as a preparation stage for all other natural language processing tasks (Aliwy, 2012). Tokenization is the task of separating out words from running text into units. These units could be characters, words, numbers, sentences or any other appropriate unit (Alotaiby, Alkharashi and Foda, 2009). The definition of a word here is not the exact syntactic form, which is why it called as “token”. In this work, the tokenization process is responsible for defining word boundaries such as white spaces from tweets.

4.3.4 Normalization:

Normalization is the process of unification of different forms of the same letter. This step includes transformed tweets into a single canonical form that it might not have had before. We eliminated the diacritical markings, non-letters, letter Hamza (ء). Also, replaced أ and إ with ا, replaced final ى with ي, and replaced final ة with ه.

4.3.5 Handling Negation:

Sentiment analysis of Arabic is still in its early stage. The most common linguistic aspect that affects sentiment analysis is negation. Negation often changes the sentiment orientation of a sentence. For example, the following two sentences, “this is a good

university” and “this is not a good university”, will have the same polarity when the negation item “not” is ignored in sentiment analysis. The positive sentiment associated with the word “good” is inverted into negative sentiment for the phrase “not good” and may not necessarily be as negative as the sentiment associated with the word “bad”. Therefore, in our work, negation items and their scope in the sentence have to be taken into account during sentiment classification. Determining negation in a sentence is not an easy task due to the compound nature of negation. Negation words such as ‘not’ and ‘no’ do more than merely demonstrate negation in the sentence, but also possess further semantic meanings. The appearance of these words does not always indicate negation, particularly in the Arabic language. The negation word scan in one instance be used to express negation and to express other meanings. In addition, the negation style can be expressed in sentence without using any of the negation words. In Arabic, negation may be expressed by using a wishing style such as "ياريت لو كان المطعم دا رخيص"which mean “wish if the price of this restaurant was cheap” in English.

In this sentence, the word ‘cheap’ can express positive polarity concerning the restaurant, due to the fact that it is cheap. However, the actual intention of the expression is the restaurant was not cheap. Hence this sentence conveys, in reality, a negative polarity. Many other works study the effect of negation in detail in the English language while few Arabic studies touch this issue because this field is still at an early stage. Most of the previous works also in Arabic sentiment analysis neither include the negation concept in sentiment analysis nor clarify the negation words list that they rely on. In addition, most of the works that include the negation theory use the semantic based approach to resolve the sentiment in Arabic text, not machine learning based approaches.

There are two styles of negation. The first style uses negation terms, called explicit negation. The second style is implicit negation that does not use negation terms or words. Instead, some of the words or forms in a sentence carry a negation meaning.

Explicit negation is a negation style that is used to negate the sentence using one of the negation words. The negation terms, tools, items, or words in the Modern Stranded Arabic are "**لا، لم، لما، لن، ما، ليس، غير**".

Table 4.4 below shows transliteration and the English meaning of these words.

Table 4.4 Arabic Negation Words

Arabic Negation Word	English Meaning
لا	No or Not
لم	Not
لما	Not
لن	Not
ما	Not
ليس	Not
غير	But

One of the challenges in this research that, the tools which available to the sentiment analysis of Arabic haven't include solution for handling negation.

In this work, a new extension was implemented using java programming language to handling negation. Then the extension was imported to the RapidMiner tool.

The scope of this work will be focused only on one type of negation, which is explicit negation.

4.3.6 Removing Stopwords:

Stopwords are frequently occurring, insignificant words that appear in texts. Words like (في ، على, من) are considered stopwords, which mean "in", "on", "from "in English. These words carry no information. In our work, Stopwords are filtered out prior to processing tweets.

4.3.7 Stemming:

This step includes removed any affixes (prefixes that added to the beginning of the word, infixes that added to the middle of the word, or/and suffixes that added to the ending of the word) from words to reduce these words to their stems or roots under the assumption that words with the same stem are semantically related. Table 4.5 below shows an example root "لعب" and a set of derivations can be obtained from this root.

Table 4.5: Some derivations of the root "لعب"

يلعب	ملعب	لاعب	ملعوب	لعبة
Play	**Playground**	**Player**	**Played**	**Game**

There are two major approaches that are followed for Arabic stemming (Stemming and light stemming). In this worked light stemming is used. Table 4.6 below show an example of tweet in a preprocessing stage.

Table 4.6: Example of preprocessing a tweet

Preprocessing Step	Tweets After Preprocessing
The original tweet	أنا ارفض صفوف العيش شدييييد دي شنو البهدلة دي**!!**
Data cleaning	أنا ارفض صفوف العيش **شدييييد** دي شنو البهدلة دي
Removing Duplicated Characters	أنا ارفض صفوف العيش شديد دي شنو البهدلة دي
Tokenization	أنا، ارفض، صفوف، العيش، شديد، دي، شنو، البهدلة، دي
Handling Negation	أنا، ارفض، صفوف، العيش، شديد، دي، شنو، البهدلة، دي

Preprocessing Step	Tweets After Preprocessing
Normalization	**انا** ارفض صفوف العيش شدييد **دي شنو** البهدلة **دي**
Stopword Removal	ارفض صفوف العيش شدييد البهدلة
Stemming	رفض صف عيش شدييد بهدل

4.3.8 Term Weight

Term weighting is one of preprocessing methods used for enhanced text document presentation as feature vector. Term weighting helps to locate important terms in a document collection for ranking purposes. There are several term weightings schemes the popular term weighting schemes are Boolean model, Term Frequency (TF), Inverse Document Frequency (IDF), and Term Frequency-Inverse Document Frequency (TF-IDF). Choosing an appropriate term weighting scheme is more important for text categorization (Qiu *et al.*, 2010).

4.3.8.1 Boolean Model

The Boolean model is the simplest retrieval model based on Boolean algebra and set theory. Boolean model indicates to absence or presence of a word with Booleans 0 or 1 respectively.

4.3.8.2 Term Frequency

Term frequency TF (*t, d*) is the number that the term *t* occurs in the document *d*.

The TF measures the importance of term i_t within the particular document d_j can be calculated by equation (Al-smairi, 2012):

$$TF_{i,j} = \frac{n_{i,j}}{\sum n_{k,j}}$$

Equation 4.1: Term frequency (TF)

Where:

n, The number of occurrences of the considered term (t_i) in the document d_j.

$\sum k\ j\ n$, Sum of number of occurrences of all terms in document d_j.

4.3.8.3 Inverse Document Frequency

The inverse document frequency (IDF) is one of the most widely used term weighting schemes for estimating the specificity of term in a document collection. It is based on the idea that if a term appears in only a few documents in the collection, then such a term is expected to be a good discriminator of these documents. The IDF weight of a term t can be calculated from document frequency using the formula (Al-smairi, 2012):

$$IDF_t = \log\left(\frac{N}{n}\right)$$

Equation 4.2: Inverse Document Frequency (IDF)

Where:

N: number of documents.

n: number of documents with word i.

The IDF of a term is low if it occurs in many documents and high if the term occurs in only a few documents.

4.3.8.4 Term Frequency-Inverse Document Frequency

Term Frequency and Inverse Document Frequency (TF-IDF), is a popular method of preprocessing documents in the information retrieval community. TF-IDF works by determining the relative frequency of words in a specific document compared to the inverse proportion of that word over the entire document corpus. Intuitively, this

calculation determines how relevant a given word is in a particular document (Al-Kabi *et al.*, 2014). The TF-IDF calculated by using the formula:

$$\text{TF-IDF} = TF_{i,j} \cdot IDF_{t}$$

$$\text{TF-IDF} = \frac{n_{i,j}}{\sum n_{k,j}} \cdot \log\left(\frac{N}{n}\right)$$

Equation 4.3: Term Frequency and Inverse Document Frequency (TF-IDF)

In this research, TF-IDF term weighting schema is applied to our dataset, because it is the most popular weighting schema and many researches such as (M. Saad, 2010) show that it gives a good result.

4.4 Sentiment Classification:

The data is divided into training and testing dataset. Training dataset used to build the classification models based on SVM, NB, KNN and DT classifiers. The data classified based on its polarity to positive and negative classes. Testing dataset is used to predict the polarity of the tweets. Table 4.7 below show sample of classification tweets.

Table 4.7: Sample of Classified Tweets.

Tweet	In English	Class
الفلم حلو حلاة	This movie is nice.	Positive
الزول دا قلبو مقطووووع	This is afraid person	Negative
البت دي فاكاها في روحها شديد	This is a conceited girl	Negative
اكل ظاااابط	This is delicious food	Positive

Tweet	In English	Class
اللوشن طلع ماسورة	This lotion is not good	Negative
البلد خربت	The country is ruined	Negative
نزفت دم يا السودان	The blood of Sudan is bleeding	Negative
سقطت ولسه بنعاني عشان نستقر	It fell and we still suffer for stability	Negative
أجمل ما في الثورة انو الشعب اتوحد في مطلبو بسقوط النظام الفاسد	The most beautiful thing in the revolution is that the peoples united in demanding the fall of the corrupt regime	Positive

4.5 Evaluation:

There are different measures that can used to measure classification accuracy. The basic measures are: accuracy, precision, recall and F-measure.

For the evaluation of classification results, well-known measures were addressed. The basic measurements are the counts of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) with respect to each class c of each instance. These depend on whether the class predicted by the classifier matches the expected prediction. Table 4.8 below show a confusion matrix which is computed by creating two categories, it is a matrix where test cases are distributed as follows:

Table 4.8: Confusion Matrix for Two Classes Positive and Negative

	Predicted Class	
Actual Class	**Pos**	**Neg**
Pos	**TP**	**FN**
Neg	**FP**	**TN**

True positive (TP): refers to positive instances that are correctly labeled.

False Negative (FN): are the positive instances that are incorrectly labeled.

False Positive (FP): are the negative instances that are incorrectly labeled.

True negative (TN): refers to negative instances that are correctly labeled.

The most basic measure is accuracy (Acc). the accuracy can be calculated by a simplified equation below.

$$Accuracy = \frac{\text{number of TP} + \text{number of } TN}{\text{number of } TP + FP + FN + TN}$$

Equation 4.4: Accuracy Measure

Accuracy is a good measure when classes are distributed uniformly in the collection. However, as class imbalances grow more pronounced, high accuracy might be attained by a classifier that has a bias towards the majority class.

Precision and recall are often used as an alternative, providing a more detailed analysis of the classifier's behavior with respect to each class c.

Precision measures the relative frequency of correctly classified examples that were predicted to belong to c as the equation below.

$$Precision = \frac{number\ of\ true\ positives}{number\ of\ true\ positives + false\ positives}$$

Equation 4.5: Precision Measure

Recall is the percentage of the total sentences for the given topic that are correctly classified. It can calculate as follows:

$$Recall = \frac{true\ positives}{true\ positives + false\ negatives}$$

Equation 4.6: Recall Measure

The harmonic mean of precision and recall is called the F-measure. It is calculated as the equation below.

$$F = 2 \times \frac{precision \times recall}{precision + recall}$$

Equation 4.7: F-score Measure

In this thesis, the four measures were calculated for every classifier, to evaluate the correctness of classifying tweets as positive or negative class.

4.6 RapidMiner Tool

In this research RapidMiner was used which is a java-based open source data mining and machine learning software. It has a graphical user interface (GUI) where the user can design his machine learning process without having to code (RapidMiner, 2019). Then all process is transformed into an XML (extensible Markup Language) file. RapidMiner includes many operators that support text mining such as Text Processing package. It includes more operators such as tokenization, stemming and filtering stop words. data

loading and transformation, data preprocessing and visualization, modeling, evaluation, and deployment. RapidMiner is written in the Java programming language. The tool can deal with the Arabic language that’s why we have chosen it.

CHAPTER FIVE

RESULTS AND DISCUSSION

Chapter Five
Results and Discussion

5.1 Introduction

This chapter describes several experimentation results with the datasets described in Chapter 4. Various machine learning techniques were used which are described in Chapter 2. The obtained results when applied SVM, NB and KNN techniques are also described in this chapter.

5.2 Results

Two datasets were used in this thesis as mentioned before (in chapter 4). Two experiment were done through the two datasets.

5.2.1 Results of Experiment 1

This experiment was done over the first dataset. SVM, NB and KNN Classifier are used to classify tweets into positive and negative classes.

For experiments implementation, RapidMiner was chosen since it has no limits for the number of instances and contains operators for text processing.

The dataset includes tweets of two classes: positive and negative. Therefore, for comparing the performance of method, Accuracy, precision, recall and F-measure were chosen as evaluation measures. The results of each classifier in our experiments is presented in this section.

Tables 5.1, 5.2 and 5.3 below show True Positive and True Negative for the SVM, NB, KNN classifiers.

Table 5.1: True Positive and True Negative for the SVM

	TP	TN
Predicted Positive	234	61
Predicted Negative	374	1185

Table 5.2: True Positive and True Negative for the Naive Bayes

	TP	TN
Predicted Positive	518	731
Predicted Negative	90	515

Table 5.3: True Positive and True Negative for the K-Nearest Neighbor

	TP	TN
Predicted Positive	180	63
Predicted Negative	428	1183

True positive rate (also called the sensitivity of a test) is defined as the positive class. a highly true positive indicates that correctly identifies the positive sentiments as shown in table 5.1, 5.3 above which TP achieved by SVM and KNN classifiers was high. A highly TP can be useful for ruling out the positive sentiments if the text is negative. In table 5.2 the TP was less than TN and this is due to the give multi negative sentiment the same weight by the NB classifier.

Figure 5.1, 5.2 and 5.3 below shows the confusion metrics for the three classifiers.

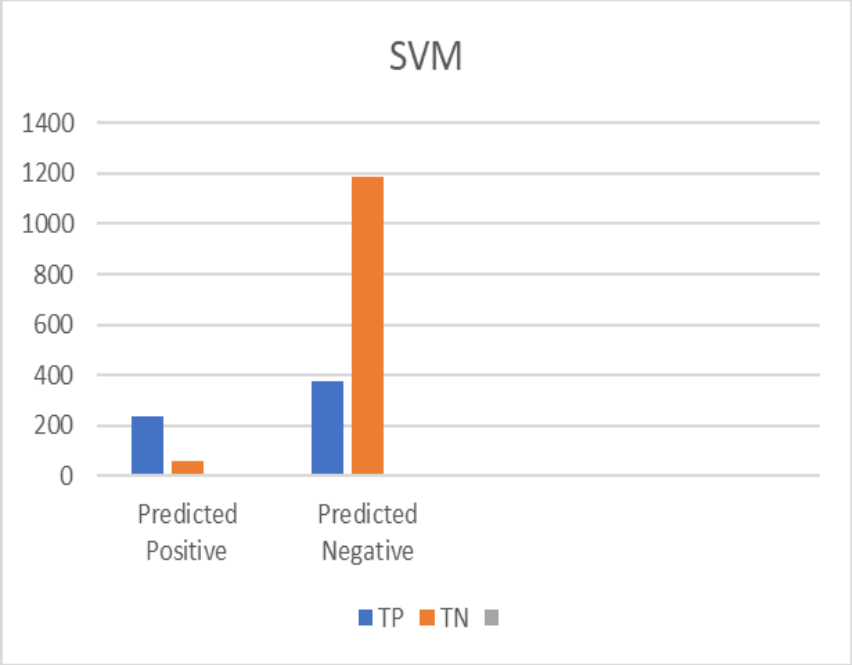


Figure 5.1: Confusion Metrics For SVM

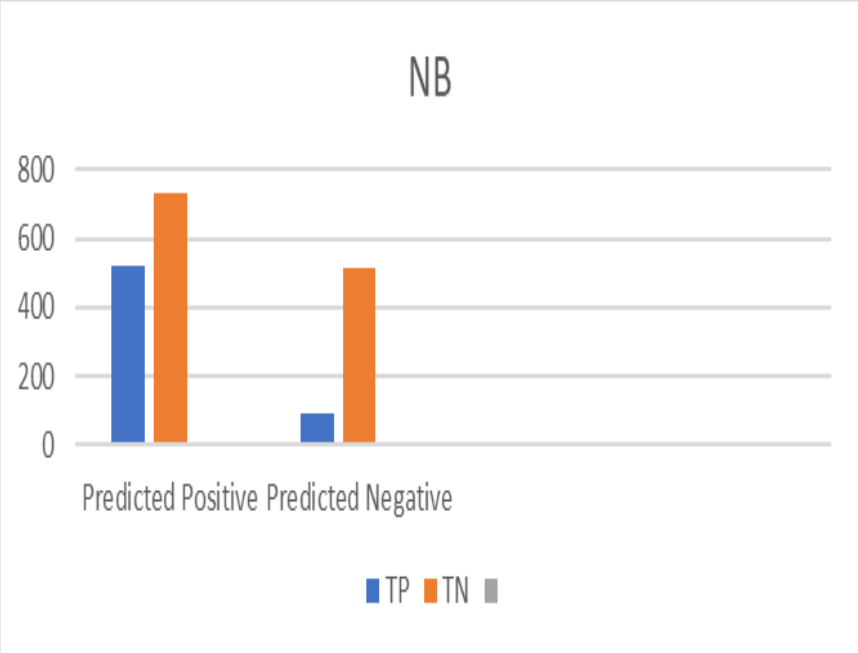


Figure 5.2: Confusion Metrics For NB

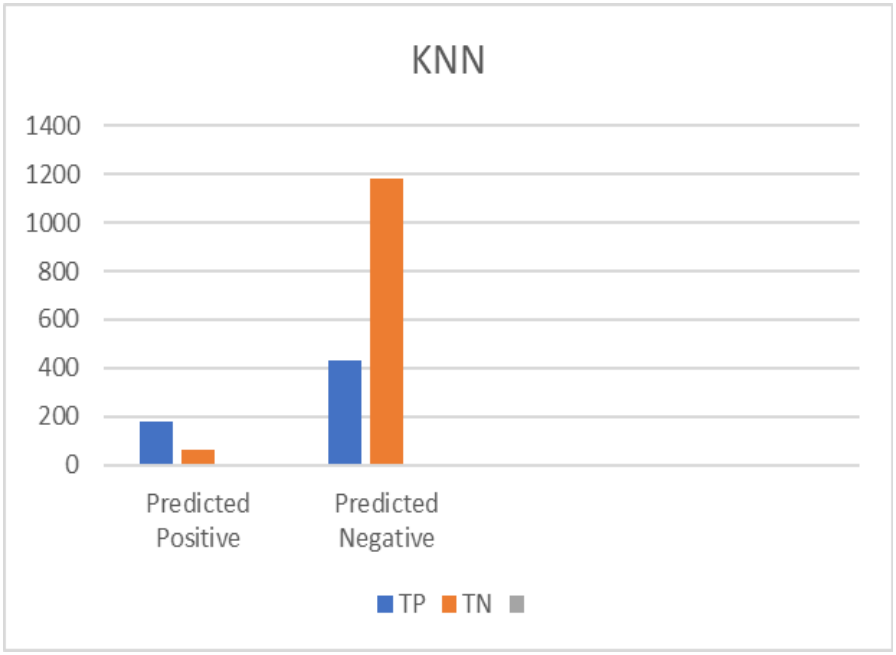


Figure 5.3: Confusion Metrics For KNN

From above figures we notice that, true negative is greater than all other parameters in SVM and KNN classifiers. That's because the negative sentiments are greater than the positive sentiments in the dataset. False negative in NB is greater than all other parameters and these was affected in Recall, while gives highest precision.

Table 5.4 shows Class Precision, Recall, Accuracy and F-Measure for the three classifiers

Table 5.4: Class Precision, Recall, Accuracy and F-Measure for the SVM, NB, KNN

	SVM	NB	KNN
Precision	**76.01%**	**85.12%**	**73.43%**
Recall	**95.10%**	**41.33%**	**94.94%**
Accuracy	**76.5%**	**55.71%**	**73.5%**
F-Measure	**84.4%**	**55.5%**	**82.8%**

From table 5.4 above, SVM classifier achieved good results for Recall, Accuracy and F-measure which equal to 95.10%, 76.5%, 48.4% respectively. Where Naive Bayes achieved good result for Precision which equal to 85.12%.

Figure 5.1 below shows a comparative between the results of the SVM, NB and K_NN classifiers.

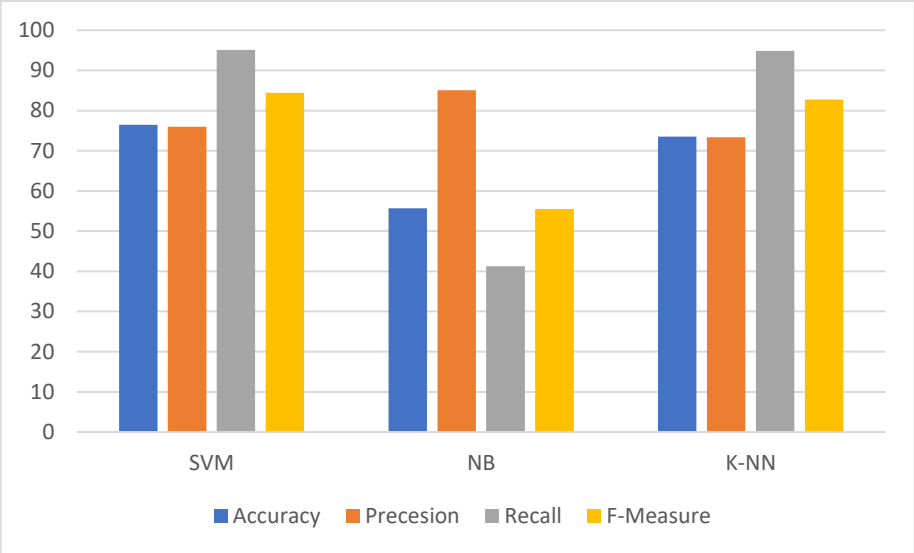


Figure 5.4: Evaluation Measures for SVM, NB, KNN

From figure 5.4 above we found that the best Accuracy, Recall and F-measure was achieved by Support Vector Machine. While the best Precision was achieved by Naïve Bayes.

5.2.2 Results of Experiment 2

This experiment was done over the second dataset which related to Sudanese revolution. This dataset was collected from twitter using Twitter API which consist of 6268 tweets with a good balance of positive and negative sentiments. Three different classifiers were used on the dataset namely; SVM, NB and Decision Tree (DT) to classify the tweets based on its polarity into positive or negative. Precision, Recall, Accuracy and F-measure were calculated for the dataset.

In this work cross-validation was performed to evaluate the classification of tweets using SVM, NB and DT classifiers with different k-folds (number of folds).

Table 5.5 below, shows the Precision, Recall, Accuracy and F-measure results of the SVM classifiers with different K-folds experiments.

Table 5.5: Precision, Recall, Accuracy and F-measure of SVM with K-folds Cross-validation

No of folds	Precision	Recall	Accuracy	F-Measure
K=5	74.3%	96.3%	75.2%	83.9%
K=6	73.2%	95.8%	73.7%	83.0%
K=7	74.2%	96.2%	75.0%	83.8%
K=8	73.8%	95.7%	74.3%	83.3%
K=9	74.0%	95.7%	74.6%	83.5%
K=10	74.2%	95.9%	75.0%	83.8%

From table 5.5 above we notice that, the best results of Precision, Recall, Accuracy and F-measure were achieved by SVM when used 5-folds Cross-validation (K=5). Figure 5.5 below shows the comparison between the four measures.

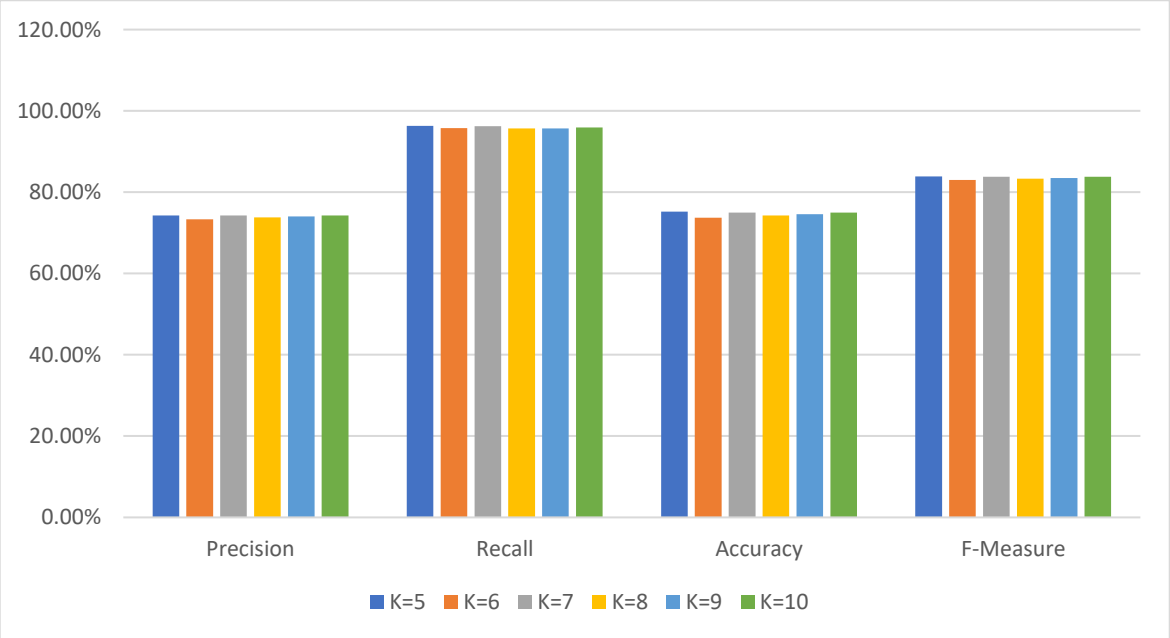


Figure 5.5: Comparison between the results of SVM based on K-folds cross validation

Table 5.6 below, shows the Precision, Recall, Accuracy and F-measure results of the NB with different K-folds experiments.

Table 5.6: Precision, Recall, Accuracy and F-measure of NB with K-folds Cross-validation

No of folds	Precision	Recall	Accuracy	F-Measure
K=5	86.7%	27.9%	48.7%	42.2%
K=6	85.7%	29.7%	49.4%	44.0%
K=7	88.4%	31.4%	51.1%	46.3%
K=8	87.2%	30.3%	50.2%	44.9%
K=9	86.0%	29.3%	49.4%	43.6%
K=10	86.9%	29.6%	49.7%	44.2%

From table 5.6 above we notice that, the best results of Precision, Recall, Accuracy and F-measure were achieved by NB when used 7-folds Cross-validation (K=7). Figure 5.6 below shows the comparison between the four measures.

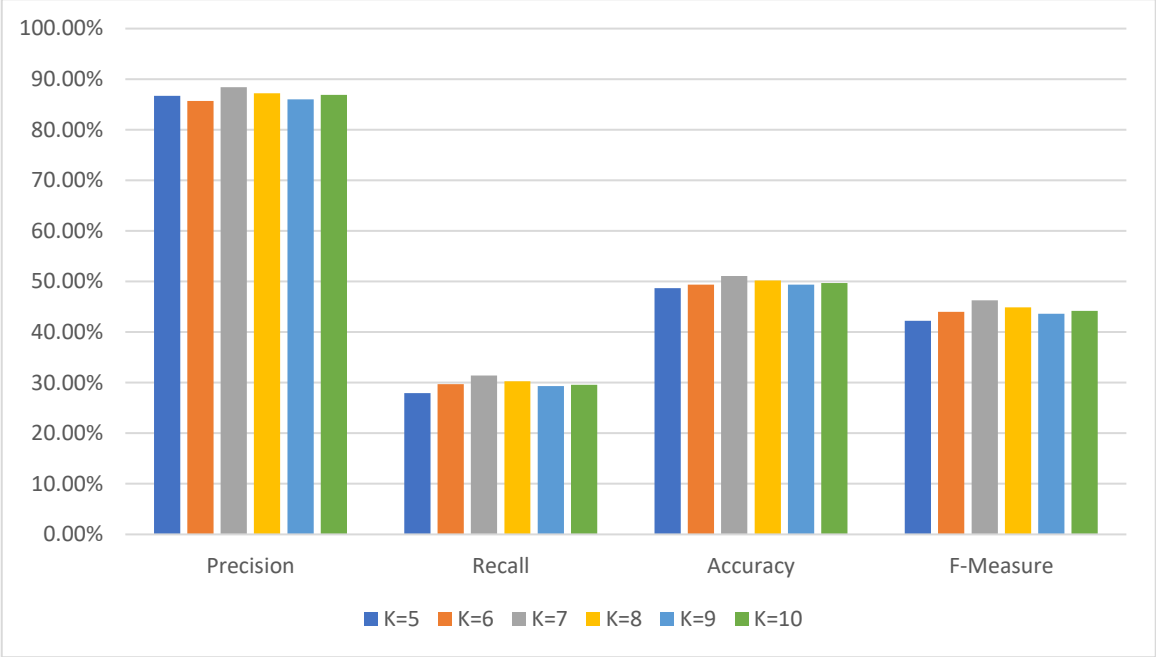


Figure 5.6: Comparison between the results of NB based on K-folds cross validation

Table 5.7 shows the Precision, Recall, Accuracy and F-measure results of the DT with different K-folds experiments.

Table 5.7: Precision, Recall, Accuracy and F-measure of DT with K-folds Cross-validation

No of folds	Precision	Recall	Accuracy	F-Measure
K=5	67.8%	99.8%	68.0%	80.7%
K=6	67.8%	99.6%	68.0%	80.7%
K=7	67.8%	99.8%	68.1%	80.8%
K=8	67.9%	99.9%	68.2%	80.9%
K=9	67.8%	99.8%	68.1%	80.8%
K=10	67.8%	99.8%	68.1%	80.8%

From table 5.7 above we notice that, the best results of Precision, Recall, Accuracy and F-measure were achieved by DT when used 8-folds Cross-validation (K=8). Figure 5.7 below shows the comparison between the four measures.

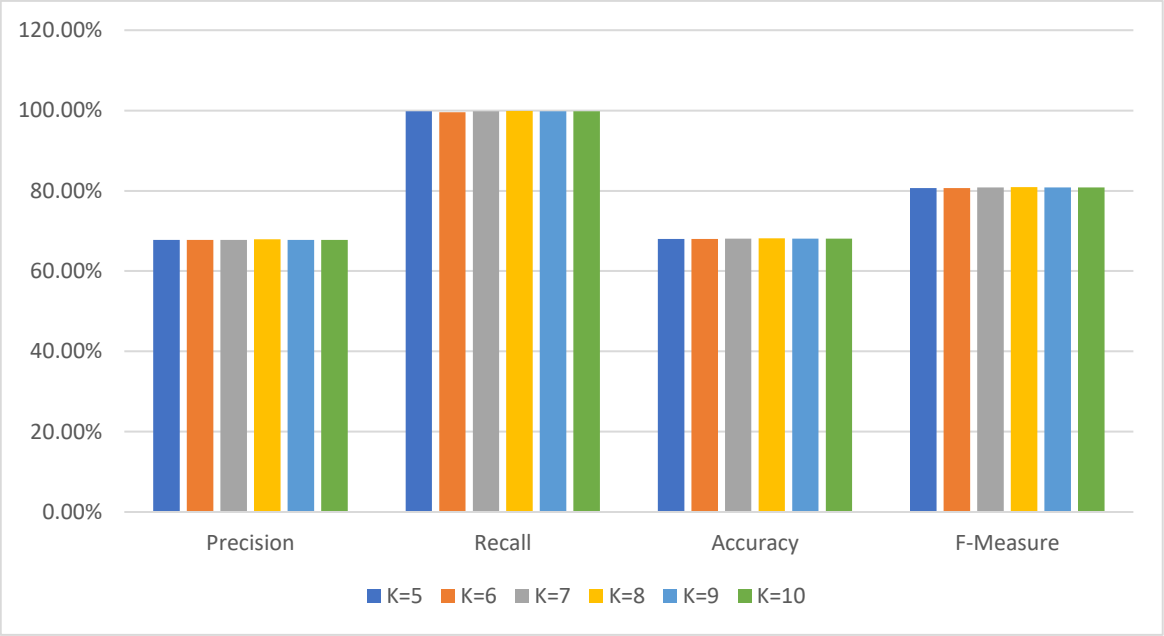


Figure 5.7: Comparison between the results of DT based on K-folds cross validation

Table 5.8 shows a comparison between SVM, NB and DT classifiers based on the best results of Precision, Recall, Accuracy and F-measures.

Table 5.8: Precision, Recall, Accuracy and F-Measure for the SVM, NB and DT classifiers

Classifier	Precision	Recall	Accuracy	F-Measure
SVM	**74.3%**	**96.3%**	**75.2%**	**83.9%**
NB	**88.4%**	**31.4%**	**51.1%**	**46.3%**
DT	**67.9%**	**99.9%**	**68.2%**	**80.9%**

From table 5.8 above we notice that, Support Vector Machine achieved good results for Accuracy and F-measure which equal to 75.2%, 83.9% respectively. While Naive Bayes achieved good results for precision which equal to 88.41%, and Decision Tree achieved good results for recall which equal to 99.9%.

Figure 5.8 below shows a composition of the results of the three classifiers in detail.

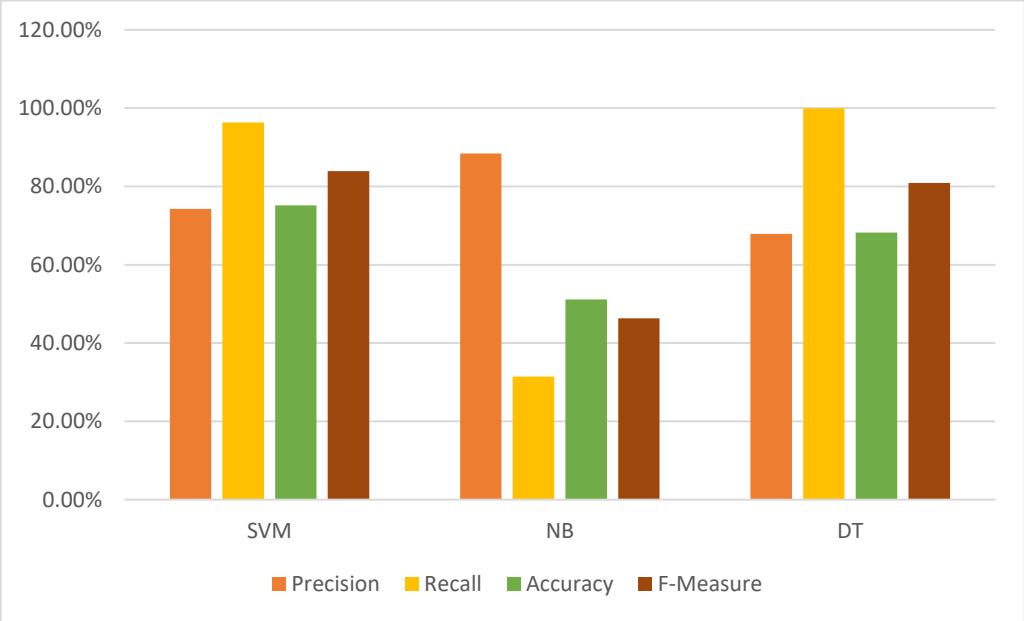


Figure 5.8: Precision, Recall, Accuracy and F-measure for the SVM, NB, DT classifiers.

From the figure above we found that the best Accuracy and F-measure was achieved by SVM. While the best Precision was achieved by NB. In addition, the best Recall was achieved by DT classifier.

The results show that, SVM achieved the best Accuracy and F-Measure and it equals 75.2%, 83.9% respectively. While Naive Bayes achieved best Precision and it equals 88.41%. Also, the best Recall was achieved by Decision Tree and it equals 99.9%.

In addition, based on these datasets which collected from Twitter, the percentages of positive and negative opinions toward the government was calculated. 9.4% represents the percentage of positive opinions related the government, while 90.6% represents the percentage of negative opinions related the same government.

To the best of our knowledge, the current work is the first to deal with detect the sentiments and classify tweets related to the Sudanese revolution. Also, the percentages of the positive and negative opinions could be very important and valuable for identifying the kind of opinions that the twitter users are sharing. It is also needed to take into account that this is not a sample of the whole Sudanese population but a subset of social networks users.

CHAPTER SIX

CONCULOSION AND FUTURE WORKS

Chapter Six
Conclusion and Future Works

6.1 Conclusion

Opinion Mining or Sentiment Analysis is a field of science used to extract the knowledge from huge amount of user's comments and topics. It has recently become one of the growing areas of research related to text mining and natural language processing. Research in sentiment analysis for the Arabic language has been very limited compared to other languages.

This thesis considered sentiment analysis in Arabic tweets which are written in MSA or Sudanese dialect. A new lexicon for Sudanese dialect was built, which consists of 2500 sentiment. To the best of our knowledge, this lexicon is the first lexicon for Sudanese dialectical Arabic. The SVM, Naïve Bayes, KNN and DT classifiers were applied to detect the polarity of the tweets. The results of the first experiment show that, SVM achieved the best Accuracy, Recall and F-measure and it equals 95.1%, 76.5% and 84.4% respectively. While Naïve Bayes achieved best Precision and it equals to 85.1%. The results of the second experiment show that, SVM achieved the best Accuracy and F-Measure and it equals 75.2%, 83.9% respectively. While Naive Bayes achieved best Precision and it equals 88.41%. Also, the best Recall was achieved by Decision Tree and it equals 99.9%. In addition, based on this dataset, the percentages of positive and negative opinions toward the government was calculated. 9.4% represents the percentage of positive opinions related the government, while 90.6% represents the percentage of negative opinions related the same government.

To the best of our knowledge, the current work is the first to deal with detect the sentiments and classify tweets related to the Sudanese revolution. Also, the percentages of the positive and negative opinions could be very important and valuable for identifying the kind of opinions that the twitter users are sharing. It is also needed to take into account that this is not a sample of the whole Sudanese population but a subset of social networks users.

6.2 Future Works

Some of the future work that could done to find more result on the topic of this thesis could be:

- A good starting point for future research may include adding a Multi-layer classification such as very positive, very negative, strong positive, strong negative, weak positive, weak negative and neutral.
- Doing more experiments for the datasets with more than three classifiers.
- Analysis of tweets that includes positive and negative opinions at the same time (complex opinions). Then, this type of opinion should not be classified as positive or negative only.
- Build a combined model by using different classifiers to enhance the accuracy more than the achieved results.
- Build a specific NLP tool for the Sudanese Dialect. Because it has unique vocabularies and structure. Therefore, it needs a special morphology tagger, parser and analyzer.
- Incorporate the effect of modifiers like exaggerate "مبالغة", intense"شديد", etc.
- Working with other data types such as images and voice.

The future work mentioned above is not comprehensive, but it gives some ideas of the possible future actions to take. Moreover, the complexity of Dialectical Arabic as a target language in sentiment analysis makes these tasks more challenging. These challenges

should encourage researchers to become involved in the project of developing ideas to solve these problems.

Bibliography

Fan, G.-F. *et al.* (2019) ‘Application of the Weighted K-Nearest Neighbor Algorithm for Short-Term Load Forecasting’, *Energies*. Multidisciplinary Digital Publishing Institute, 12(5), p. 916.

Harrat, S. *et al.* (2017) ‘Machine translation for Arabic dialects (survey*)’, Information Processing & Management.*

Neethu, M.S. and Rajasree, R. (2013) ‘Sentiment analysis in twitter using machine learning techniques’, *In 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT)* (pp. 1-5). IEEE.

Aghila, G. (2010) ‘A Survey of Na\" ive Bayes Machine Learning approach in Text Document Classification’, *arXiv preprint arXiv:1003.1795*.

Aisopos, F., Papadakis, G. and Varvarigou, T. (2011) ‘Sentiment analysis of social media content using N-Gram graphs’, in *Proceedings of the 3rd ACM SIGMM international workshop on Social media*. ACM, pp. 9–14.

Al-Harbi, S. *et al.* (2008) ‘Automatic Arabic text classification’.
Al-Kabi, M. *et al.* (2018) ‘Evaluating social context in arabic opinion mining.’, *Int. Arab J. Inf. Technol.*, 15(6), pp. 974–982.

Al-Kabi, M. N. *et al.* (2014) ‘Opinion mining and analysis for Arabic language’, *IJACSA) International Journal of Advanced Computer Science and Applications*, 5(5), pp. 181–195.

Al-Kabi, M. N., Abdulla, N. A. and Al-Ayyoub, M. (2013) ‘An analytical study of arabic sentiments: Maktoob case study’, in *8th International Conference for Internet Technology and Secured Transactions (ICITST-2013)*. IEEE, pp. 89–94.

Al-smairi, A. H. (2012) ‘Arabic Text Classification Using Learning Vector Quantization’, *Arabic Text Classification Using Learning Vector Quantization*. the islamic university.

Aliwy, A. H. (2012) ‘Tokenization as Preprocessing for Arabic Tagging System’, *International Journal of Information and Education Technology*. IACSIT Press, 2(4), p. 348.

Alotaibi, S. S. (2015) ‘Sentiment analysis in the Arabic language using machine learning’. Colorado State University. Libraries.

Alotaiby, F., Alkharashi, I. and Foda, S. (2009) ‘Processing large Arabic text corpora: Preliminary analysis and results’, in *Proceedings of the second international conference on Arabic language resources and tools*. Citeseer, pp. 78–82.

Alsaleem, S. (2011) ‘Automated Arabic Text Categorization Using SVM and NB.’, *Int. Arab J. e-Technol.*, 2(2), pp. 124–128.

Arlot, S. and Celisse, A. (2010) ‘A survey of cross-validation procedures for model selection’, *Statistics surveys*. The author, under a Creative Commons Attribution License, 4, pp. 40–79.

Ceron, A. *et al.* (2014) ‘Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens’ political preferences with an application to Italy and France’, *New media & society*. Sage Publications Sage UK: London, England, 16(2), pp. 340–358.

Duwairi, R M, Ahmed, N. A. and Al-Rifai, S. Y. (2015) ‘Detecting sentiment embedded in Arabic social media–a lexicon-based approach’, *Journal of Intelligent & Fuzzy Systems*. IOS Press, 29(1), pp. 107–117.

Duwairi, R. M., Ahmed, N. A. and Al-Rifai, S. Y. (2015) ‘Detecting sentiment embedded in Arabic social media - A lexicon-based approach’, *Journal of Intelligent and Fuzzy Systems*, 29(1), pp. 107–117. doi: 10.3233/IFS-151574.

Ehrlich, K. and Carboni, I. (2005) ‘Inside social network analysis’, *Boston College*, 13.

El-Beltagy, S. R. and Ali, A. (2013) ‘Open issues in the sentiment analysis of Arabic social media: A case study’, in *2013 9th International Conference on Innovations in Information Technology (IIT)*. IEEE, pp. 215–220.

El-halees, A. (2011) ‘Arabic Opinion Mining Using Combined’.

El-Halees, A. M. (2011) ‘Arabic opinion mining using combined classification approach’, *Arabic opinion mining using combined classification approach*. Naif Arab University for Security Sciences.

Elhawary, M. and Elfeky, M. (2010) ‘Mining Arabic business reviews’, *Proceedings - IEEE International Conference on Data Mining, ICDM*, pp. 1108–1113. doi: 10.1109/ICDMW.2010.24.

Fan, G.-F. *et al.* (2019) ‘Application of the Weighted K-Nearest Neighbor Algorithm for Short-Term Load Forecasting’, *Energies*. Multidisciplinary Digital Publishing Institute, 12(5), p. 916.

Fisher, D. *et al.* (2012) ‘Interactions with big data analytics’, *interactions*. ACM, 19(3), pp. 50–59.

Go, A., Bhayani, R. and Huang, L. (2009) ‘Twitter sentiment classification using distant supervision’, *CS224N Project Report, Stanford*, 1(12), p. 2009.

Harrat, S., Meftouh, K. and Smaili, K. (2017) ‘Machine translation for Arabic dialects (survey)’, *Information Processing & Management*. Elsevier.

Hedar, A. R. and Doss, M. (2013) ‘Mining social networks arabic slang comments’, in *IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*.

Helmy, T. and Daud, A. (2010) ‘Intelligent agent for information extraction from Arabic text without machine translation’, in *Proceedings of the 1st International Workshop on Cross-Cultural and Cross-Lingual Aspects of the Semantic Web*, p. C3LSW2010.

Itani, M. M. *et al.* (2012) ‘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)*. IEEE, pp. 192–197.

Jagdale, R. S., Shirsat, V. S. and Deshmukh, S. N. (2016) ‘Sentiment analysis of events from Twitter using open source tool’, *IJCSMC*, 5(4), pp. 475–485.

Jain, A. and Jain, V. (2019) ‘Sentiment classification of twitter data belonging to renewable energy using machine learning’, *Journal of Information and Optimization Sciences*, 40(2), pp. 521–533. doi: 10.1080/02522667.2019.1582873.

Al-Kabi, M., Alsmadi, I., Khasawneh, R.T. and Wahsheh, H., 2018. Evaluating social context in Arabic opinion mining. *Int. Arab J. Inf. Technol.*, *15*(6), pp.974-982.

Java, A. (2008) ‘Mining social media communities and content’.

Kanakaraj, M. and Guddeti, R. M. R. (2015) ‘Performance analysis of Ensemble methods on Twitter sentiment analysis using NLP techniques’, in *Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015)*. IEEE, pp. 169–170.

Kanaris, I. *et al.* (2007) ‘Words versus character n-grams for anti-spam filtering’, *International Journal on Artificial Intelligence Tools*. World Scientific, 16(06), pp. 1047–1067.

Kobayashi, N., Inui, K. and Matsumoto, Y. (2007) ‘Opinion mining from web documents: Extraction and structurization’, *Information and Media Technologies*. Information and Media Technologies Editorial Board, 2(1), pp. 326–337.

Lin, C. and He, Y. (2009) ‘Joint sentiment/topic model for sentiment analysis’, in *Proceedings of the 18th ACM conference on Information and knowledge management*. ACM, pp. 375–384.

Lin, L. *et al.* (2014) ‘Opinion mining and sentiment analysis in social networks: A retweeting structure-aware approach’, in *2014 IEEE/ACM 7th International Conference on Utility and Cloud Computing*. IEEE, pp. 890–895.

Liu, M., Yang, S. and Chen, Q. (2012) ‘Sentiment classification on Chinese reviews based on ambiguous sentiment confined library’, in *2012 IEEE 2nd International Conference on Cloud Computing and Intelligence Systems*. IEEE, pp. 1470–1473.

Liu, T.-Y. (2009) ‘Learning to rank for information retrieval’, *Foundations and Trends® in Information Retrieval*. Now Publishers, Inc., 3(3), pp. 225–331.

Lotte, F. *et al.* (2007) ‘A review of classification algorithms for EEG-based brain–computer interfaces’, *Journal of neural engineering*. IOP Publishing, 4(2), p. R1.

Maynard, D., Bontcheva, K. and Rout, D. (2012) ‘Challenges in developing opinion mining tools for social media’, *Proceedings of the@ NLP can u tag# usergeneratedcontent*, pp. 15–22.

Melville, P., Gryc, W. and Lawrence, R. D. (2009) ‘Sentiment analysis of blogs by combining lexical knowledge with text classification’, in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pp. 1275–1284.

Mohammed, N. *et al.* (2014) ‘Opinion Mining and Analysis for Arabic Language’, *International Journal of Advanced Computer Science and Applications*, 5(5), pp. 181–195. doi: 10.14569/ijacsa.2014.050528.

Mohammed, T. A. E. (2016) ‘Review of sentiment analysis for classification Arabic tweets’, *International Journal of Emerging Technology and Advanced Engineering*, 6(3), pp. 47–53.

Mollett, A., Moran, D. and Dunleavy, P. (2011) ‘Using Twitter in university research, teaching and impact activities’. LSE Public Policy Group, London School of Economics and Political Science.

Mustafa, A., A., S. and Sohail, S. (2017) ‘Sentiment Classification of Twitter Data Belonging to Saudi Arabian Telecommunication Companies’, *International Journal of Advanced Computer Science and Applications*, 8(1). doi: 10.14569/ijacsa.2017.080150.

Nadali, S., Murad, M. A. A. and Kadir, R. A. (2010) ‘Sentiment classification of customer reviews based on fuzzy logic’, in *2010 International Symposium on Information Technology*. IEEE, pp. 1037–1044.

Neethu, M. S. and Rajasree, R. (2013) ‘Sentiment analysis in twitter using machine learning techniques’, in *2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT)*. IEEE, pp. 1–5.

O’Connor, B. *et al.* (2010) ‘From tweets to polls: Linking text sentiment to public

opinion time series', in *Fourth International AAAI Conference on Weblogs and Social Media*.

Pak, A. and Paroubek, P. (2010) 'Twitter as a corpus for sentiment analysis and opinion mining.', in *LREc*, pp. 1320–1326.

Pang, B. and Lee, L. (2004) 'A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts', in *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, p. 271.

Pang, B. and Lee, L. (2008) 'Opinion mining and sentiment analysis', *Foundations and Trends® in Information Retrieval*. Now Publishers, Inc., 2(1–2), pp. 1–135.

Qiu, Z. *et al.* (2010) 'Term weighting approaches for mining significant locations from personal location logs', in *2010 10th IEEE International Conference on Computer and Information Technology*. IEEE, pp. 20–25.

Rajeswari, R. P., Juliet, K. and Aradhana, D. (2017) 'Text classification for student data set using naive bayes classifier and KNN classifier', *Int. J. Comput. Trends Technol*, 43, pp. 8–12.

RapidMiner (no date). Available at: http://rapid-i.com/.

Ritterman, J., Osborne, M. and Klein, E. (2009) 'Using prediction markets and Twitter to predict a swine flu pandemic', in *1st international workshop on mining social media*. ac. uk/miles/papers/swine09. pdf (accessed 26 August 2015), pp. 9–17.

Rushdi-Saleh, M. *et al.* (2011) 'Bilingual experiments with an arabic-english corpus for opinion mining', in *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pp. 740–745.

Ryding, K. C. (2005) *A reference grammar of modern standard Arabic*. Cambridge university press.

Saad, M. (2010) 'The Impact of Text Preprocessing and Term Weighting on Arabic Text

Classification', p. 112. Available at: http://site.iugaza.edu.ps/msaad/files/2012/05/mksaad-Arabic-text-classification-MSc-Thesis-2010-rev9.pdf.

Saad, M. K. (2010) 'The impact of text preprocessing and term weighting on arabic text classification', *The Impact of Text Preprocessing and Term Weighting on Arabic Text Classification*. the islamic university.

Saleh, M. R. *et al.* (2011) 'Experiments with SVM to classify opinions in different domains', *Expert Systems with Applications*. Elsevier, 38(12), pp. 14799–14804.

Slamet, C. *et al.* (2018) 'Web scraping and Naïve Bayes classification for job search engine', in *IOP Conference Series: Materials Science and Engineering*. IOP Publishing, p. 12038.

Soliman, T. H. *et al.* (2014) 'Sentiment analysis of Arabic slang comments on facebook', *International Journal of Computers & Technology*, 12(5), pp. 3470–3478.

Soliman, T. H. A. and Ali, M. M. (2013) 'MINING SOCIAL NETWOR KS ' ARABIC SLANG COMMENTS', 2013, pp. 22–24.

Soni, S. and Sharaff, A. (2015) 'Sentiment analysis of customer reviews based on hidden markov model', in *Proceedings of the 2015 International Conference on Advanced Research in Computer Science Engineering & Technology (ICARCSET 2015)*. ACM, p. 12.

Turney, P. (no date) 'Semantic Orientation Applied to Unsupervised Classification of Reviews', in *Proceedings of ACL-02, 40th Annual Meeting of the Association for Computational Linguistics*, pp. 417–424.

Vargas, S. *et al.* (2016) 'Comparing overall and targeted sentiments in social media during crises', in *Tenth international AAAI conference on web and social media*.

Wilson, T., Wiebe, J. and Hoffmann, P. (2009) 'Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis', *Computational linguistics*. MIT Press, 35(3), pp. 399–433.

Xu, J., Ding, Y.-X. and Wang, X.-L. (2007) 'Sentiment classification for Chinese news

using machine learning methods', *Journal of Chinese Information Processing*, 21(6), pp. 95–100.

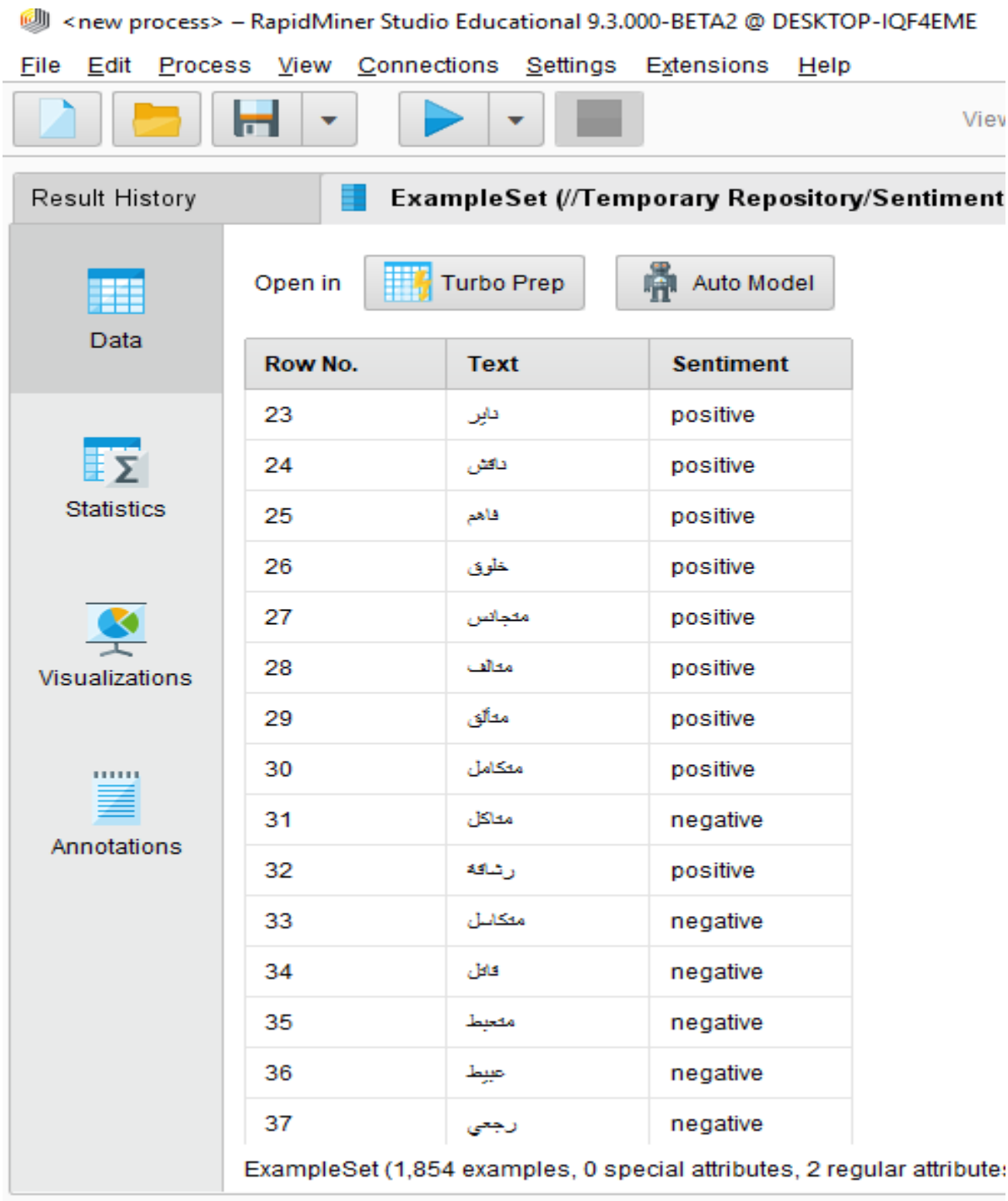
Zhai, Z. *et al.* (2009) 'Sentiment classification for Chinese reviews based on key substring features', in *2009 International Conference on Natural Language Processing and Knowledge Engineering*. IEEE, pp. 1–8.

Zhang, H. (2004) 'The optimality of naive Bayes', *AA*, 1(2), p. 3.

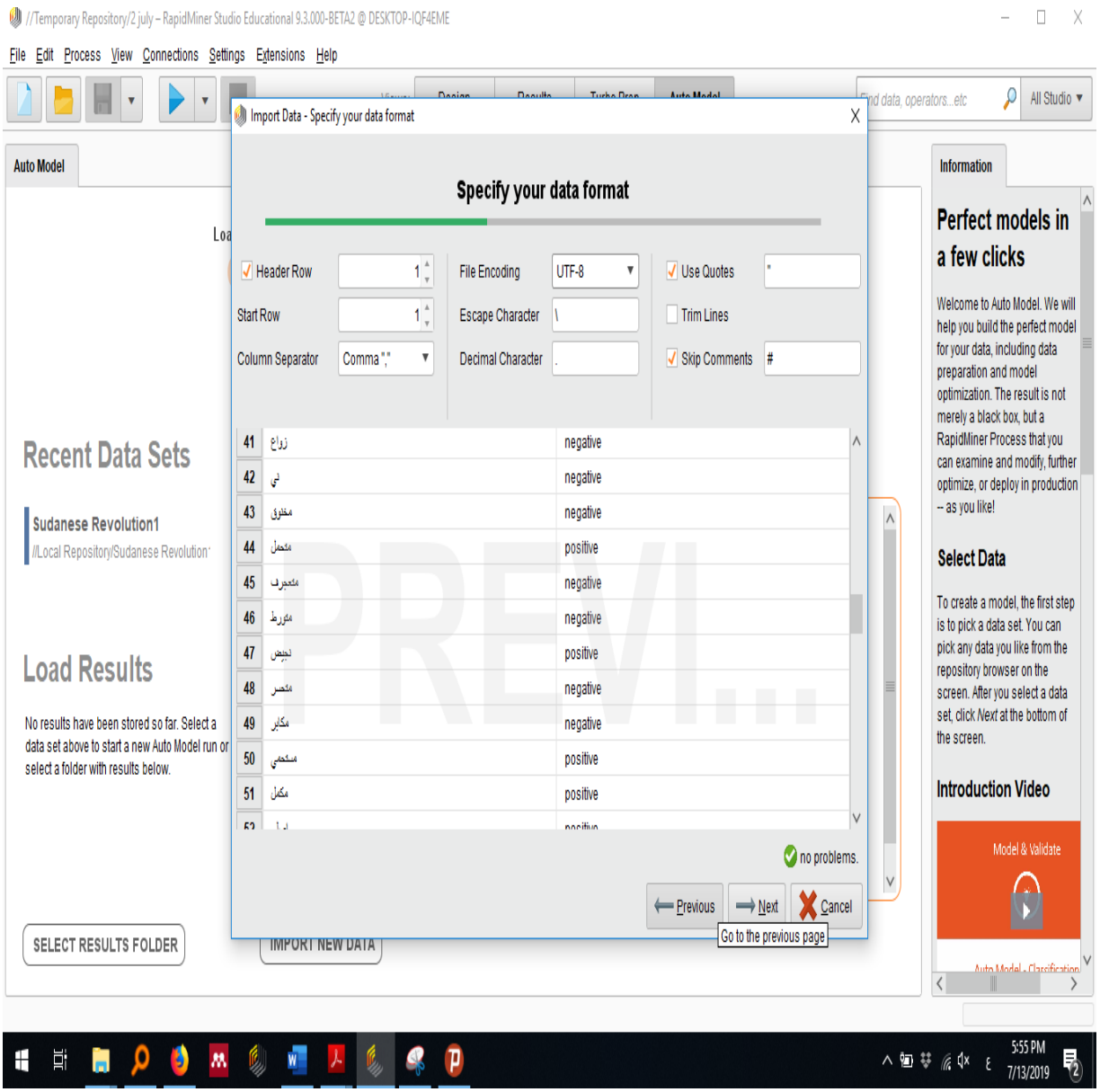
APPENDIX A

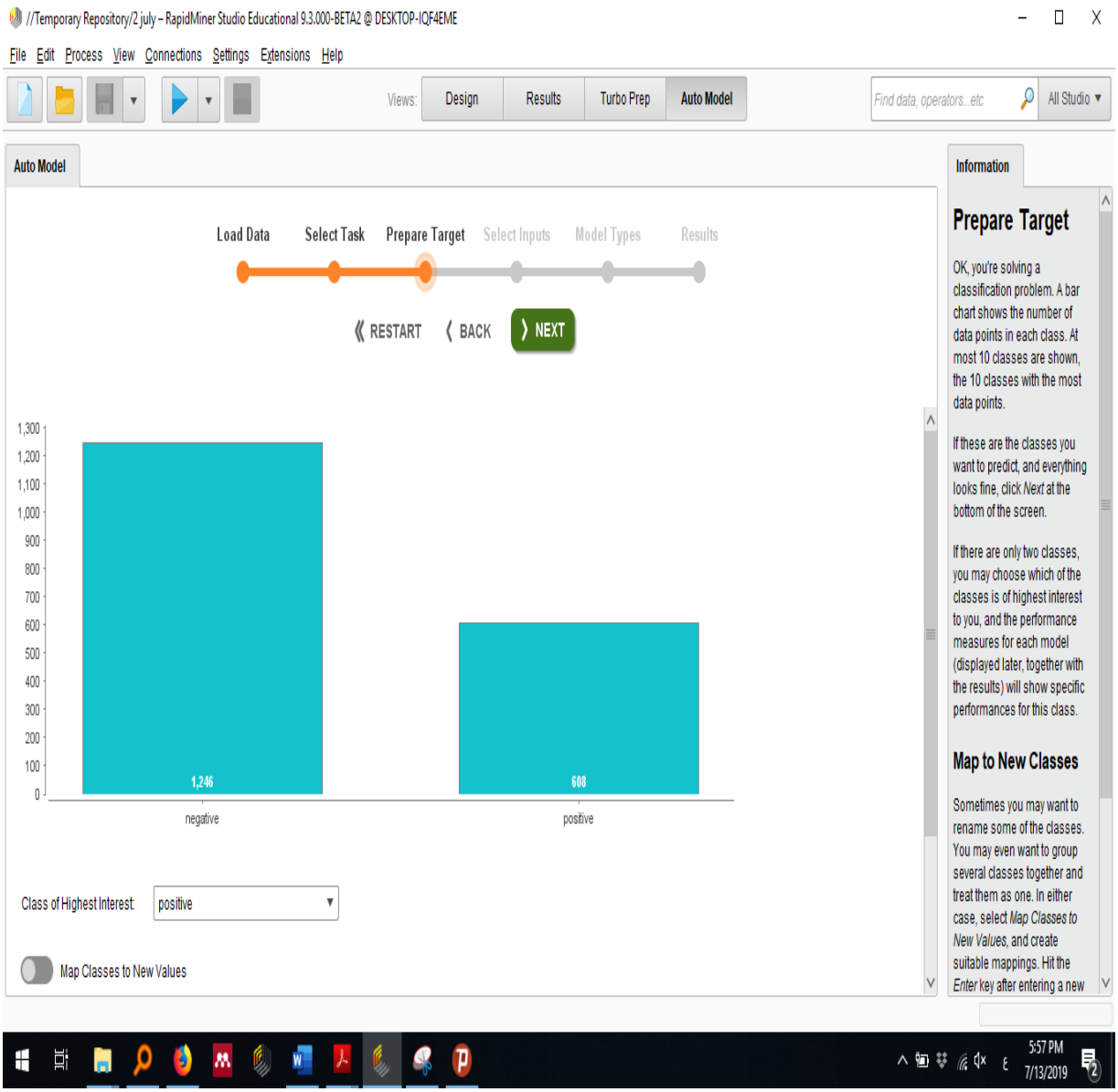
SAMPLE OF DATASET

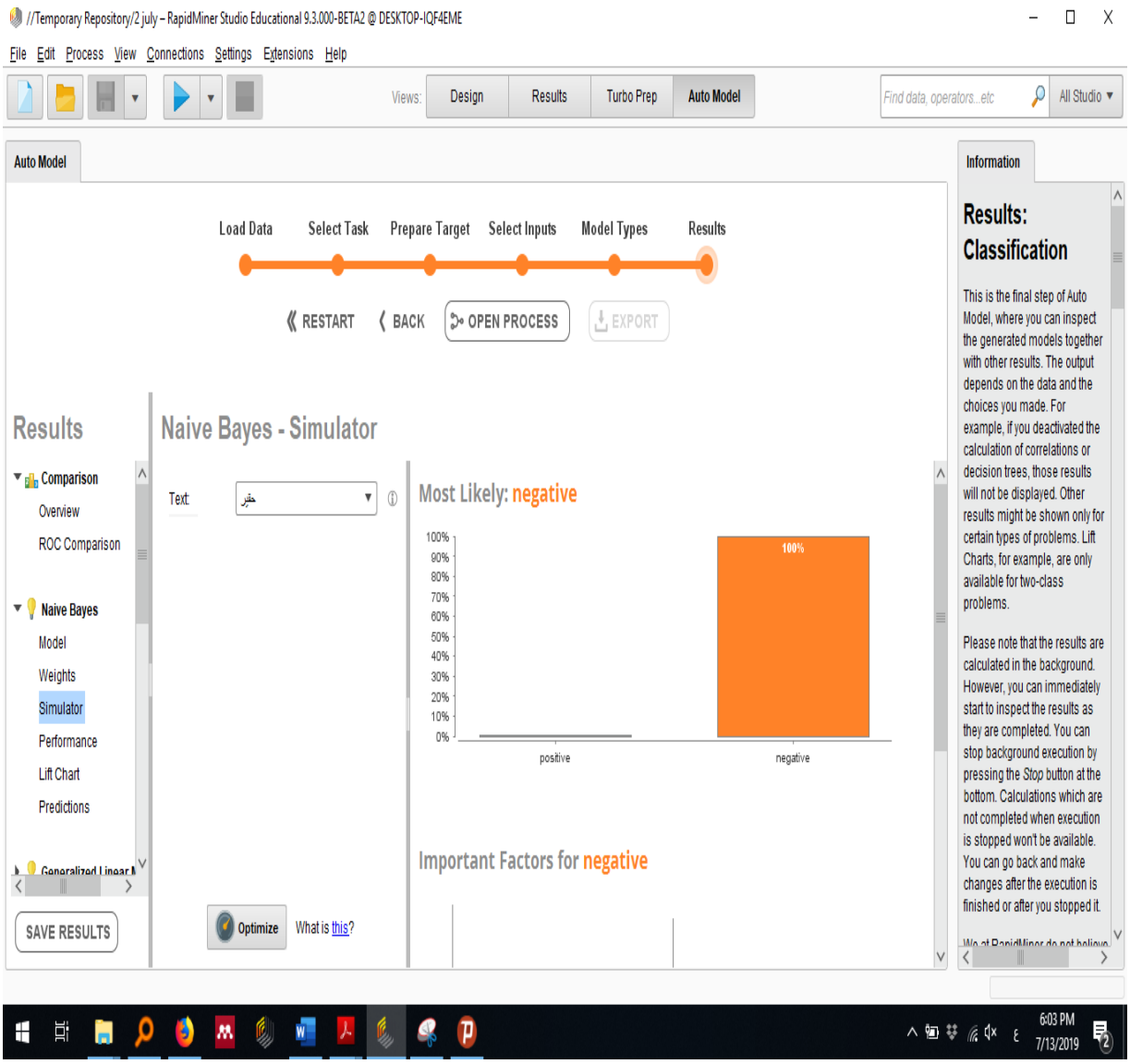
APPENDIX A: SAMPLE OF DATASET

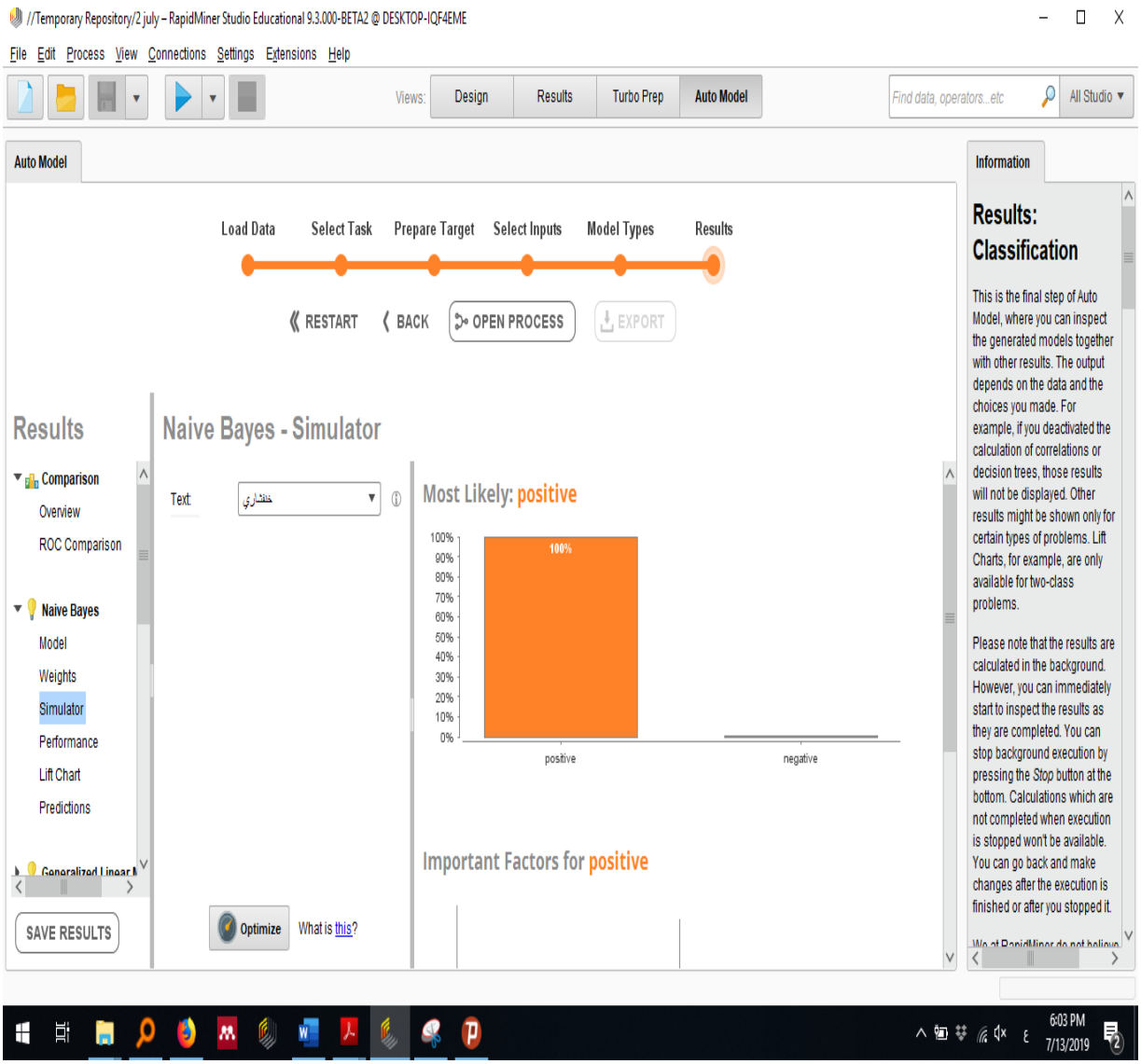


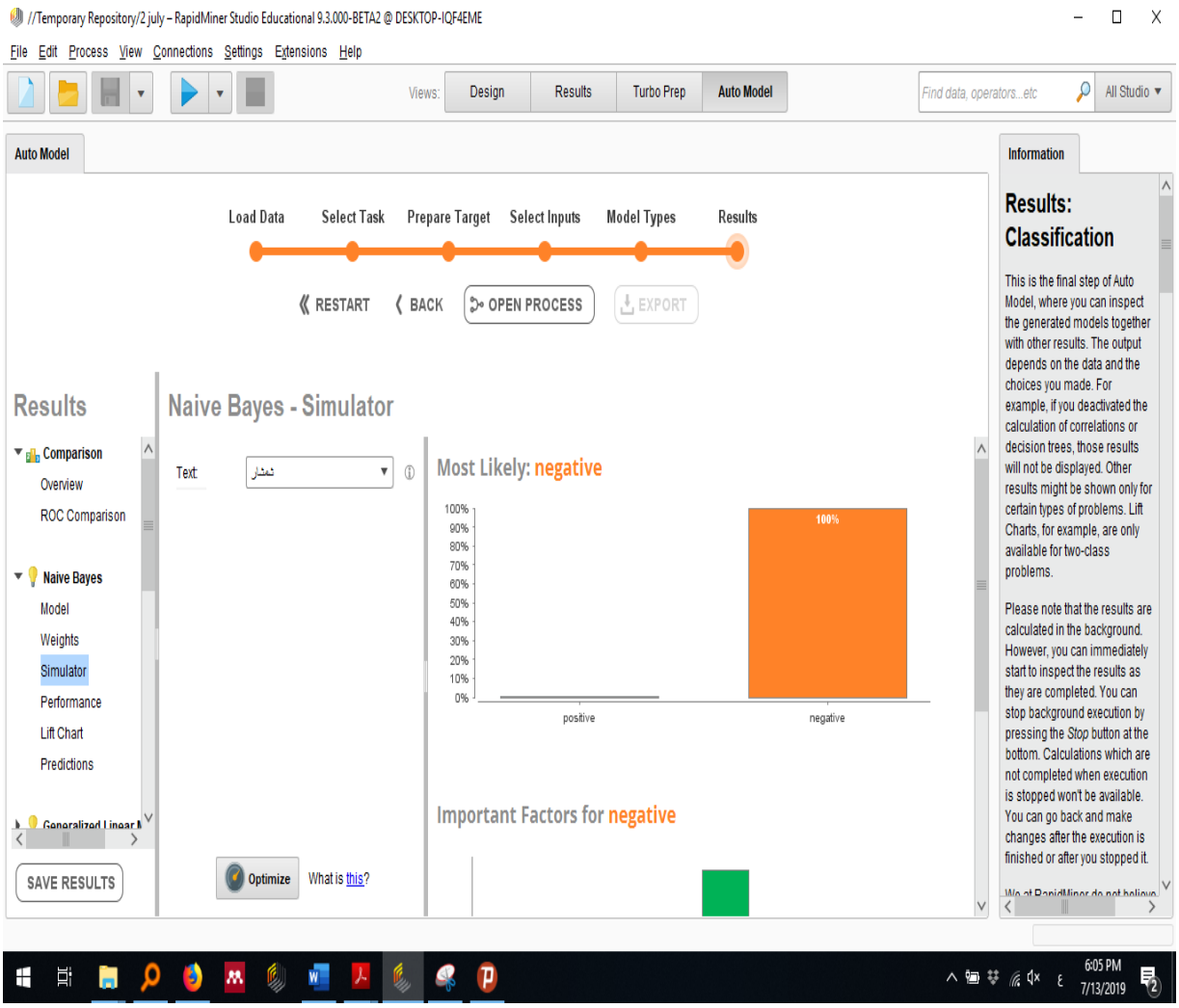
Row No.	Text	Sentiment
23	دابر	positive
24	دافش	positive
25	فاهم	positive
26	خلوق	positive
27	متجانس	positive
28	مخالف	positive
29	متألق	positive
30	متكامل	positive
31	مخاكل	negative
32	رشاقه	positive
33	متكاسل	negative
34	فاشل	negative
35	متحبط	negative
36	حبيط	negative
37	رجعي	negative


//Temporary Repository/2 july – RapidMiner Studio Educational 9.3.000-BETA2 @ DESKTOP-IQF4EME
File Edit Process View Connections Settings Extensions Help
Import Data - Specify your data format
Specify your data format
Header Row
Start Row
Column Separator
Comma ","
File Encoding
UTF-8
Escape Character
Decimal Character
Use Quotes
Trim Lines
Skip Comments
41 زواع negative
42 تي negative
43 مخوق negative
44 متحمل positive
45 متعجرف negative
46 مفرط negative
47 لحجص positive
48 متصر negative
49 مكابر negative
50 مسحمي positive
51 مكمل positive
no problems.
Previous
Next
Cancel
Go to the previous page
Auto Model
Recent Data Sets
Sudanese Revolution1
//Local Repository/Sudanese Revolution
Load Results
No results have been stored so far. Select a data set above to start a new Auto Model run or select a folder with results below.
SELECT RESULTS FOLDER
IMPORT NEW DATA
Information
Perfect models in a few clicks
Welcome to Auto Model. We will help you build the perfect model for your data, including data preparation and model optimization. The result is not merely a black box, but a RapidMiner Process that you can examine and modify, further optimize, or deploy in production -- as you like!
Select Data
To create a model, the first step is to pick a data set. You can pick any data you like from the repository browser on the screen. After you select a data set, click Next at the bottom of the screen.
Introduction Video
Model & Validate
5:55 PM
7/13/2019


//Temporary Repository/2 july – RapidMiner Studio Educational 9.3.000-BETA2 @ DESKTOP-IQF4EME
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model
Find data, operators...etc
All Studio
Auto Model
Load Data
Select Task
Prepare Target
Select Inputs
Model Types
Results
RESTART
BACK
NEXT
1,300
1,200
1,100
1,000
900
800
700
600
500
400
300
200
100
0
1,246
608
negative
positive
Class of Highest Interest:
positive
Map Classes to New Values
Information
Prepare Target
OK, you're solving a classification problem. A bar chart shows the number of data points in each class. At most 10 classes are shown, the 10 classes with the most data points.
If these are the classes you want to predict, and everything looks fine, click Next at the bottom of the screen.
If there are only two classes, you may choose which of the classes is of highest interest to you, and the performance measures for each model (displayed later, together with the results) will show specific performances for this class.
Map to New Classes
Sometimes you may want to rename some of the classes. You may even want to group several classes together and treat them as one. In either case, select Map Classes to New Values, and create suitable mappings. Hit the Enter key after entering a new
5:57 PM
7/13/2019


//Temporary Repository/2 july – RapidMiner Studio Educational 9.3.000-BETA2 @ DESKTOP-IQF4EME
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model
Find data, operators...etc
All Studio
Auto Model
Load Data
Select Task
Prepare Target
Select Inputs
Model Types
Results
RESTART
BACK
OPEN PROCESS
EXPORT
Results
Comparison
Overview
ROC Comparison
Naive Bayes
Model
Weights
Simulator
Performance
Lift Chart
Predictions
SAVE RESULTS
Naive Bayes - Simulator
Text:
Most Likely: negative
100%
positive
negative
Important Factors for negative
Optimize
What is this?
Information
Results: Classification
This is the final step of Auto Model, where you can inspect the generated models together with other results. The output depends on the data and the choices you made. For example, if you deactivated the calculation of correlations or decision trees, those results will not be displayed. Other results might be shown only for certain types of problems. Lift Charts, for example, are only available for two-class problems.
Please note that the results are calculated in the background. However, you can immediately start to inspect the results as they are completed. You can stop background execution by pressing the Stop button at the bottom. Calculations which are not completed when execution is stopped won't be available. You can go back and make changes after the execution is finished or after you stopped it.
6:03 PM
7/13/2019


//Temporary Repository/2 july – RapidMiner Studio Educational 9.3.000-BETA2 @ DESKTOP-IQF4EME
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model
Find data, operators...etc
All Studio
Auto Model
Load Data
Select Task
Prepare Target
Select Inputs
Model Types
Results
RESTART
BACK
OPEN PROCESS
EXPORT
Results
Comparison
Overview
ROC Comparison
Naive Bayes
Model
Weights
Simulator
Performance
Lift Chart
Predictions
SAVE RESULTS
Naive Bayes - Simulator
Text:
Optimize
What is this?
Most Likely: positive
100%
positive
negative
Important Factors for positive
Information
Results: Classification
This is the final step of Auto Model, where you can inspect the generated models together with other results. The output depends on the data and the choices you made. For example, if you deactivated the calculation of correlations or decision trees, those results will not be displayed. Other results might be shown only for certain types of problems. Lift Charts, for example, are only available for two-class problems.
Please note that the results are calculated in the background. However, you can immediately start to inspect the results as they are completed. You can stop background execution by pressing the Stop button at the bottom. Calculations which are not completed when execution is stopped won't be available. You can go back and make changes after the execution is finished or after you stopped it.
6:03 PM
7/13/2019


//Temporary Repository/2 july – RapidMiner Studio Educational 9.3.000-BETA2 @ DESKTOP-IQF4EME
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model
Find data, operators...etc
All Studio
Auto Model
Load Data
Select Task
Prepare Target
Select Inputs
Model Types
Results
RESTART
BACK
OPEN PROCESS
EXPORT
Results
Comparison
Overview
ROC Comparison
Naive Bayes
Model
Weights
Simulator
Performance
Lift Chart
Predictions
SAVE RESULTS
Naive Bayes - Simulator
Text:
Optimize
What is this?
Most Likely: negative
100%
positive
negative
Important Factors for negative
Information
Results: Classification
This is the final step of Auto Model, where you can inspect the generated models together with other results. The output depends on the data and the choices you made. For example, if you deactivated the calculation of correlations or decision trees, those results will not be displayed. Other results might be shown only for certain types of problems. Lift Charts, for example, are only available for two-class problems.
Please note that the results are calculated in the background. However, you can immediately start to inspect the results as they are completed. You can stop background execution by pressing the Stop button at the bottom. Calculations which are not completed when execution is stopped won't be available. You can go back and make changes after the execution is finished or after you stopped it.
6:05 PM
7/13/2019

APPENDIX B

QUSETIONERS USING GOOGLE FORM

APPENDIX B: QUESTIONER USING GOOGLE FORM

مصطلحات باللغة العامية السودانية

Form description

* كلمات او عبارات باللغة الدارجية تعبر عن ايجابيات فرد (مثلا مسالم خطير ..)

Long answer text

* كلمات او عبارات باللغة الدارجية تعبر عن سلبيات فرد (مثلا مغاط كضاب ..)

Long answer text

* كلمات او عبارات باللغة الدارجية تعبر عن إيجابيات منتج (رخيص ..)

Long answer text

* كلمات او عبارات باللغة الدارجية تعبر عن سلبيات منتج (غالي ..)

Long answer text

* كلمات او عبارات باللغة الدارجية تعبر عن إيجابيات فلم (مُمتع ..)

Long answer text

* كلمات او عبارات باللغة الدارجية تعبر عن سلبيات فلم (خيالي ..)

Long answer text

* كلمات او عبارات باللغة الدارجية توصف طعام معين(ايجابيات و سلبيات) (مثلا لذيذ، مدّود..)

Long answer text

* كلمات او عبارات اخرى باللغة الدارجية (مثلا مفطوم اللبن ما بسكتو اللولاي)

Long answer text

APPENDIX C

PUBLISHED PAPERS

APPENDIX C: PUBLISHED PAPERS

#	Paper Title	Journal	Publication Date
1	Emotion and Opinion Retrieval from Social Media in Arabic Language: Survey	*2017 Joint International Conference on Information and Communication Technologies for Education and Training and International Conference on a Computing in Arabic (ICCA-TICET)* (pp. 1-8). **IEEE**	2017
2	Sentiment Analysis of Arabic Tweets in Sudanese Dialect	*International Journal of New Technology and Research* ***(IJNTR)***	2019
3	Classification of Twitter Data Belonging to Sudanese Revolution Using Text Mining Techniques	*International Journal of New Technology and Research* ***(IJNTR)***	2019