
GAZİ ÜNİVERSİTESİ
Gazi
1926

SENTIMENT ANALYSIS IN IRAQI ARABIC DIALECTS BASED ON DISTRIBUTED REPRESENTATIONS OF SENTENCES AND MACHINE LEARNING APPROACH

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
GAZİ UNIVERSITY

BY

Anwar Adnan Mzher ALNAWAS

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF DOCTOR OF PHILOSOPHY
IN
COMPUTER ENGINEERING

MAY 2019

SENTIMENT ANALYSIS IN IRAQI ARABIC DIALECTS BASED ON DISTRIBUTED REPRESENTATIONS OF SENTENCES AND MACHINE LEARNING APPROACH

(Ph. D. Thesis)

Anwar Adnan Mzhe ALNAWAS

GAZİ UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

May 2019

ABSTRACT

Sentiment analysis is a sub-discipline of computer science involved in computational linguistics and data mining. The purpose of Sentiment analysis is the inference of individuals 'and communities' feelings and thoughts about a topic from textual documents. In the field of Sentiment analysis, which has become an interesting research topic for researchers in recent years, there are many studies on English in the scientific literature. However, not enough studies have yet been published on Arabic. Arabic; it is an important language in terms of number of speakers, history, and religious heritage. The official language in Arabic consists of classical and modern standard Arabic. Classical Arabic represents the language of the Qur'an. Modern Standard Arabic is used in newsletters and education. Although the use of Arabic on the Internet is increasing, these two types are not used in social networking environments. Local dialects used in daily practice are more preferred. Therefore Sentiment Analysis of the Arabic texts based on dialects, is an important research topic. In this doctoral dissertation, Sentiment Analysis is conducted in the Arabic-Iraqi dialect. In the first stage of the study, three types of data were collected. These are: data sets classified from previous studies, unclassified Iraqi Arabic dialect and classified Iraqi Arabic dialect. The second stage is the pre-processing stage. At this stage, unnecessary terms from the datasets have been eliminated to minimize complexity and standardize text format. In the third stage, features were extracted to represent a word as a vector using Doc2Vec model. In the fourth step, the vectors created were trained through four machine learning algorithms to create a sentiment estimation model. Lastly, the sentiment predictive model was evaluated. Moreover, at the experimental phase, the effects of variable parameters and the background corpora on classification performance was evaluated.

Science Code	:	92429
Key Words	:	Data mining, Text classification, Iraqi Arabic dialect, Sentiment analysis, NLP, Doc2Vec
Page Number	:	86
Supervisor	:	Assoc. Prof. Dr. Nursal ARICI
Co-Supervisor	:	Prof. Dr. Mehmet Hakkı SUÇIN

CÜMLELERİN DAĞITILMIŞ TEMSİLLERİ VE MAKİNE ÖĞRENMESİ YAKLAŞIMINA DAYALI IRAK LEHÇELERİNDE DUYGU ANALİZİ

(Doktora Tezi)

Anwar Adnan Mzher ALNAWAS

GAZİ ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

Mayıs 2019

ÖZET

Duygu Analizi, hesaplamalı dilbilimi ve veri madenciliği içinde yer alan bilgisayar bilimlerinin bir alt disiplindir. Duygu analizinin amacı, kişilerin veya toplulukların bir konu hakkındaki duygu ve düşüncelerinin metinsel dökümanlardan çıkarılmasıdır. Son yıllarda araştırmacılar için ilginç bir araştırma konusu haline gelen duygu analizi alanında bilimsel literatürde İngilizce için birçok çalışma bulunmaktadır. Bununla birlikte, Arapça için henüz çok fazla çalışma yayınlanmamıştır. Arapça; konuşmacıların sayısı, tarihi ve dini miras açısından önemli bir dildir. Arapçada resmi dil, klasik ve modern standart Arapçadan oluşur. Klasik Arapça, Kuran dilini temsil eder. Modern Standart Arapça, haber bültenlerinde ve eğitimde kullanır. İnternette Arapça kullanımı giderek artmakla birlikte, sosyal ağ ortamlarında bu iki tür kullanılmaz. Günlük pratik hayatta kullanılan yerel lehçeler daha çok tercih edilir. Bu nedenle, lehçelere dayalı Arapça içerikli metinlerden Duygu Analizi çalışmaları gittikçe önem kazanan araştırma konularından biridir. Bu doktora tezinde, Arap Irak lehçesinde Duygu Analizi çalışması gerçekleştirilmektedir. Çalışmanın ilk aşamasında üç tür veri kümesini toplanmıştır. Bunlar: önceki çalışmalardan sınıflandırılmış veri setleri, sınıflandırılmamış Irak Arapça lehçesi ve sınıflandırılmış Irak Arapça lehçesidir. İkinci aşama ön işleme aşamasıdır. Bu aşamada, karmaşıklığı en aza indirmek ve metin biçimini standartlaştırmak için veri kümelerinden gereksiz terimler ortadan kaldırılmıştır. Üçüncü aşamada, özelliklerin çıkarılması ve bir kelimeyi Doc2Vec modelini kullanarak vektör olarak temsil edilmesi sağlanmıştır. Dördüncü aşamada, bir duygu tahmin modeli oluşturmak için oluşturulan vektörler dört makine öğrenme algoritmasıyla eğitilmiştir. Beşinci aşamada, duygu tahmin modeli değerlendirilmiştir. Ayrıca deneysel çalışmada, değişken parametrelerin külliyat (derlem) sınıflandırma performansına etkisi de incelenmiştir.

Bilim Kodu	:	92429
Anahtar Kelimeler	:	Veri madenciliği, Metin sınıflandırma, Arapça Irak lehçesi, Duygu Analizi, DDİ, Doc2vec
Sayfa Adedi	:	86
Danışman	:	Doç. Dr. Nursal ARICI
İkinci Danışman	:	Prof. Dr. Mehmet Hakkı SUÇIN

ACKNOWLEDGEMENTS

Firstly, I am ever grateful to Allah, the Creator, and the Guardian, and to whom I owe my very existence. Foremost, I would like to offer my thanks and appreciation to my supervisor Assoc. Prof. Dr. Nursal ARICI and co-supervisor Prof. Dr. Mehmet Hakkı SUÇIN for their continued support and guidance during my PhD study. Which, without them would not have been completed. Besides my advisors, I would like to express my thanks to my thesis committee members Prof. Dr. Recep DEMİRCİ and Assist. Prof. Dr. Mustafa SERT for their advice, support, and helpful feedback. Furthermore, I would like to thank presidency for Turks abroad and related communities, Southern Technical University-Iraq, and ministry of higher education and scientific research-Iraq for sponsoring and funding my study. As well as academical personal and administrative staff of Gazi University to their help for me during my study. Best gratitude to my dear father, mother, wife, children, sisters, and brother for their unlimited support, patience, and encouragement. At last but not least, I thank all my friends and colleagues who have not forgotten me. Thank you to everyone who gave me a good word during the difficult periods of study, which reduced the stress.

TABLE OF CONTENTS

Page

Page

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

Figure **Page**

ABBREVIATIONS AND SYMBOLS

The symbols and abbreviations used in this study are presented below along with explanations.

Abbreviation	Explanation
2D	Two-Dimensional
Acc	Accuracy
AM	Average of Margins
AMT-SIMP	Amazon Mechanical Turk Simple
API	Twitter's Application Interface
AROMA	A Recursive Deep Learning Model for Opinion Mining in Arabic
ASTD	Arabic Sentiment Tweets Dataset
ATB	Arabic Tree Bank
ATT	Attraction
BOW	Bag-Of-Words
BTO	Binary-Term Occurrence
CNN	Conventional Neural Network
CSV	Comma-Separated Values
F1	F-score
FN	False Negative
FP	False Positive
GH-LG	Gold Human linguistically-motivated and genre nuanced
GH-SIMP	Gold Human Simple
HAAD	Human Annotated Arabic Dataset
HTL	Hotel
HTML	Hyper Text Markup Language
IJMES	International Journal of Middle East Studies
IAD	Iraqi Arabic Dialects
KNN	K-Nearest Neighbor Classifiers
LABR	Large-Scale Arabic Book Reviews

Abbreviation	Explanation
LG	linguistically-motivated and genre nuanced
LR	Logistic Regression
LSTM	Long short-term memory
ML	Machine Learning
MNB	Multinomial Naïve Bayes
MOV	Movie
MPQA	Multi-Perspective Question Answering
MSA	Modern Standard Arabic
NB	Naïve Bayes
NER	Name Entity Recognition
NLP	Natural Language Processing
OCA	Opinion Corpus for Arabic
P	Precision
PATB	Penn Arabic Tree Bank
PhD	A Doctor of Philosophy
POS	Part Of Speech
PROD	Product
PV-DBOW	Distributed Bag of Words of Paragraph Vector
PV-DM	Distributed Memory Model of Paragraph Vectors
QALB	Qatar Arabic Language Bank
R	Recall
RAE	Arabic using Recursive Auto Encoder
RES	Restaurant
RF	Random Forest
RMSE	Root Mean Square Error
RNTN	Recursive Neural Tensor Networks
SA	Sentiment Analysis
SIMP	Simple
SO	Semantic Orientation
SVM	Support Vector Machine
T1	Aspect Term Extraction

Abbreviation	Explanation
T2	Aspect Term Polarity
T3	Aspect Category Identification
T4	Aspect Category Polarity
TF-IDF	Term Frequency Inverse Document Frequency
TN	True Negative
TP	True Positive
U-IAD	Unlabeled- Iraqi Arabic Dialects
URL	Uniform Resource Locator
WF	Web Forum
Word2Vec	Continuous Vector Representations Of Words
WTP	Wikipedia Talk Pages
GUI	Graphical User Interface

1. INTRODUCTION

Use of social networks, blogs, and forums has enabled millions of people to share their reviews or comments on the Internet. These reviews and comments can be classified into several topics and sectors, such as goods, marketing, political and others. Many governments, companies, and other entities require analysis of these reviews to be made. While companies search for customer comments about their products, governments search for thoughts of people to make the right decision about them. In recent years, there has been an increase in the number of users of social platforms. For instance, Facebook accommodates more than 1.5 billion daily active users who share billions of different items of information, such as text, images, reactions, etc. Twitter hosts 330 million active accounts that send about 500 million Tweets everyday (Bagadiya, 2018).

The user reviews, which are available on the business pages of online platforms, are considered a good source to collect people's opinion about various topics. Therefore, when customers intend to buy a product, they seem to be interested in collecting information from comments of other people, to know about their opinions (Al-Ayyoub, Khamaiseh, Jararweh, and Al-Kabi, 2019). At this point, manually collecting comments of people and extracting them out of a huge number of comments is time-consuming and it might be difficult, especially with the rapidly growing of Web 2.0 technology. For that reason, the operative solution appears, in the last era of this problem, to be the Sentiment Analysis (SA) (García-Pablos, Cuadros, and Rigau, 2018).

SA is a field of research in Natural Language Processing (NLP) and aims to identify the polarity of text (negative or positive) (Ponomareva, 2014). SA may also be called "opinion mining", which studies people's perceptions and emotions they exhibit towards various incidents. SA is a subtask of text mining that includes tasks of processing huge numbers of comments and reviews. However, SA automatically extracts the opinion of a person he exhibits towards a matter. SA has created a new area in text analysis, which changed the concept of study from the traditional process of fact and information of text to sentiment applications. SA has attracted attentions of many people, both in business and academic fields. Rather, SA attempts to detect opinions / sentiments of people from the way they write.

Many fields are included in SA, such as NLP, Machine Learning (ML), and computational linguistics (Bhadane, Dalal, and Doshi, 2015; Soliman, Eissa, and El-Beltagy, 2017).

The Arabic language is one of the most commonly spoken languages around the world. More than 290 million people speak and write in Arabic in more than 22 countries (UNESCO, 2017). Arabic is a morphologically rich language (Abu-Errub, Odeh, Shambour, and Hassan, 2014) and has many challenges that need a special process (Alhumoud, Altuwaijri, Albuhairi, and Alohaideb, 2015). Therefore, Arabic NLP has become attractive to researchers because of its complexity and the lack of available resources (Alnawas and Arıcı, 2018a).

The importance of the Arabic language has been addressed . It can be seen that robust efforts are being made for the essential tools of Arabic NLP, for example, the morphological analyzer, part of speech tagger, and syntactical parser. According to Farghaly and Shaalan (2009), the field of Arabic NLP is still at an early phase of progress. However, studies made in Arabic SA have begun to come to existence.

With regard to the problems encountered in SA, ML approaches are widely used and typically, a supervised statistical approach is used for learning and creating a prediction model. Supervised statistical classifiers such as Support Vector Machine (SVM), Decision Tree (DT) Naïve Bayesian (NB) Classifiers, etc. are used in SA studies (Ibrahim, Abdou, and Gheith, 2015b). In addition, unsupervised approaches are also used for SA, which classifiers are applied linguistic rule-based that derived from a language and sentiment lexicon. ML algorithms require a fixed-length feature vector. A common fixed-length feature representation is bag-of-words (BOW). BOW has many limitations such as; the word order is lost and has a very little sense about the semantics. To overcome this limitation, Mikolov, Chen, Corrado, and Dean (2013a) proposed a novel model for computing continuous vector representations of words (Word2Vec). Word2Vec model proposed in two architectures: Continuous Bag-of-Words (CBOW) and Skip-gram. These architectures aim to minimize computational complexity for learning distributed representations.

Le and Mikolov (2014), proposed Distributed Representations of Documents (Doc2Vec). Doc2Vec is inspired by the recent work in distributed representations of words using neural networks (Collobert and Weston, 2008; Mnih and Hinton, 2009; Turian, Ratinov, and Bengio, 2010; Mikolov et al., 2013a).

Objectives and motivation

Based on the previous studies carried out on social media and its importance, the objectives of this study are summarized as follows: Firstly, Iraqi Arabic Dialects (IAD) was addressed to extract opinion from its text. Secondly, PV-DBOW architecture was used to extract features and represent words and sentences as vectors. Thirdly, four ML algorithms were used to train vectors to generate SA prediction models. Finally, comparison of classifiers and Doc2Vec architecture was performed.

Many reasons were conduct us to perform research in IAD sentiment analysis. The first reason is the consideration of a large number of Iraqi users navigating over the Internet. There is a significant increase in Internet users, and the number of users increased from 14 million to 19 million between 2016 and 2017 (IWS, 2017). The second reason is the importance of social media that plays a significant role in the IAD. Out of the number of Iraqi Internet users, 17 million are Facebook subscribers (StatCounter, 2017). Lastly, IAD has been representing importance according to the historical dimension and its connection to the diversity of races, religions, and cultures in Iraq (Alnawas and Arıcı, 2019).

Originality of thesis

Many SA studies are conducted in the literature, but the majority of these studies were devoted to English Language. Arabic is a morphologically rich language and has many challenges that need a special process. Therefore, the techniques that were used for the English language were not directly applicable to Arabic Language. This study is contributed to SA in Arabic in two aspects. First, IAD was addressed for SA task as the first study carried out for this dialect. Second, distributed representations of the sentences were used for IAD. Doc2Vec was modelled to represent sentences of IAD, the Doc2Vecs that have not been used in Arabic SA. Doc2Vec technique is encouraged by the current studies for English to learn about vector representations of words using neural networks. Doc2Vec is utilized to overcome the limitations of traditional techniques for representing documents since it loses the order of the word, it ignores grammatical structure and it is lexicon-dependent.

Thesis outline

The following sections explain the studies carried out within the scope of the thesis. The sections are structured as follows: Section 2. reviews the related studies carried out in Arabic SA in MSA and Arabic dialects Section 3 presents the background of the Arabic language and IAD. Section 4 presents the proposed methodology used for achieving the goals of this thesis. Section 5 illustrates the experimental study that was applied for evaluating the performance of the proposed methodology and presents results. Section 6 summarizes and discusses the main points of this thesis and suggestions for future work.

2. LITERATURE REVIEW

While the Arabic language is considered as one of the mostly used languages on the Internet, it has attracted less attention with regard to NLP studies, especially SA, compared to other languages such as English (Naili, Chaibi, and Ghezala, 2017). There is a lack in Arabic SA studies due to the morphology and nature of the Arabic language, as well as the lack of linguistic resources such as corpora and lexicons (Jarrar, Habash, Alrimawi, Akra, and Zalmout, 2017). In this section of the study, SA studies conducted in Arabic are highlighted and several papers in this field are discussed as follows.

Heikal, Torki, and El-Makky (2018) proposed a hybrid model to predict the sentiment analysis of Arabic tweets. Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) models are combined to enhance the Arabic sentiment analysis.

Alayba, Palade, England, and Iqbal (2018) studied the benefit of constructing Word2Vec model from a large Arabic corpus in order to obtain similar words. Different machine learning methods were applied and CNN was used to expand the vocabularies.

Al-Azani and El-Alfy (2017) experimented word embedding with highly imbalanced datasets containing Arabic tweets. They compared more than one SA classifiers. Al-Sallab et al. (2017) proposed a Recursive Deep Learning Model for Opinion Mining in Arabic (AROMA) to handle the morphological complexity of the Arabic language and the lack of Arabic opinion resources.

The study of Al-Azani and El-Alfy (2017) proposed Word2Vec model for Arabic SA. More than one classifier was used to compare the performance of highly imbalanced SA datasets of tweets. A dataset consisting of Syrian tweets was used as word embedding training dataset.

In the study of Al-Sallab et al. (2017), a Recursive Deep Learning Model for Opinion Mining in Arabic (AROMA) is proposed to address the restriction of morphological complexity and the lack of opinion resources for Arabic using Recursive Auto Encoder (RAE) model. The restrictions of RAE when dealing with Arabic are indicated: the Arabic morphological complexity, input features are more complete and comprehensive for the auto-encoder and

semantic composition and express the overall meaning carry out by following the natural way constituents. The other study of Baly, Hajj, Habash, Shaban, and El-Hajj (2017) evaluates the Recursive Neural Tensor Networks (RNTN) that is based on deep learning advances for Arabic sentiment analysis.

In the study of Itani, Roast, and Al-Khayatt (2017), two dialect corpora were created: news and art. 1000 posts were collected from the Facebook page of "Al-Arabiyya" news. 1000 posts were collected from "The Voice" page on Facebook. Redundancies, time stamps, and Likes are deleted in preprocessing. Four experts in the Arabic language and dialects classified the posts manually. Five classification rules were applied: Negative, Positive, Dual, Spam and Neutral. The corpus performance results ranged between 73% and 96%.

Dahou, Xiong, Zhou, Haddoud, and Duan (2016) used publicly available datasets from (Aly and Atiya, 2013; Refaee and Rieser, 2014; ElSahar and El-Beltagy, 2015). 3,4 billion words were used to build a word embedding. The performance of the created word embedding was evaluated based on Conventional Neural Network (CNN) technique.

In the study of Altowayan and Tao (2016) word embedding model was used to extract features for sentiment analysis model. The proposed model consists of three steps; compiling a large Arabic corpus, generating word vectors (embedding) and detecting subjectivity and sentiment by training several binary classifiers. Word2Vec tool from (Mikolov, Yih, and Zweig, 2013c) was used to represent words as vectors. Proposed approach accomplished a slightly better accuracy when compared to other methods in the literature that were based on hand-crafted features.

Al-Rubaiee, Qiu, and Li (2016) used Twitter as a source to create a "Mubasher". Mubasher is a corpus of product reviews collected from the Tweeter page of Mubasher company. Tweets are written in MSA and local Saudi Arabian dialects. Over a period of 57 days, 2051 tweets were collected. Two experts manually classified the tweets into three categories (positive, negative and neutral). After deleting irrelevant tweets, the corpus was remained with 1331 tweets. RapidMiner was used as a tool for preprocessing tasks. Tokenization, removing stop words, and light stem were applied. TF-IDF and BTO (Binary-Term Occurrence) feature selection schemes were used. The performance of corpus was evaluated based on NB and SVM. The best results were achieved using SVM.

Hathlian and Hafezs (2016) used 3700 tweets to create a corpus. Only 1550 tweets appeared to be related to a specific topic. User names, pictures, hashtags, URLs, emoticons, and all non-Arabic words were deleted from tweets. Spelling mistakes and standardized word-writing formulas in tweets were treated in the normalization phase. Weka suite software was used to extract features based on unigrams, bigrams, and trigrams. NB and SVM classifiers were used to test the quality of corpus. The best results were achieved for both the NB and SVM classifiers based on unigram model.

Sghaier and Zrigui (2016) collected reviews from several web pages like reviewzat1, jawal1232 and jumia3. The reviews were on five products: Cameras, notebook PCs, mobile phones, tablets, and televisions. Three of the experts classified 250 reviews manually into 125 positive and 125 negative ones. Emotions and symbols were converted to words by a small "symbol to word" program developed by authors. Stop words, special characters, non-Arabic words, and numbers were deleted in the preprocessing phase. Features were extracted based on Unigram, bigrams, trigrams. SVM, NB, and K-Nearest Neighbor (KNN) were used to evaluate the performance of corpus. The best performance was obtained using SVM and NB algorithms rather than using KNN algorithm.

In the study of Cherif, Madani, and Kissi (2015a), 625 review comments were collected from the TripAdvisor site. Comments were categorized manually into five categories: excellent, very good, middling, weak, and horrible. After removing non-relevant comments, the dataset contained 250 positive comments and 250 negative comments. Tokenization and stemming were applied. A new mathematical approach was proposed to determine the polarity of opinion. A linear program was utilized to calculate the weights of each comment and then calculated the label. The model was evaluated based on two terms: Root Mean Square Error (RMSE) and Average of Margins (AM).

In the study of Al-Smadi, Qawasmeh, Talafha, and Quwaider (2015b), a large-scale Arabic Book Reviews Dataset (LABR) that had been collected by (Aly and Atiya, 2013) was used as a source for Human Annotated Arabic Dataset (HAAD). HAAD was used as an annotated corpus for aspect-based sentiment analysis of Arabic text. The classification of comments was performed in two phases, first phase was carried out by a group of seven students who were studying in the course of Natural Language Processing at Jordan University of Science and Technology who also helped researchers classify the comments that were selected out

of the LABR. The second phase was carried out by a native Arabic speaker who held a PhD degree in computer science.

The classified comments contained information related to four tasks: Aspect Term Extraction (T1), Aspect Term Polarity (T2), Aspect Category Identification (T3) and Aspect Category Polarity (T4). At the end, the HAAD contained 1513 comments classified into positive and negative polarity aspect terms. In order to evaluate T1 and T3, F1 was calculated, where the results were 23% for T1and 15% for T3. To evaluate T2 and T4, the accuracy of the approach was measured. The accuracies of T2 and T4 were 29% and 42% respectively.

In the study of Cherif, Madani, and Kissi (2015b), 625 reviews from TripAdvisor site were classified manually into five categories: excellent, very good, middling, weak and horrible. NLP was applied to delete repeated characters in the words and delete comments that did not represent any opinion. The comment that appears in more than one category is classified in which it appears with a percentage 80%. A hybrid approach based on two of ML algorithms: SVM and KNN was proposed. SVM was used to classify comments into five groups. KNN was used in the second step to obtain satisfying results. The proposed approach showed the best results of F-measure up to %97 on average.

In the study of Shoukry and Rafea (2015), their corpus included words in Egyptian dialect. 20000 tweets were retrieved from Twitter. Tweets that contained only one opinion were considered in their study. Two of the experts classified 4800 tweets into three categories: 1600 positive, 1600 negative and 1600 neutral tweets. Images, non-Arabic words, hashtags, and URLs were deleted. Tokenization, stemming and removal of stop words were applied in the preprocessing phase. Three models of N-Grams were used to select features. The hybrid approach was proposed based on ML (NB and SVM) and Semantic Orientation (SO). In the proposed approach, SO score was added onto ML score and each sentiment word found was multiplied by the inverse of its SO weight. The results obtained by the proposed approach showed improved performance than using either ML or SO approaches.

In the study of Ibrahim, Abdou, and Gheith (2015a), 4000 reviews were collected from different resources such as tweets, product reviews, hotel reservation comments, and TV program comments. The reviews were written in MSA and Egyptian dialect. These reviews were used to build a MIKA corpus. The reviews were classified into three categories:

positive, negative and neutral. Every review was assigned a power of sentiment, 1 to 10 for the positive, and (-1 to -10) for the negative, and 0 for the natural.

In the study of AL-Smadi, Al-Ayyoub, Al-Sarhan, and Jararweh (2015a), 10000 comments were collected from Facebook. All of the comments were related to the Israel-Gaza conflict in 2014. "Breaking news from Gaza" page was used as a source for comments. The irrelevant comments were excluded. A group of three members classified 2265 comments manually into three categories (positive, negative and neutral). The group consisted of graduate students and two senior researchers. They used BRAT tool from (Stenetorp et al., 2012) for classification process. The preprocessing tasks: Tokenization, stemming, segmentation, Part Of Speech (POS), punctuation, removing stop word, and N-Gram were applied using AraNLP tool from (Althobaiti, Kruschwitz, and Poesio, 2014). The features were extracted based on Name Entity Recognition (NER) using a tool that was proposed by (Al-Rfou, Kulkarni, Perozzi, and Skiena, 2015). T1 and T2 tasks proposed in the study of (Al-Smadi et al., 2015b) were applied. The baseline results for T1 was F1=37% and for T2, Accuracy= 61%.

In the study of Duwairi and Qarqaz (2014), Twitter and Facebook were used to create a corpus. 10,000 tweets and 500 Facebook comments were collected. Many of these comments were excluded because they were written in Latin characters or contained only emoticons and symbols. The tool from (Duwairi, Marji, Shaban, and Ershaidat, 2012) was used to classify Tweets, while Facebook comments were classified manually by the authors of the paper. The created corpus had 2591 comments and tweets (1073 positive, 1518 negative ones). RapidMiner was used for preprocessing tasks such as Tokenizing, Stemming, deleting stop words, and generating N-Grams. Arabic SA process was proposed based on three supervised MLs, which were NB, SVM, and KNN classifiers. 10-fold cross-validation was used to split the data into training and testing sets. SVM achieved the best performance with accuracy equaling to 75%.

In the study of Duwairi, Marji, Sha'ban, and Rushaidat (2014), 350,000 tweets were collected using Twitter's Application Interface (API). Tweets are classified manually using small tool. The tool showed the tweets to the user one by one and the user could choose one of the categories (positive, negative, neutral, and not applicable) for the tweets. The filtering process for the tweets was based on specific criteria: i) each tweet must contain at least 100

characters. ii) The tweet must not contain more than four hashtags. iii) The tweet must not contain mentions and links. iv) Tweet must be non-duplicate or non-retweets. Further processes were included such as converting Emoticons, dialects (Jordanian dialect) and Arabizi to MSA, removing repetitions and links. RapidMiner extension was developed to match the tasks of the study. The proposed framework included three ML algorithms: NB, SVM, and KNN. NB achieved a good performance result compared to the other algorithms.

In the study of Abdulla, Ahmed, Shehab, and Al-Ayyoub (2013), 2000 tweets were used to create a corpus. The tweets were written in MSA and the Jordanian dialect that covered different topics. Three experts manually classified the tweets into two categories: 1000 positive and 1000 negative tweets. Repeated letters and stop-words were removed, and the letters were normalized. In this study, corpus-based and lexicon-based approaches were proposed. SVM, NB, DT, and KNN were used as classifiers. Root stemming, light stemming and no stemming were tested on each classifier. SVM with light-stemming showed the best performance by accuracy equaling to 87%.

In the study of Aly and Atiya (2013), 220000 reviews are used to build a Large-scale Arabic Book Reviews (LABR) corpus for Arabic SA. The reviews were collected from “Goodreads” website. The collected dataset contained approximately 70% unwanted reviews such as reviews that were not written in Arabic or were not related to Arabic books. The unwanted reviews were deleted. The user of the website were able to add a rating of 1-5 as evaluation score with their review.

The authors assumed reviews with ratings (4 or 5) as positive, (1 or 2) as negative and review with rating 3 were considered neutral and not included in the polarity classification. Unigrams, bigrams, and trigrams with/without TF-IDF were used to extract features. Multinomial Naive Bayes (MNB), NB (for binary counts), and SVM based on Scikit-Python (Pedregosa et al., 2011) were used as classifiers. The authors used two approaches; balanced and unbalanced classification of reviews. SVM performed a better in use unbalanced approach, while in the balanced approach MNB was slightly better than SVM.

In the study of Abdul-Mageed and Diab (2012), a multi-genre corpus for Arabic SA “AWATIF” was represented. The authors used three sources to build the corpus: i) Penn Arabic TreeBank (PATB) (Maamouri, Bies, Buckwalter, and Mekki, 2004). ii) About 5342

sentences were selected from Wikipedia Talk Pages (WTP). iii) 2532 interrelated conversations were taken from seven Web forum (WF) pages. Simple (SIMP) and linguistically motivated and genre nuanced (LG) were identified as labelling guidelines. In SIMP form, to help people classify the sentences into one of three categories (positive, negative, or natural); the authors prepared two examples for each of these types to help annotators. In LG task, they exposed annotators to a linguistics background and explained the nuances of the genre to which each dataset belonged to. With the two types of guidelines the annotators were able to classify sentences under three conditions; GH-LG, GH-SIMP and AMT-SIMP. In GH-LG condition, the expert students specialized in linguistics (referred to Gold Human (GH)) classified sentences using the LG guidelines. In GH-SIMP condition, the GH team worked under SIMP conditions. AMT-SIMP, Amazon Mechanical Turk (AMT) was used to crowd-source with SIMP conditions.

In the study of Shoukry and Rafea (2012), about 4000 tweets were retrieved to create a corpus. Tweets that had one opinion were selected. Three experts manually classified 1000 tweets into two categories: 500 positive and 500 negative. User name, pictures, hashtags, URL, and all non-Arabic words were deleted in the preprocessing phase. Using Weka Suite Software, unigrams and bigrams models are utilized to extract features. NB and SVM were used to validate the performance of corpus. The results showed that SVM was more accurate than NB.

Rushdi-Saleh, Martín-Valdivia, Ureña-López, and Perea-Ortega (2011) presented Opinion Corpus for Arabic (OCA). The reviews were collected from several Arabic blog sites and web pages. Preprocessing tasks were applied to standardize the text format. HTML tags and special characters were deleted. Spelling mistakes were corrected manually. The authors used RapidMiner to Tokenize, delete Arabic stop words and obtain stem. OCA contained 500 movie reviews. 250 of them were considered as positive reviews and 250 as negative reviews. Different experiment tasks were applied to evaluate the corpus. Unigram, Bigrams, Trigrams, and TF-IDF were used to extract features. SVM and NB were used to compare the performance of OCA. In general, their results were promising.

Tables 2.1, 2.2 and Figure 2.1 summarize previous studies on Arabic language (MSA and dialects). Also, highlight the data resources and technique that used in these studies.

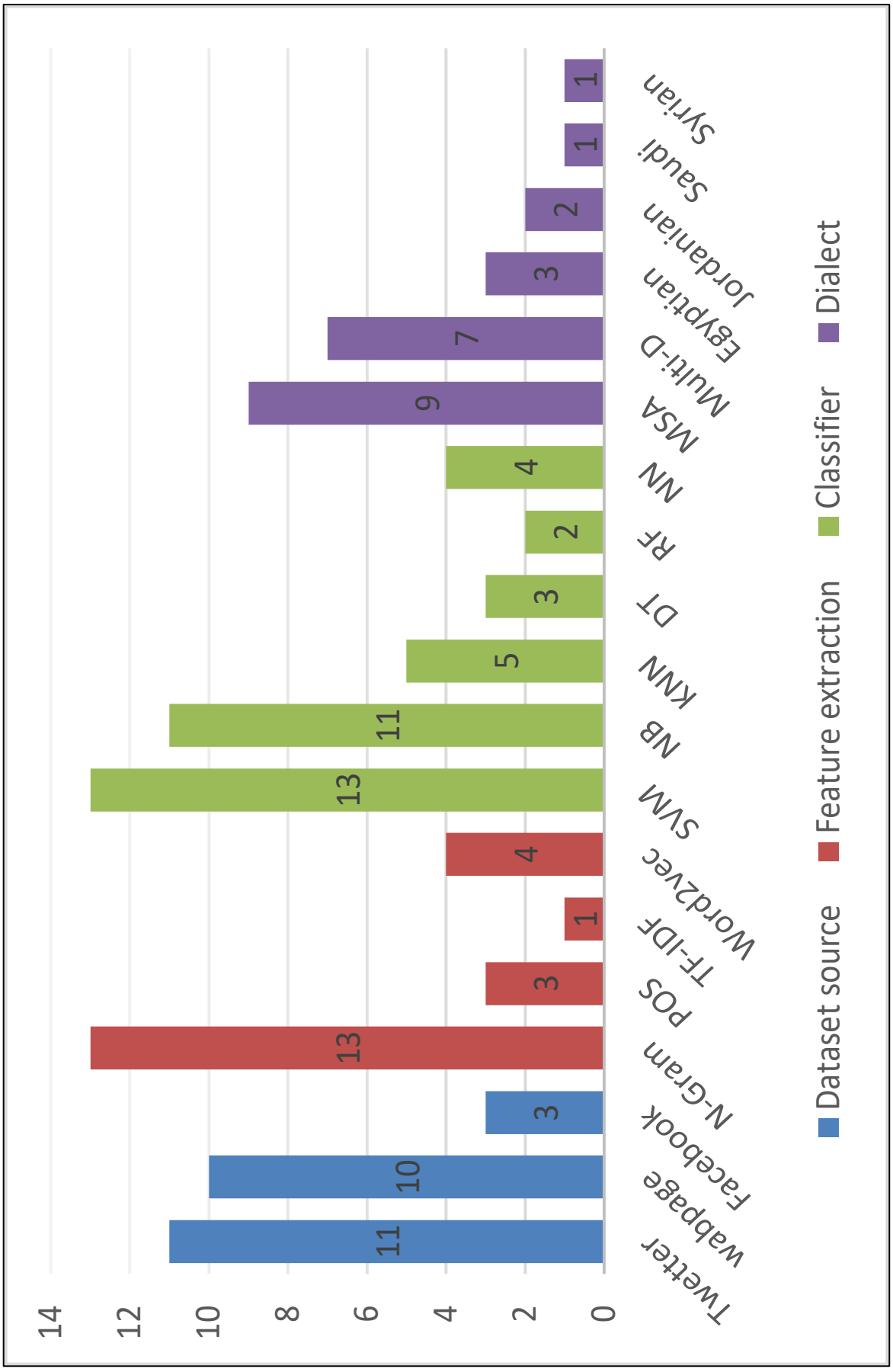

Figure 2.1. Statistics of literature studies

Table 2.1.The sentiment analysis for MSA.

Studies	Level	Feature extraction	Dataset source	Classifier	Tool	Evaluation	
						Criteria	Result
Cherif et al. (2015a)	Sentences	Low-level Light stemming	TripAdvisor website.	SVM	N/A	RMSE	0,83
						AM	0,57
Al-Smadi et al. (2015b)	Aspect-based	N-Grams	LABR	T1;T2;T3;T4	BRAT	F1-	0,23 T1
							0,15 T3
						Accuracy	0,29 T2
							0,42 T4
Cherif et al. (2015b)	Sentences	Light-stemming	TripAdvisor website	SVM+ KNN	N/A	Av. F1	0,97
AL-Smadi et al. (2015a)	Aspect-based	N-Grams, POS, NER	Facebook	CRF, J48	AraNLP	F1	0,37 T1
							0,61 T2
Duwairi and Qarqaz (2014)	Sentences	N-Grams,	Twitter and Facebook	NB, SVM, KNN	RapidMiner	Macro-Precision	0,66 NB 0,75 SVM 0,70 KNN
Duwairi et al. (2014)	Sentences	N-Grams,	Twitter	NB, SVM, KNN	RapidMiner	Accuracy	0,75 NB
							0,71 SVM
							0,51 KNN
Aly and Atiya (2013)	Sentences	N-Grams,	book readers social network www.goodreads.com	MNB, NB, SVM	Scikit-learn	F1	0,42 MNB
							0,21 NB
							0,41 SVM
Abdul-Mageed and Diab (2012)	Sentences	Did not extracted	PATB, WTP, WF	SIMP, LG	Manual annotation	Kappa (k)	0,82 ATB
							0,79 WTP
							0,79 WF
Rushdi-Saleh et al. (2011)	Sentences	N-Grams,	Movieswebpages	NB, SVM	RapidMiner	Accuracy	0,90 SVM
							0,89 NB

Table 2.2. The sentiment analysis for Arabic dialects

Studies	Level	Features extraction	Dialect	Dataset source	Classifier	Tool	Evaluation	
							Criteria	Result
Heikal et al. (2018)	Sentence	Word2Vec	Multi dialects	ASTD	CNN + LSTM	N/A	Accuracy	0,65
							F1	0,64
Alayba et al. (2018)	Sentence	Word2Vec	Multi dialects	Twitter	CNN	N/A	Accuracy	0,92
Itani et al. (2017)	Sentence	Did not extracted	Multi dialects	Facebook	Manual tagging	Manual tagging	Accuracy	0,73-0,96
Al-Azani and El-Alfy (2017)	Sentence	Word2Vec	Syrian	Twitter	Ensemble classifiers	Python Sci-kit-learn package	Average F1	0,63
Al-Sallab et al. (2017)	Sentence	AROMA	Multi dialects	ATB, QALB, Twitter	AROMA	AROMA	Accuracy	0,86 ATB 0,79 QALB 0,76 Twitter
							F1	0,84 ATB 0,75 QALB 0,68 Twitter
Al-Rubaiee et al. (2016)	Sentence	TF-IDF and BTO	Saudi Arabian	Twitter	NB, SVM	RapidMiner	Accuracy	0,83 NB
								0,79 SVM
							Precision	0,78 NB
								0,98 SVM

Table 2.2. (continued) The sentiment analysis for Arabic dialects

Studies	Level	Features extraction	Dialect	Dataset source	Classifier	Tool	Evaluation	
							Criteria	Result
Altowayan and Tao (2016)	Sentence	Word2Vec	Multi dialects	ASTD, LABR	SVC, RF, NB, NuSVC LR. SGD		Accuracy	0,81 SVC 0,80 RF 0,64 NB 0,81 NuSVC 0,81 LR 0,78 SGD
							F1	0,80 SVC 0,79 RF 0,58 NB 0,81 NuSVC 0,81 LR 0,80 SGD
Hathlian and Hafezs (2016)	Sentence	N-Grams	Multi dialects	Twitter	NB, SVM	Weka	Accuracy	0,84 NB
								0,84 SVM
							F1	0,83 NB
								0,84 SVM
Sghaier and Zrigui (2016)	Sentence	N-Grams	Multi dialect	reviewzat 1, jawal123 2, jumia3 web sites	NB, SVM	Weka	Precision	0,94 NB
								0,94 SVM
							F1	0,93 NB
								0,94 SVM
Shoukry and Rafea (2015)	Sentence	N-Grams	Egyptian Dialect	Twitter	ML + SO	Weka	Accuracy	0,80
							F1	0,97

Table 2.2. (continued) The sentiment analysis for Arabic dialects

Studies	Level	Features extraction	Dialect	Dataset source	Classifier	Tool	Evaluation	
							Criteria	Result
Ibrahim et al. (2015a)	Sentences	POS N-Grams	Egyptian dialectal	tweets, product review, hotel reservation comments and TV program comments	Manual annotation	N/A	Kappa	0,97
Abdulla et al. (2013)	Sentence	Unigram, Bigram	Jordanian dialect	Twitter	SVM, NB, DT, and KNN	Rapid Miner	Accuracy	0,87 SVM
								0,81 NB
								0,51 DT
								0,50 KNN
Shoukry and Rafea (2012)	Sentence	Unigram, Bigram	Egyptian dialect	Twitter	NB, SVM	Weka	Accuracy	0,65 NB
								0,72 SVM
							F1	0,65 NB
								0,72 SVM

3. BACKGROUND

In this section of the thesis, the main concepts in SA were explained. The historical background of the Arabic language was discussed. Moreover, Characteristics of the Iraqi Arabic dialects were highlighted.

3.1. Sentiment Analysis

SA is a computational study that is carried out to extract people's opinions, conditions, and feelings they exhibit towards incidents and circumstances (Abdul-Mageed, 2017). SA processes can be defined as multidisciplinary tasks, which use techniques from ML, NLP and computational linguistics to achieve various discovery tasks to identify and extract subjective information at different text-granularity levels (Chen et al., 2018). Every year, many SA studies are published, which discuss different aspects and scopes of the problem.

SA can be performed at three levels based on granularities, which are:

Document-level: At this level, the whole text is processed as one piece and is allocated to one class. This level considers that the document expresses only one opinion. It is supposed that the document is holding one opinion about a single entity only. In forums and blogs, this approach is not proper, because the document may include opinions of different entities.

Sentence-level: This level deals with each sentence as a separate text to identify whether the sentence has an opinion or not. This level of granularity is highly context-dependent.

Aspect-based level: Its purpose is to discover the opinions in a single sentence that contains more than one aspect (entity). In this level, there are two key tasks: aspect extraction and aspect sentiment classification. In aspect extraction task, the entity is specified. In aspect sentiment classification task, the sentiment polarity of various aspects is specified.

In general, the main tasks of SA process can be divided into four main tasks, as represented and summarized in Figure 3.1.

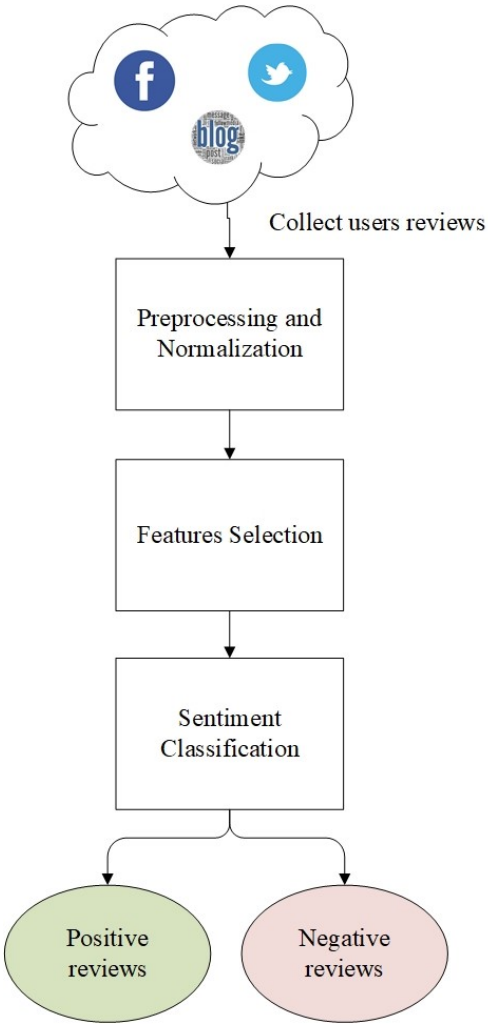

Figure 3.1. The main tasks of SA process

3.1.1. Collecting user's reviews

In SA field, the datasets used are an important issue. The sources of the reviews are mainly reviews made on web sites or social media pages. The datasets are not only made of product reviews but may also contain reviews on stock markets, news articles, or political debates. For example in political debates, it could visualize out people's opinions on specific election candidates or political parties. The election results can also be predicted from political posts. The social media pages and micro-blogging sites are considered a very useful source of information. Many people share and discuss their opinions about a specific topic freely. They are also used as main data sources in the SA.

3.1.2. Preprocessing and normalization

The preprocessing and normalization techniques for SA datasets are crucial tasks that usually compromise the success of SA process and algorithms (Tubishat, Idris, and Abushariah, 2018). In general, the algorithms benefit from standardization of the datasets. The preprocessing and normalization tasks are used various functions and transformer to change raw data into a structured form that is more proper for the estimation SA algorithms. If standardization of datasets is not implemented as required by the algorithms, they might individual features behave badly. The features, in this case, do not appear as standard, normally distributed data. Mainly NLP is used in these tasks. The main tasks that are used in preprocessing can be summarized as in Figure 3.2.

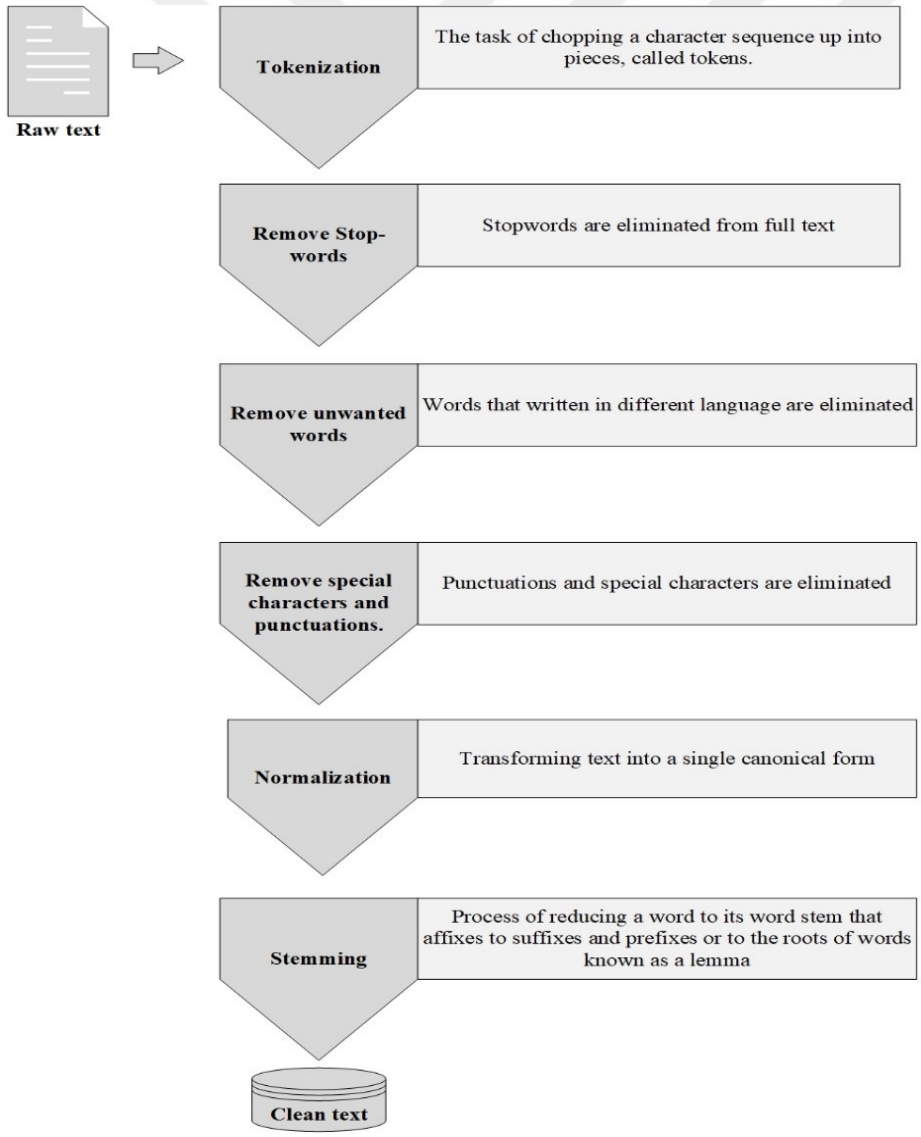

Figure 3.2. The main preprocessing tasks

3.1.3. Features selection

Many of SA techniques are based on ML methods. The feature extraction of texts plays an important role in these methods. The similarity of sentences can be calculated by text vector, a scheme that is based on feature extraction. In SA, Term Frequency-Inverse Document Frequency (TF–IDF) term weighting and word embedding are the most commonly used methods of feature extraction (Duwairi, Ahmed, and Al-Rifai, 2015). Some of the techniques used in extraction features are:

Terms frequency: This model is calculated as the frequency of words and the number of times they appear. Some approaches use one or zero weight to indicate the appearance of words. The other approach uses the value of term frequency to give weight to the importance of the word.

Parts of speech (POS): This approach looks for adjectives in the text; it is an important indicator of the evidence on the opinion.

Opinion words and phrases: This form searches for common words used to express opinions such as words of love and hate. Opinions, in this case, could be implicit in context.

Negations: This model looks for words that change the polarity from negative to positive or vice versa such as “not good” is equivalent to bad.

3.1.4. Sentiment classification

Methods can be divided into ML and linguistics-based approaches or a hybrid approach can be used. Figure 3.3 illustrates the main approaches of SA.

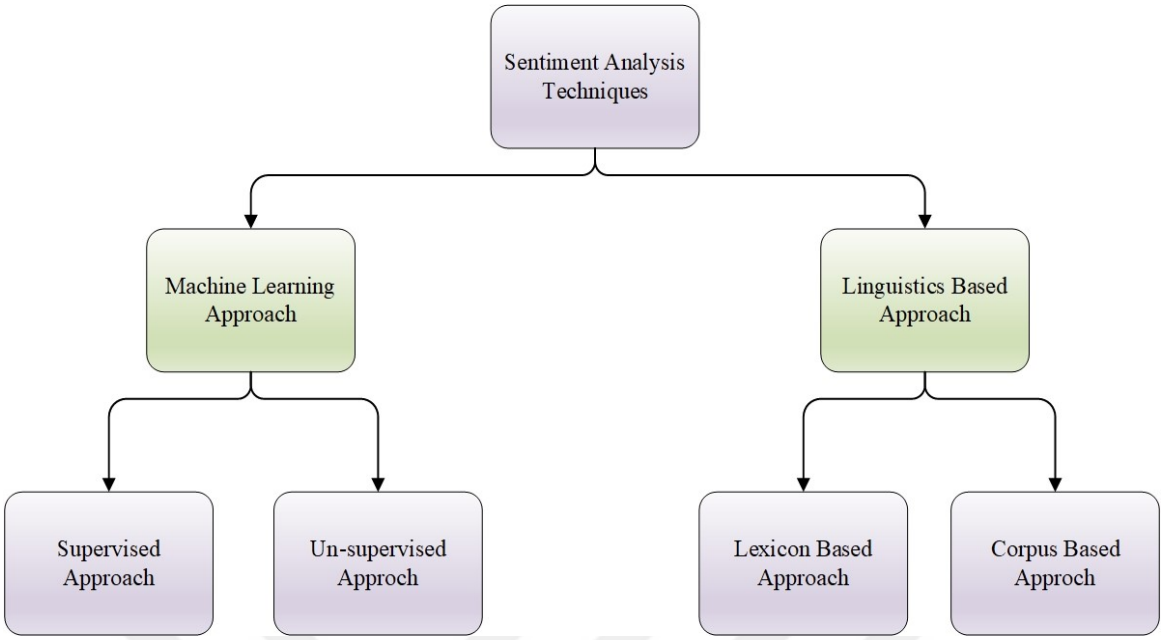

Figure 3.3. Sentiment classification techniques.

The ML approaches can be divided into supervised and unsupervised learning methods. In supervised ML approaches, a large number of training documents are used. The unsupervised ML approaches may be useful when it is difficult to find an appropriate number of training documents.

Linguistics-based approaches depend on the definition of the sentiments lexicon and can be divided into two main approaches; lexicon and corpus-based approaches. Lexicon based approach is based on finding synonyms and antonyms for words in the dictionary, and also uses the roots of words to obtain high accuracy. This approach uses a set of sentences that are classified into appropriate polarities. The process of preparing the corpus can be performed manually or through learning.

3.2. Arabic Language and Iraqi Arabic Dialects

The Arabic language is classified in one of the Semitic languages. Over 290 million people in 22 Arab countries in the Middle East and North Africa speak Arabic. The Arabic language is one of the five mostly spoken languages in the world (WorldAtlas, 2018) and, one of the five mostly used languages on the internet (InternetWorldStats, 2019).In addition, it is one of the formal languages in the united nations organization.

The Arabic language is written from right to left direction and it has a 28-letter alphabet. In addition, these letters can be turned into ninety elements by adding special vowels and shapes (ḥarakāt) (Abuata and Al-Omari, 2015).

Many researchers were considered the history and development of Arabic language, where the term "pre-Classical Arabic" appeared in these researches. Al-Sharkawi (2016: 207) presented three definitions to demystify of this term. The first definition considered that this term represents the age before the standardization of classical Arabic in the eighth and ninth centuries CE. Second, cconsidered that term refers to the variety of pre-Islamic poetry and the direct predecessor of Classical Arabic before its standardization. Thirdly, pre-Classical covers the total varieties of Arabic in pre-Islamic times after its epigraphic phase and it is an alternative term for old Arabic and pre-Islamic Arabic designations.

Versteegh (2014: 34-35) reviewed in his book the beginnings of the Arabic language, where he studied the texts (inscriptions) in other languages related to Arabic because of the official language of that period, Aramaic. However, the spoken colloquial language in that period was Arabic. A group of inscriptions belongs to the first century BCE was found in south Arabia, from Qaryat al-Faʾw (280 km north of Najrān), in Sabaean script, contains a Language that is closely related to Arabic. He also referred to two Nabataean scripts, which contains some features similar to Arabic. The first inscription found in northern Jordan "ʾUmm al-Jimāl" (± 250 CE). the second inscription found in Saudi Arabia "al-Ḥijr" (267 CE).

At the beginning of the era of Islam, there were two sources of literary Arabic available, the Qurʾān and the pre-Islamic poems (Versteegh, 2014: 60). The Qur'ān (The Holy Islamic book), as a holy book, was to varying degrees an accessible model of daily importance and relevance to the layman. Through this situation, came to be known as the Classical language into the consciousness of the common people on daily basis (Holes, 2004; Al-Sharkawi, 2016: 131). After the spread of Islam, millions of non-Arabs entered into the Arab empire, especially after the era of conquests, there was an urgent need to communicate among different linguistic communities in addition to communicating with the Arabs. Versteegh (2014: 61) indicated three reasons lead to standardize of the Arabic language as the language of the empire. "First, the Arab community was speaking in different dialects. Second, the policy of the central government was aimed at controlling its subjects not only in economic

and religious but also in linguistic matters. Third, the changed situation called forth a rapid expansion of the lexicon, which had to be regulated in order to achieve some measure of uniformity".

Nowadays, Arabic Language is used in two forms: (i) Modern Standard Arabic (MSA). (ii) Dialectal Arabic. MSA is derived from the language of the Quran and is usually used in schools, media, newspapers, literature, formal speeches, etc. MSA consists of a vocabulary size of more than 1.5 million words (Nasser, 2018: 9). The grammars of MSA, in general, follow the grammatical rules of the language of the Quran.

Dialectal Arabic is the second form of the Arabic language, which is used as the daily language in Arab countries. Versteegh (2014: 47) reviewed the theories on the development of the Arabic language and he mentioned, the pre-Islamic period there was already diglossia. Al-Sharkawi (2016) discussed dialects of Arabic in pre-Islamic period. He arranged his study on dialects geographically. He surveyed the dialects of the western part of the Arabian Peninsula. Then move to the southwest, then Central Arabia, Najd. Moreover, A number of tribes that were able to obtain linguistic sources around them reviewed in his research.

Moreover, it is possible to produce a text of dialectal Arabic using spelling rules that are similar to that of used in MSA, which are regularly phonetic. There is an understandable level of mutual reasonability among the dialects, but the ability to understand other dialects depends on person's own dialect, interaction with other Arab cultures, and his/her awareness of literary works from other countries. For example, in Iraq, it is not a problem to understand the Egyptian dialect, thanks to the popularity of the Egyptian movies and television shows. On the other hand, it is difficult to understand the Algerian dialect, especially in its spoken form.

Some recent studies have shown that the difference in Arabic dialects is not confined to the Arab countries but rather to some neighboring countries. Acat (2015) indicated the difference between the Arabic spoken in Turkey, where three areas speak different Arabic dialects, namely the Mersin – Adana – Hatay region, the Urfa region and the Diyarbakir – Mardin – Siirt region. On the other hand, some researchers point out that the importance of studying Arabic dialects has increased significantly. Suçin (2015) motioned there is a need

to study these dialects in Turkish universities "an example" to avoid difficulties for teachers and students of the Arabic language for non-native speakers.

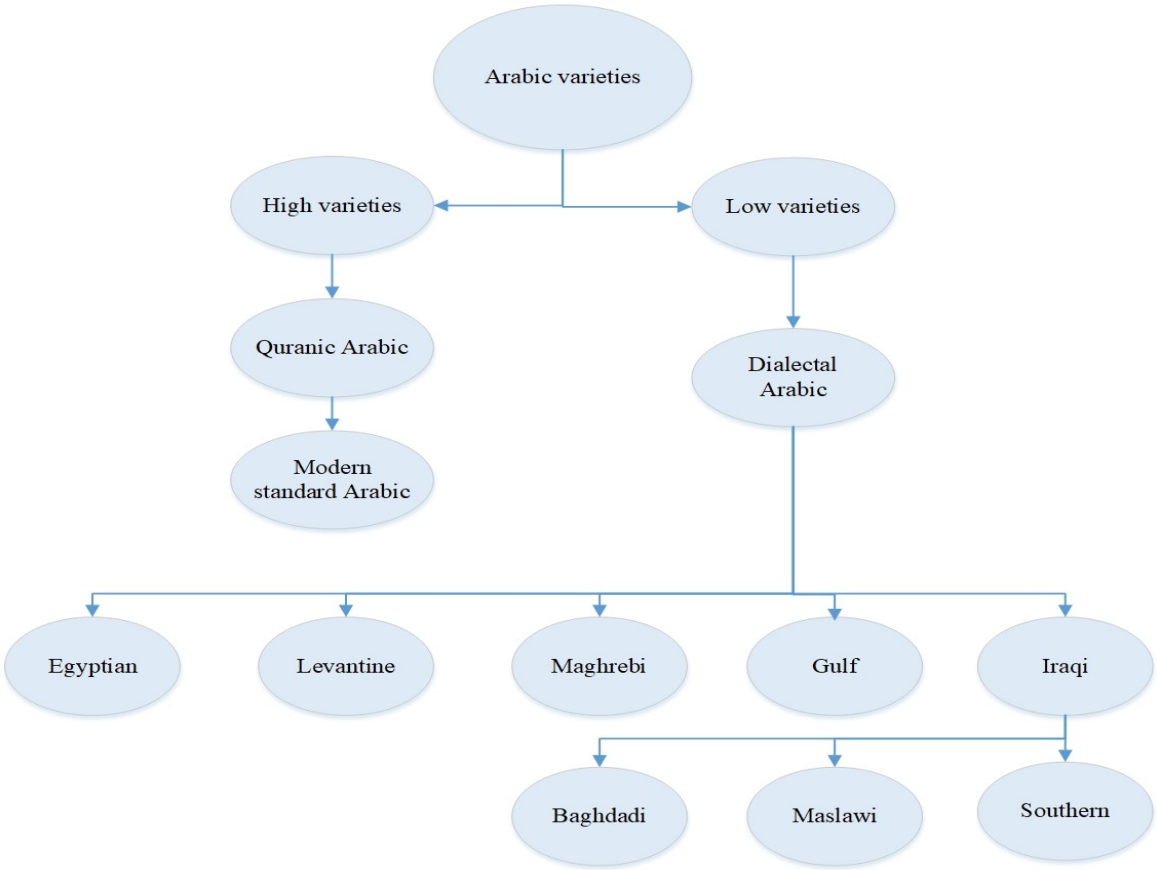

Figure 3.4. The qualitative derivation and variation of the Arabic language

The Arabic dialects can be broken down to five main groups as follows:

Egyptian**:** this dialect is the one the Egyptian people use. It is most widespread dialect over Arab countries due to the spread of the Egyptian films and series industry (Haeri, 2003).

Levantine**:** It is related to Aramaic, spread in the countries of the eastern coast of the Mediterranean (Bassiouney, 2009).

Maghrebi**:** It is a dialect that is incomprehensible by other speakers, especially regions in the Middle East. The French and Berber languages have an influence on it. This dialect is spoken primarily in Morocco, Tunisia, Libya, and Algeria.

Gulf**:** It is the closest dialect to MSA, and it is perhaps due to the fact that MSA was developed from an Arabic variety originated in the Gulf region (Versteegh, 2014).

Iraqi**:** This dialect is spread in Iraq and some neighboring Syrian cities. It is closer to Gulf dialects, though it has characteristics of its own in terms of phonology and morphology (Alnawas and Arıcı, 2019).

3.2.1. Iraqi Arabic dialects

Iraq consists of a community structure in which different languages are spoken. The majority of people speak Arabic, Kurdish, Turkmen, and other languages. Iraqi dialect is often close to MSA. The Iraqis can pronounce the MSA sound of the vocals that exit. IAD is the dialect spoken by Arabic speakers who live in or near Iraq. IAD is historically the result of interaction between Arabic, Turkish, Persian, Kurdish, Turkmen, and other languages. IAD is the accumulation of several linguistic layers that have passed through the history of Iraq: Sumerian, Akkadian, Babylonian, Assyrian, Aramaic, and then Arabic (Al-Bazi, 2005).

IAD is often close to MSA and the Iraqis can pronounce the MSA sounds of the vocals that exit. In addition to obtaining the vocabulary from Kurdish, Persian, Turkish, and English, IAD uses many words that do not exist in MSA. These words were received from other languages or adapted from the MSA. These words, although frequently used in everyday life, are not seen in standard Arabic books.

With the increasing of online communities, this dialect has taken on a broad scope as written texts in the Internet environment. IAD consists of three sub-dialects: *Moslawi*, *Baghdadi* and *Southern*. The *Moslawi* dialect is spread in northern Iraq (Mosul city); also, Tikrit city has a dialect that is close to this dialect. *Baghdadi* dialect is used in the Iraqi capital Baghdad and some neighboring cities like Diyala, Samara, Anbar, Babylon and in some areas of Tikrit. Southern dialect is prevalent in the southern cities such as Missan, Basra, Thi Qar, Qadisiyah, Karbala, and Najaf.

Blanc (1964) introduced the first division of the spoken dialects in Iraq. He divided the Iraqi dialects into two categories, *qiltu* and *gilit*. the linguistic division formulation of Blanc, partly regional, partly social. While Jews and Cristian speak *qəltu*, nomadic sedentarized

nomadic and Bedouin populations speak *gilit* dialects. The dialects of sedentary Muslim populations (city dwellers) follow geographical distribution: north of the Samarra Falluja line *qiltu* dialect prevail, the south of the same line *gilit* dialect prevail. In Baghdad, *gələt* speaking by Muslim, and *qiltu* speaking by Jews and Cristian.

Grigore has reinforced studies on the Iraqi dialect with many research papers especially for Baghdadi dialect. In study of Grigore (2014) the verb of perception *shāf* "to see" in Baghdadi Arabic is analyzed in terms of polysemy and semantic extension. In the study of Grigore (2005) an analytical study of conditional sentences in the Baghdadi dialect was presented. The study aimed to infer a system that may govern the choice of verbal and mood conditions spread in such sentences.

Ingham (2009) studied an interesting dialect in south Mesopotamian the so called *gilit* dialect. Ingham recorded material from three sources in spring of 1977, which showed a similar type of dialect. The sources of this material were from two areas in Nāṣiriyya in Iraq and one from Al-Rawḍatayyin in Kuwait. The sources showed a similar type of dialect.

Hassan (2015) indicated that the number of the studies on Iraqi Arabic dialects has recently increased, but at the same time little attention has been paid to the spoken South Iraqi Arabic. Most of studies focused mainly on the spoken varieties of Baghdadi Arabic and the northern language area. Where the spoken South Iraqi Arabic was again neglected.

To understand IAD, the following linguist terms were analyzed: phonology, morphology, and lexicon that is summarized from (Al-Bazi, 2005; Alnawas and Arıcı, 2019) studies.

In this thesis, the transliteration system of the International Journal of Middle East Studies (IJMES) is used to represent Arabic letters using Latin symbols. The Appendix-1 presents IJMES transliteration system.

Phonology

The IAD sounds are defined as consonants and vowels. Consonants are divided into two categories: *voiced* and *voiceless*. The voiced consonants are those, when pronounced, the

vocal cords are responded. The voiceless consonants have no respond from the vocal cords. Table 3.1. shows Consonants sounds in IAD with examples.

Table 3.1. IAD consonants

Voiceless					Voiced				
پ	/p/	پانكة	/pānka/	Fan	ب	/b/	باب	/bāb/	Door
ت	/t/	تنكة	/tanka/	Box	د	/d/	دروح	/darūḥ/	Go
ک	/k/	نكتة	/nukta/	Joke	گ	/g/	گال	/gāl/	Said
ف	/F/	فلكة	/filka/	Bend	ڤ	/v/	ايڤان	/Īvān/	Ivan
ث	/th/	ثوم	/thōm/	Garlic	ذ	/dh/	ذيب	/dhīb/	Wolf
س	/s/	سمرة	/samra/	Brunette	ز	/z/	زورق	/zawraq/	A boat
ش	/sh/	شمرة	/shamra/	Attitude	ژ	/zh/	ژيان	/zhiyān/	Girl’s name
چ	/ch/	چنة	/channa/	As if	ج	/j/	جنة	/janna/	Heaven
ه	/h/	هاي	/hāy/	This	أ	/ā/	أول	/awwal/	First
ح	/ḥ/	حليب	/ḥalīb/	Milk	ع	/ʿ/	عرفت	/ʿarafit/	I knew
خ	/kh/	خراب	/kharāb/	Destruction	غ	/gh/	غرام	/gharām/	Romance love
ق	/q/	قناة	/qanāt/	Canal	ر	/r/	رحت	/riḥit/	I went
ص	/s/	صور	/suwar/	Pictures	ل	/l/	لازم	/lāzim/	Must
ط	/ṭ/	طار	/ṭār/	Flew	م	/m/	ماكو	/māko/	There is not
ظ	/ẓ/	ظليت	/ẓallēt/	I remained	ن	/n/	انطاني	/anṭāny/	He gaves me
ض	/ḍ/	ضعت	/ḍiʿit/	I got lost	ي	/y/	يسوي	/ysawwy/	He does
					و	/w/	ويا	/wayā/	With him

In sub-IAD, alternative consonants are used as diaphones. When diaphones are used, the meaning of the words does not change. Table 3.2 shows these cases.

Table 3.2. Alternative consonants in Sub IAD

Original consonants	Alternative consonants	Sub IAD
/r/	/gh /	Moslawi
/q/	/g/	Baghdadi and Southern
/k/	/ch/	Southern
/ch/ or /j/	/q/	Southern
/g/	/q/	Baghdadi, Southern and Moslawi

We can recognize 10 pure vowels in the IAD. Table 3.3. Shows these vowels with an example for each one.

Table 3.3. The IAD Vowels

Vowels	Arabic example		In English
/i/	/riḥit/	رحت	I went
/ī/	/baʿid/	بعيد	Far
/iyy/	/mitit/	مِتِت	I died
/ai/	/taiyr/,/tayr/	طَير	Bird
/a/	/ḥaḍarit/	حضرت	I attended
/ā/	/bāmya/	باميا	Okra
/u/	/mukhtār/	مختار	Neighborhood mayor
/ū/	/mū/	موو	Not
/au/	/shaurba/	شوربة	Soup
/uww/	/khōr/	خور	Marshland

Morphology

Unlike MSA, the words in IAD end with the consonant letters rather than vowels. The grammar case is not shown at the end of words.

The past tense**:** The past tense verb in IAD starts with a 2- consonant. This is similar to ancient Mesopotamian languages and Assyrian dialects. Usually, a "أ" /a/ glottal is placed before the verbs in the past tense.

The future tense**:** The future tense indicator for IAD is "raaḥ" راح, which means "went". It is inserted before the present tense. In MSA, the prefix " s " or " sawfa " used before the present verb to indicate future tense.

The continuous tense**:** MSA does not have any structure to indicate the continuous tense. Sometimes, the adverb "now" /al-ān/ الآن is used as aspect to get the present tense. In IAD, however, the particle /da/ or /jaay/ جاي is used to refer to the continuous tense.

Table 3.4. Examples of tenses in MSA and IAD

Tense	MSA		IAD		English
The Past	ذهبت إلى البيت	/dhahaebtu ilā al-bayti/	رحت للبيت	/riḥit lilbayt/	I went home.
The future	سوف أراك سأراك	/sawfa ʾarāka/ /saʾarāke/	حشوفك (Baghdadi) رح اشوفك (Southern)	/ḥashūfak/ /rḥ ʾashūfak/	I will see you.
The Continuous	أنا آكل آلان	/anā ākul al-ān/	أني دا آكل (Baghdadi) جاي آكل (Southern)	/āni dā ākul/ /jāy ākul/	I am eating.

In IAD, there are some special cases that distinguish it from MSA, Table 3.5 shows these cases.

Table 3.5. Examples of special cases between MSA and IAD

Cases	MSA		IAD		English	Note
Plural number	عندي كتابان	/ʿindy kitābān/	عندي كتابين	/ʿindy kitābēn /	I have two books.	The numbers are always in the accusative case. (also dual number)
Relative pronoun	الشخص الذي التقيت به	/al-shakhṣ alldhī iltaqaytu bihi/	الشخص اللي شفت	/alshakhṣ illi shifit/	The person with whom I met.	Relative pronoun in the IAD is (اللي) /illi/ regardless of gender, number
Orders and commands	نم	/nam/	نام	/nām/	Go to sleep	The speakers of IAD usually keep the long vowels when they give orders and commands.
The passive form	يقتل	/yaqtulu/	ينكتل (Baghdadi) ينچـتل (southern)	/yinkitil/ /yinchitil/	He will be killed	In IAD usually measure VII (seven) verb is used to express passivity

Lexicon

IAD uses words that do not exist in MSA. These words either are reformed forms of MSA and the classical Arabic or received from other languages. They are used commonly in daily life like the language of society. The speech of the dialects is not considered a language to be written in books.

Therefore, these words cannot be found in any Standard Arabic books. This constitutes a new language that has its own phonology, morphology and grammar. These words come in multiple forms such as verbs, adjectives, and adverbs; or functional words such as question words, prepositions, demonstratives, relative pronouns, and vocative particles.

IAD verbs

The new learners of MSA cannot understand the speech of the Iraqi people as the Iraqis use the dialect verbs when speaking with each other. Learners, when learning MSA in courses or at schools have no access whatsoever to IAD verbs. Table 3.6 lists some examples of verbs that are frequently used in conversations.

Table 3.6. Examples of IAD verbs

MSA		IAD		English
يتحدث	/yataḥaddathu/	يسولف	/ysōlf/	Speak
يعمل	/yaʿmalu/	يسوي	/ysawwi/	Work
يطلب الإذن	/yatlubu ʾal-idhn /	يترخص	/yitrakhkhaṣ/	Ask for permission
يعطي	/yuʿṭy/	ينطي	/ynṭy/	Give
يخرج	/yakhruju/	يطلع	/yaṭlaʿ/	Go out
يترك	/yatruku/	يخللي	/ykhally /	Let /make
يتكلم	/yatakallam/	يحجي	/yiḥchi/	Speak
يرمي	/yarmy/	يذب	/yidhib/	Throw
يمزح	/yamzaḥu/	يتشاقا	/yitshāqā/	Is joking / is kidding
يحمل	/yaḥmilu/	يشيل	/yshīl/	Carry

IAD nouns

In IAD, there are hundreds of nouns cannot be found in MSA or other Arabic dialects. These nouns are used by Iraqis to refer to the things in their houses. They are authentic Iraqi nouns. Table 3.7 presents examples of IAD nouns.

Table 3.7. Examples of IAD nouns

MSA		IAD		English
منضدة	/minḍada/	ميز	/mēz/	Table
سرير	/sarīr /	قريولة	/quryōla/	Bed frame
أحذية	/aḥdhiya/	قنادر	/qanādir/	Shoes
شراشف	/sharāshif /	چراچف	/charāchif/	Sheets
مفك	/mifak/	درنفيس	/darnafīs/	Screwdriver

Table 3.7. (continued) Examples of IAD nouns

MSA		IAD		English
غنم	/ghanam/	طلي	/ṭily/	Sheep
خيمة	/khayma/	چادر	/chadar/	Tent
إطار	/ʾiṭār/	چرچوبة	/charchūba/	Frame
قيود	/quyūd/	كلپچات	/kalapchāt/	Handcuffs
قدح	/qadaḥ/	گلاص	/glāṣ/	A glass

Functional words in IAD

Functional words have been modified a lot as they are frequently used in sentences with less stress and they cannot be neglected in any sentence. In IAD, They are generally said softly and quick. Furthermore, they can be heard in different forms. Table 2.8 shows examples of functional words in IAD.

Table 3.8. Examples of IAD question words

MSA		IAD		English
ما؟	/ma/	شـ؟	/shə/	What ?
أين؟	/ayna/	وين؟	/wēn/	Where?
متى؟	/matā/	يمتة؟ شوكت؟	/yamta/ or /shəwakət/	When?
مَنْ ؟	/man huwa/	منو؟	/minu/	Who?
كيف؟	/kayfa/	شلون؟	/shlōn/	How?
لماذا؟	/limādha/	ليش؟ علويش؟	/lēsh / or /ʿalawēsh/	Why?
بكم؟	/bikam/	ببيش؟	/əbbaysh/	How much?
كم؟	/kam /	چم؟	/cham/	How many?

Adjectives

Many of IAD adjectives are adapted from the MSA or received from other languages. These adjectives do not use in the formal discussions since they are not Arabic. However, they are considered as high-frequency words in the IAD.

Table 3.9. Examples of IAD Adjectives

MSA		IAD		English
مجاني	/ma/	بلاش	/balāsh /	For free
تالف	/tālif/	خربان	/kharbān/	Damaged
جيد	/jayyid/	زين	/zēn/	Well, nice
طازج	/ṭāzaj/	تازة	/tāza/	Fresh/ delicate
للغاية	/lilghāya/	كلش	/killish /	A lot
كثير	/kathīr/	هواي	/hwāy/	A lot
قليل	/qalīl/	شوية	/shwayya/	A little

4. IRAQI ARABIC DIALECTS SENTIMENT ANALYSIS

In this section of the thesis, the proposed approach is presented and explained, which is implemented in order to achieve the objectives of the thesis. This approach supports two main tasks; building of a model and prediction of the sentiment. The proposed approach is related to the IAD only and has not been experimented on other dialects. The proposed model includes a set of tasks as shown in Figure 4.1.

The proposed approach consists of five main tasks; dataset collection, preprocessing, feature extraction and representing words as vectors, training vectors with ML and creating a prediction model and evaluation of the model.

4.1. Datasets

With the limited available resources of labelled IAD datasets, the proposed approach assumes the use of labelled MSA and Arabic dialects datasets (positive or negative) with unlabeled-IAD (U-IAD). The first task is collecting and defining a dataset that is used to measure the performance of the proposed approach. Datasets have three types, the first one is based on pre-labelled datasets from previous research and it is used for training the proposed approach in Doc2Vec and ML training task. The second dataset is U-IAD that contains Facebook reviews and is used in a Doc2Vec training task. The third dataset is labelled IAD that contains Facebook reviews, which are classified manually by experts. The dataset will be used in Doc2Vec training, ML training, and test tasks. Facebook reviews are collected from different pages that use IAD. Table 4.1. explain the use of the dataset in the proposed approach.

Table 4.1. The use of a dataset in the proposed approach

Dataset	Doc2Vec training task	ML training task	ML test task
Labeled IAD	Used	Used	Used
pre-labelled MSA and other Arabic dialects	Used	Used	
Unlabeled-IAD	Used		

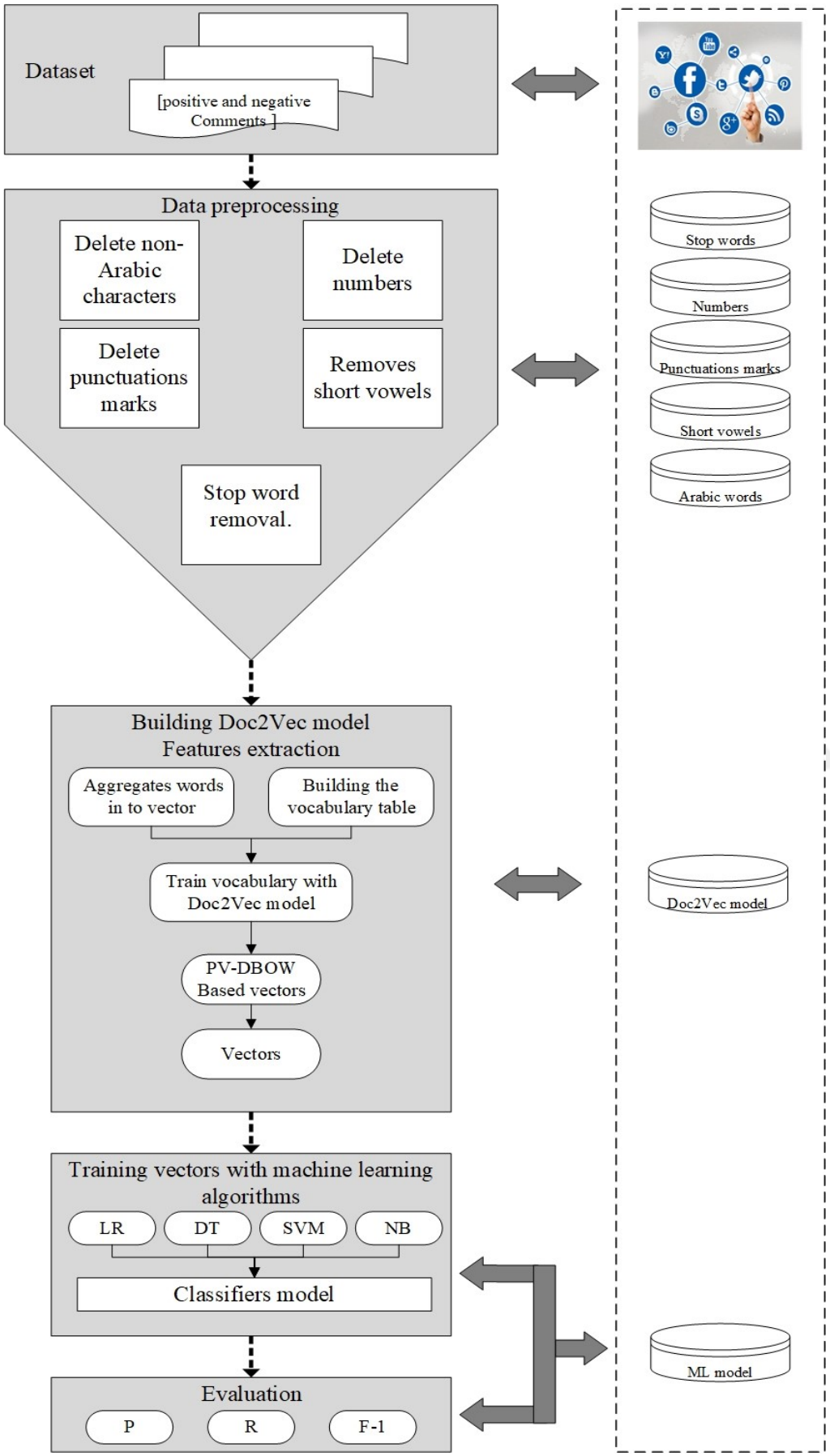

Figure 4.1. The proposed sentiment analysis process for Iraqi Arabic dialect.

4.2. Preprocessing Approach

The second task of the proposed approach discusses preprocessing. The textual data comes mostly containing noise as well as it is most probably unstructured. Preprocessing aims to eliminate unnecessary terms and standardize the text format to minimize complexity. Figure 4.2 shows the steps that were used to clean and structure the text data.

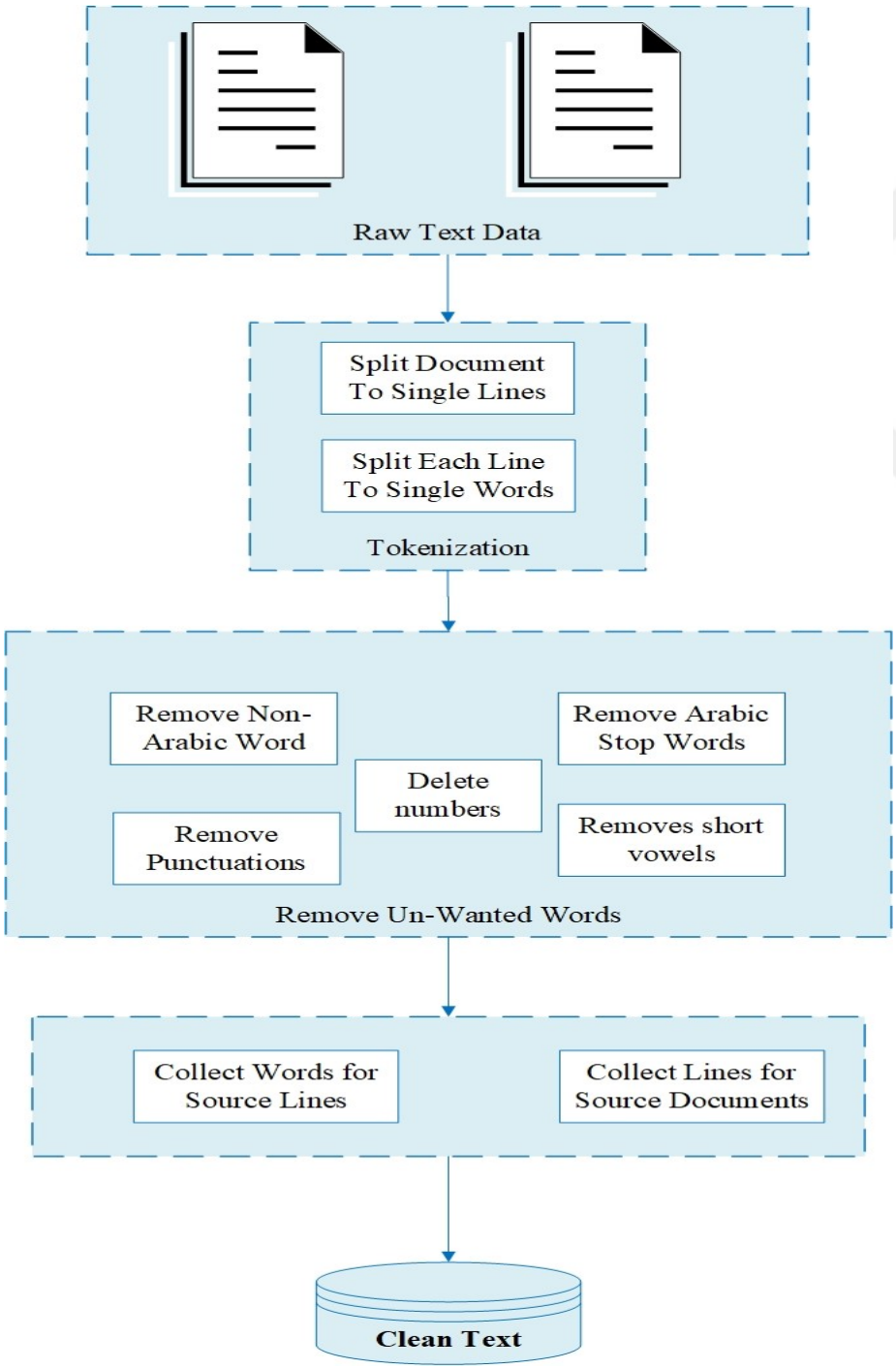

Figure 4.2. Pre-processing steps applied in this study

4.3. Doc2Vec Model

Word embedding models are effective models for representing words as vectors (Mikolov et al., 2013a; Pennington, Socher, and Manning, 2014). These models are used for learning a group of words and representing them as vectors depending on semantic of words (Hayran and Sert, 2017). In 2013, Google developed a model for word representation (Word2Vec). In this model, vectors that are close to each other together represent the words with the same semantic meaning. This model is proposed by two architectures; Continuous Bag of Words (CBOW) and Skip-Gram. Figure 4.3. highlights the difference between CBOW and Skip-Gram architectures.

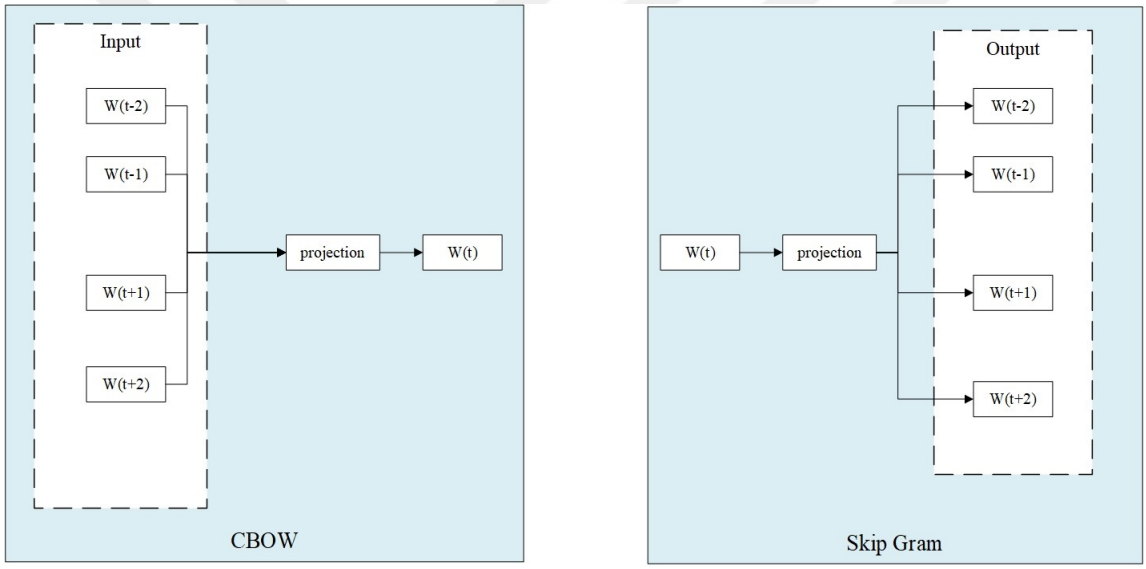

Figure 4.3. CBOW and Skip Gram architectures (Mikolov et al., 2013a)

Suppose we have a string of words (sentence): w_{t-2}, w_{t-1}, w_t, w_{t+1}, w_{t+2}. In CBOW architecture, the model receives w_{t-2}, w_{t-1}, w_{t+1}, w_{t+2} as input to predict a target word w_t. Therefore, CBOW is trained to predict the word based on the given context. In Skip Gram architecture, the Skip-Gram uses w_t to predict other words in the context.

To illustrate the working of these two architectures, the sentence “the house is very beautiful”, which has five words [the, house, is, very, beautiful] is used. In CBOW, [the, house, very, beautiful] are trained to predict the word “is”. In Skip Gram, trained “is” is used to predict [the, house, is, very, beautiful]. Figure 4.4 illustrates the two architectures with an example.

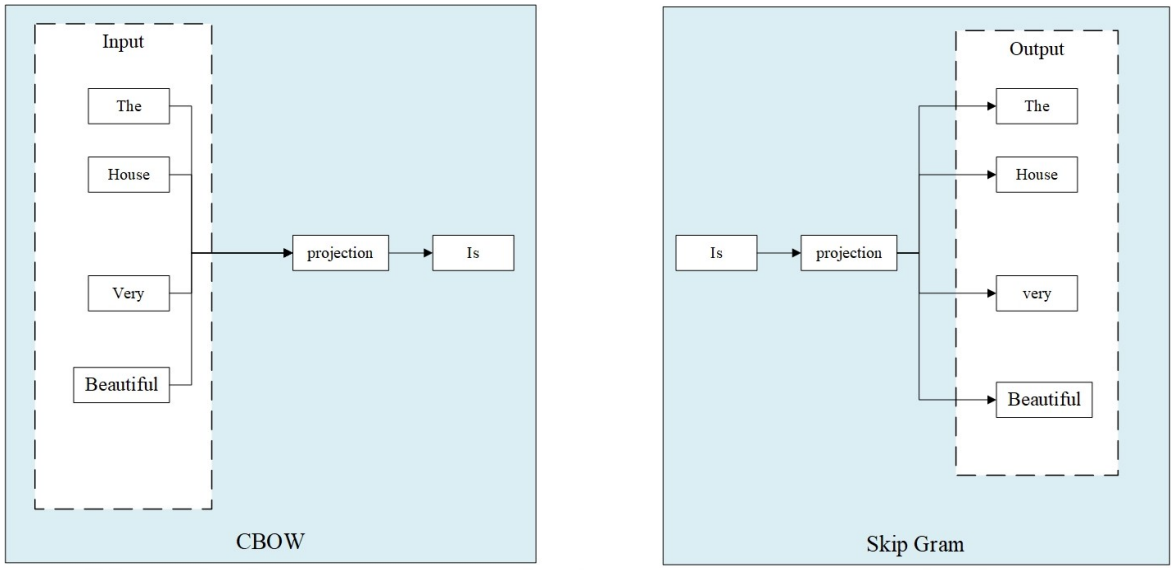

Figure 4.4. Illustrate the work of CBOW and Skip Gram.

The input into CBOW and Skip Gram models is a one-hot representation, not the word itself. For example, the four words “Baghdad”, ”Cairo”, ”Queen” and “King” can be represented as in Figure 4.5.

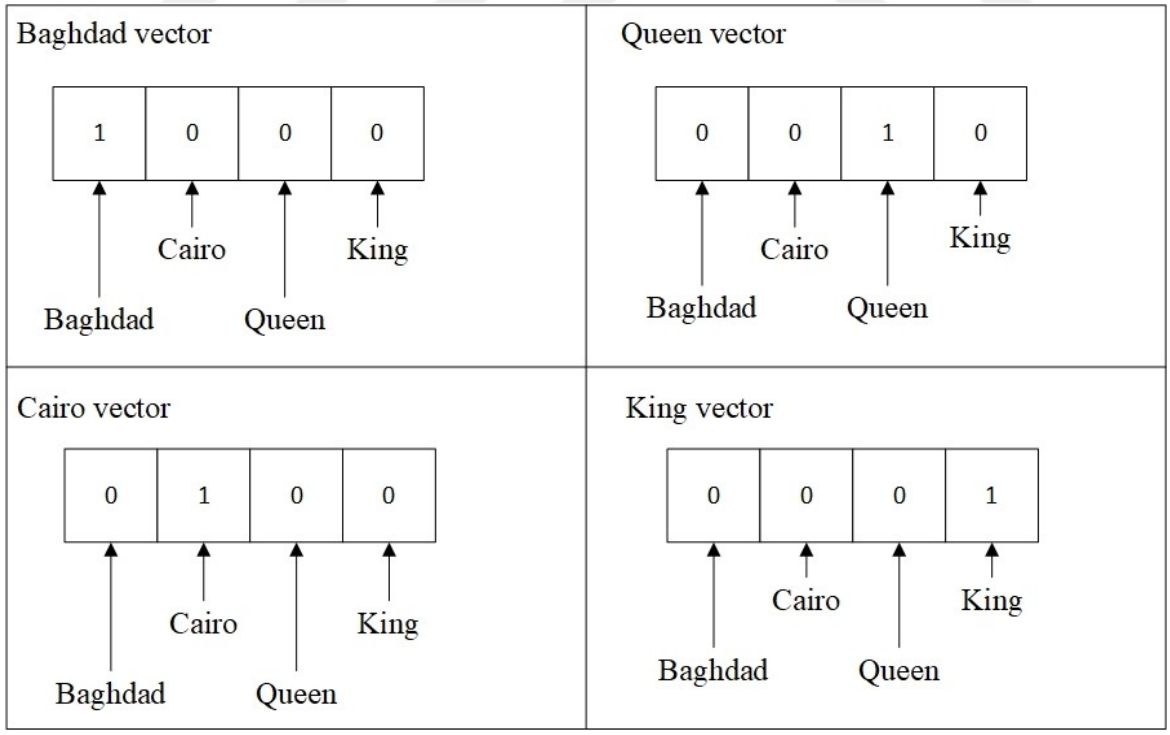

Figure 4.5. One-hot representation

The relationship between words cannot be found using one-hot representation. Word2Vec represents each word using weights. The word representation is spread over all the vector entries instead of being spread on only one entry where every entry contributes to definition of many words (Suleiman, Awajan, and Al-Madi, 2017).

In Word2Vec framework, every word is mapped to a unique vector and is represented by these vectors in a column in a matrix *W*. Depending on the location of the words in the vocabulary table, the columns of *W* matrix are indexed. The main task of this framework is to predict the next probable word in a word set. To achieve this goal, the sum of the vectors is used as a unique feature (Le and Mikolov, 2014).

Suppose we have a sequence of training words $w_1, w_2, w_3, ..., w_T$, where T represents number of words in vocabulary table. The objective of the Word2Vec model is to maximize the probability of the average log

$$\frac{1}{T} \sum_{t=k}^{T-K} \log p\ (\ w_t | w_{t-k}, \dots, w_{t+k}) \tag{4.1}$$

The prediction task is typically performed by means of a multiclass classifier, such as Softmax. There, we have

$$\mathrm{p}(\ w_t | w_{t-k}, \dots, _{t+k}\ ) = \frac{e^{y w_t}}{\sum_i e^{y i}} \tag{4.2}$$

y_{wt} represent the probability of center word w_t , and each of y_i is an un-normalized log-probability for each output word i, computed as

$$y = b + Uh(w_{t-k}, \dots, w_{t+k}; W) \tag{4.3}$$

Where *U*, *b* are the Softmax parameter, h is constructed by concatenation or average of word vectors extracted from *W*. After the training, words of similar meaning are represented in a similar location in the vector space. For example, “Baghdad” and “Cairo” are represented in convergence. “Queen” and “Baghdad” are represented far from each other. Figure 4.6 shows the vectors of the words “Baghdad” and “Queen” after Word2Vec is used.

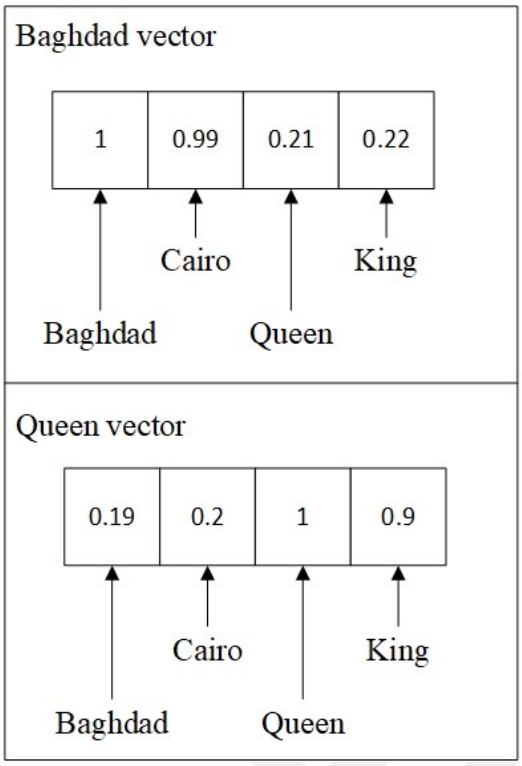

Figure 4.6. Distributed representation using Word2Vec

This difference in the representation of words carries a meaning. For example, simple vector algebra can be used with the word vectors to answer analogy questions (Mikolov et al., 2013c). The word vector can also be used to learn the linear matrix, which is used to translate words and phrases between different languages (Mikolov, Le, and Sutskever, 2013b). All of these properties have made the word vector technology attractive to many NLP such as natural language understanding (Collobert and Weston, 2008; Zhila, Yih, Meek, Zweig, and Mikolov, 2013), language modeling (Mikolov, 2012; Mnih and Teh, 2012), statistical machine translation (Mikolov et al., 2013b; Zou, Socher, Cer, and Manning, 2013), relational extraction (Socher, Huang, Pennin, Manning, and Ng, 2011) and image understanding (Frome et al., 2013).

Learning paragraph vectors approach is inspired by learning the word vectors method (Le and Mikolov, 2014). In this approach, the vector of words contributes to predicting the next word in the sentence. Although word vectors are randomly accented, they can capture semantics as an indirect result of the prediction. Le and Mikolov (2014) proposed Doc2Vec model in two architectures; Distributed Memory Model of Paragraph Vectors (PV-DM) and Distributed Bag of Words version of Paragraph Vector (PV-DBOW).

In PV-DM model, each paragraph is assigned with a unique vector. This vector is represented in the D matrix by a column. Each word is also assigned with a vector and represented by a column in the matrix W. Paragraph and word vectors represent a task to predict the next word in context by averaged or concatenated methods. By comparing PV-DM model with the word vector model, the change in equation 4.3, where *W* and *D* are used to construct *h*. In PV-DM model, the paragraph token works as a memory to retrieve what is missing from the context or subject of the paragraph. For this reason, it was called the Distributed Memory Model of Paragraph Vectors.

In the training time, paragraph and word vectors are trained using stochastic gradient descent and the gradient is obtained via backpropagation. Figure 4.7 shows the PV-DM framework.

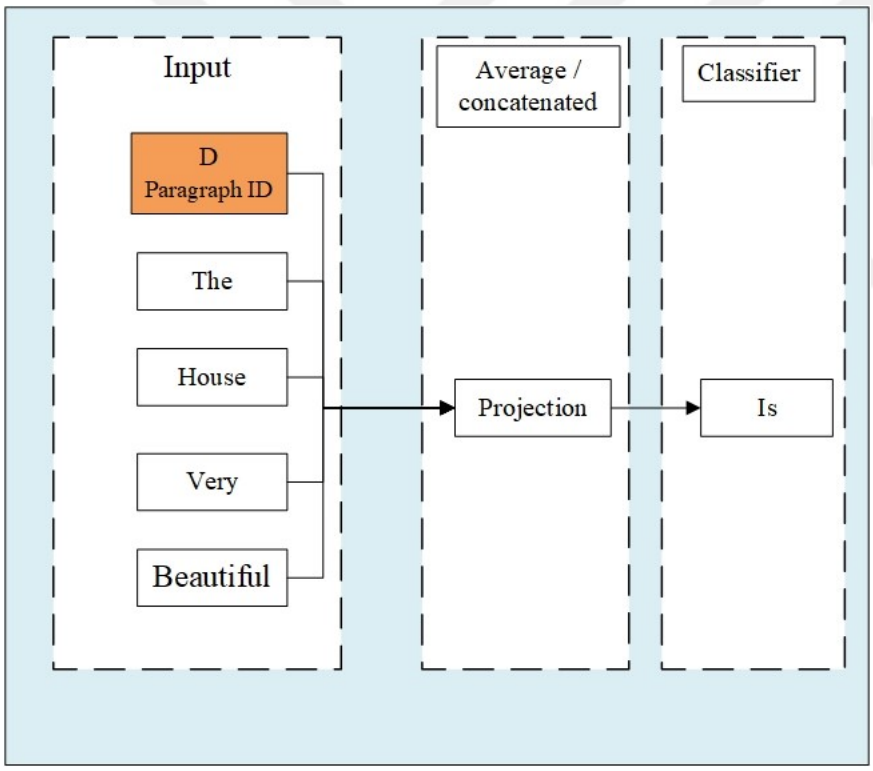

Figure 4.7. PV-DM framework.

The framework of PV-DM is similar to the framework of CBOW. The only difference is that paragraph ID is added which is assigned to a vector by D matrix. The framework illustrates the use of the context of four word vectors with the paragraph vector to predict the fifth word.

In PV-DBOW model, a paragraph vector is used to classify words throughout the document. It ignores the context of the words in the input. However, they force the model to predict the random words of the paragraph. Figure 4.8 shows the PV-DBOW framework. The advantages of this model; it is simple and it requires fewer data storage. This model is similar to Skip Gram.

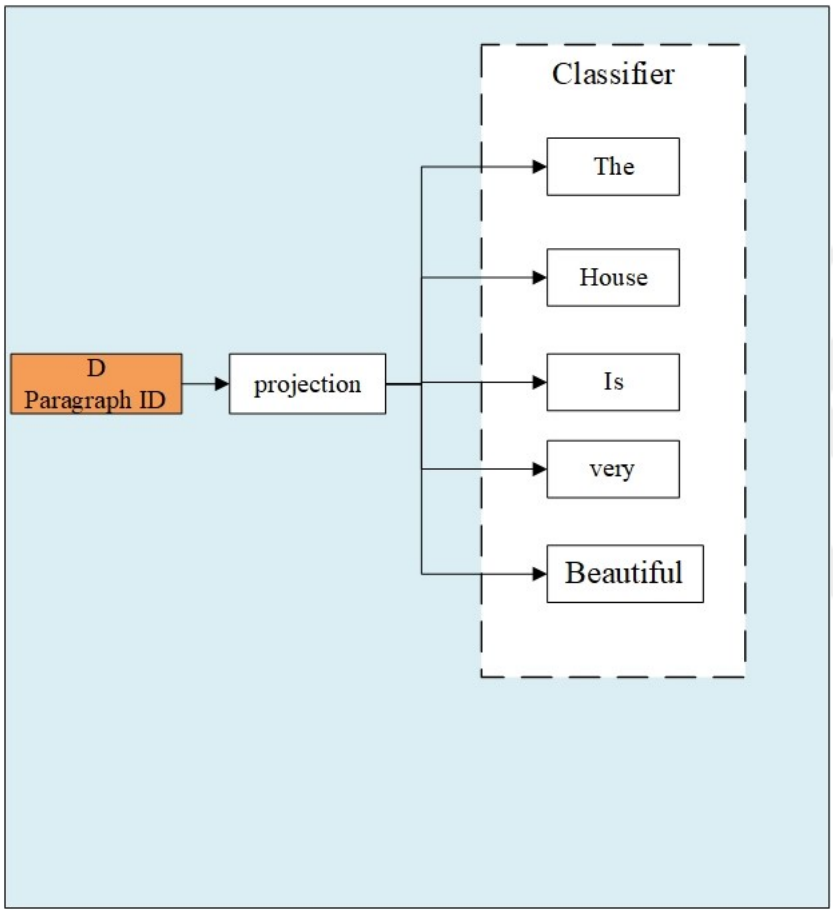

Figure 4.8. PV-DBOW framework

4.4. Machine Learning Approach

After training vectors of paragraphs and words, these vectors can be used as features. These features can be fed directly into different ML techniques. In this thesis, there are two target classes; positive and negative target classes. That means a binary classification model is needed to build. This model is trained using PV-DBOW feature vectors. In this study, four ML algorithms are used. These algorithms were selected according to their performance and reliability in SA studies (Boudad, Faizi, Thami, and Chiheb, 2017; Alnawas and Arıcı, 2019), which are LR, DT, SVM, and NB.

LR is one of the statistical methods used in machine learning. Regression analysis is appropriate to conduct when the dependent variable is dichotomous (binary). LR is also used to describe data and explain the relationship between one dependent variable and one or more independent variables. LR proposes a logistic curve, which is limited to values between 0 and 1. The curve is constructed based on the natural logarithm of the probabilities of the target variable. Furthermore, the predictors do not have to be normally distributed or have equal variance in each group. Figure 4.9 shows the optimal curve using LR.

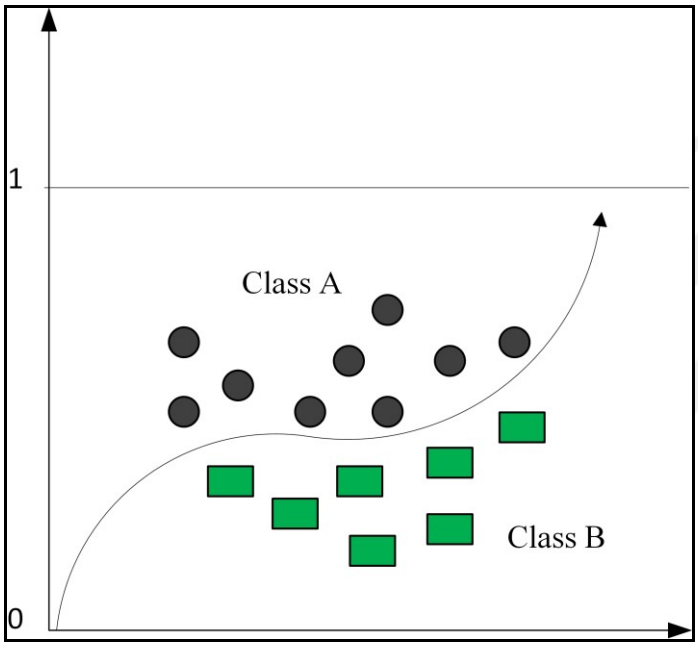

Figure 4.9. LR curve

SVM is one of the supervised learning models that have many characteristics that have motivated researchers to use it. SVM model is used for classification and regression problems. Moreover, SVM can be used efficiently to classify the text data because of the sparse nature of the text. SVM has many advantages when used with distributed word approaches such as; effective performance with large dimensional representation, the effectiveness remains stable even if the number of dimensions is greater than the number of samples, and its effects in memory because of its use a subset of training points in a decision. The basic idea of SVM is to find the optimal separation line (hyperplane) between the classes as in Figure 4.10.

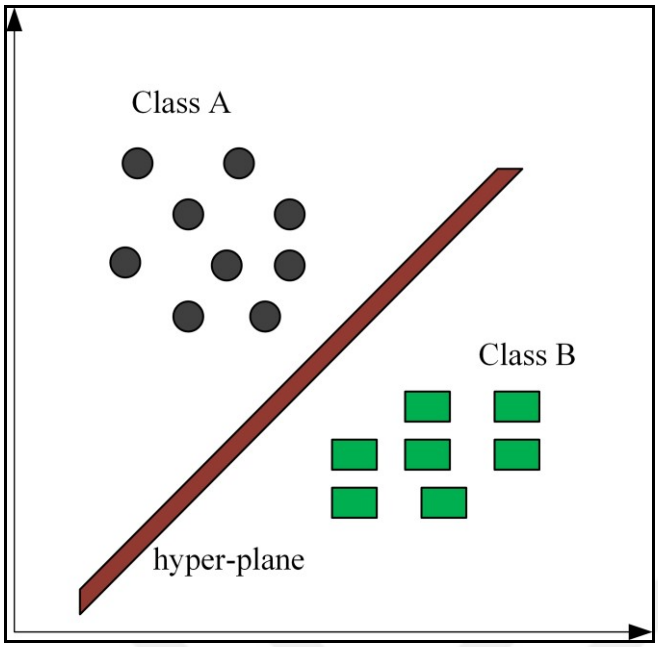

Figure 4.10. SVM hyperplane in 2D

It is one of the most commonly used methods of data mining and classification. It consists of a hierarchical structure (a set of nodes) representing training data. The split of nodes depends on a set of conditions. Each time a question is asked about the attributes, and when an answer is received, follow up on the next question until a conclusion about the class label of the record is reached. Figure 4.11 shows the hierarchical structure of the Decision Tree.

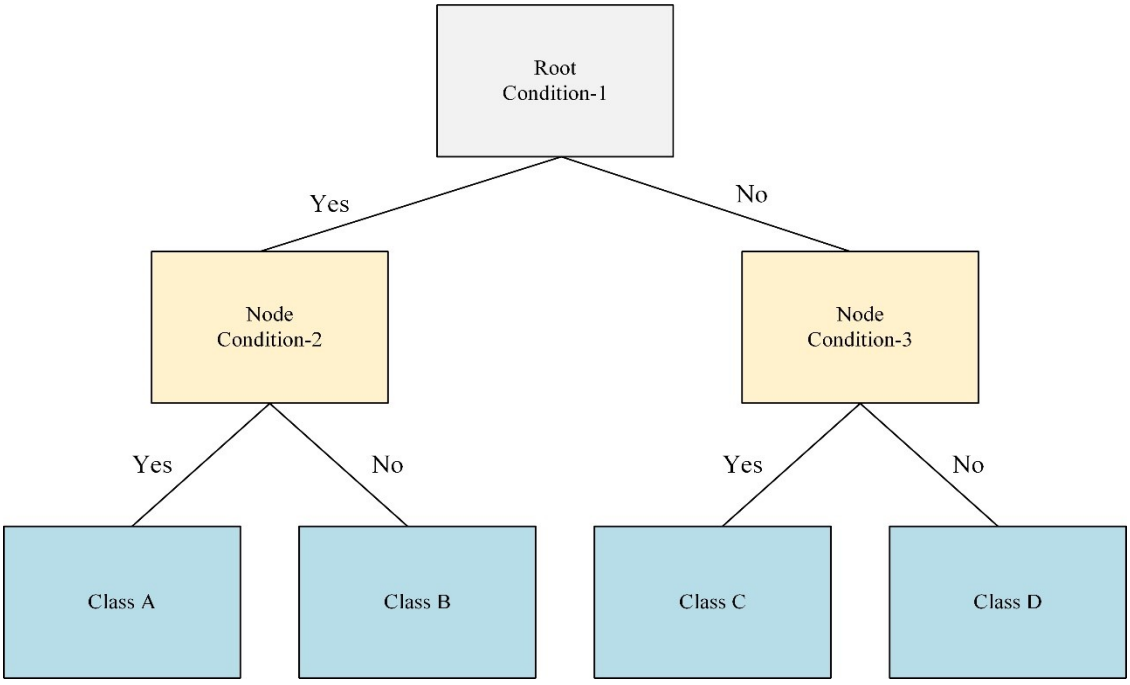

Figure 4.11. Decision Tree division depends on conditions

It is an ML technique based on “Bayes’ Theorem” with an assumption of independence among predictors. This technique assumes that having a feature in a particular category is not associated with another feature. Building a model through this method is easy and useful for large datasets. It outperforms sophisticated classification methods.

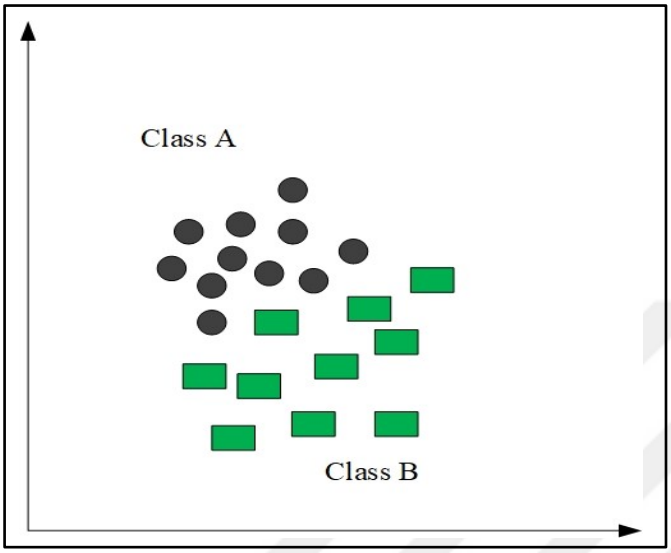

Figure 4.12. NB classification

The evaluation of model is an important part of the model development process. It helps to study the model behavior with our data and how will work in the future. The task of the confusion matrix is highlighted at this point. The performance of ML classification is measured using the confusion matrix. It is considered to be the basis for measurement of accuracy (Acc), recall (R), precision (P) and F-score (F1). With binary classification (as in this study), Confusion Matrix has four possible cases as in figure 4.13.

		Predicted Class: Negative	Predicted Class: Positive
Actual Class	Negative	True Negative (TN)	False Positive (FP)
	Positive	False Negative (FN)	True Positive (TP)

Figure 4.13. Confusion matrix model

TN: True predicted for negative instance (the instance in actual class is negative and it is predicted as negative).

FN: False predicted for negative instance (the instance in actual class is negative and it is predicted as positive)

FP**:** False predicted for positive instance (the instance in actual class is positive and it is predicted as negative)

TP**:** True predicted for negative instance (the instance in actual class is positive and it is predicted as positive).

For a binary classifier, lists of rates that are often computed from a confusion matrix are:

Accuracy**:** Indicates the correctness of the model.

$$Acc = \frac{TP+TN}{TP+TN+FP+FN} \tag{4.4}$$

Recall: Indicates the sensitivity of the model or how many TPs are returned.

$$R = \frac{TP}{TP+FN} \tag{4.5}$$

Precision**:** Indicates the result of relevancy or how often predicted TP is it correct

$$P = \frac{TP}{TP+FP} \tag{4.6}$$

F Score**:** Indicates a test's accuracy. It considers both the TP rate (recall) and precision to compute the score.

$$F1 = 2 * \frac{P*R}{P+R} \tag{4.7}$$

R, P and F1 are useful measurements of model performance when the classes are very imbalanced (Al-Azani and El-Alfy, 2017; Alnawas and Arıcı, 2019).

5. EXPERIMENTAL SETUP AND RESULTS

The proposed approach for Iraqi Arabic dialect sentiment analysis was evaluated based on a series of experiments set up. To achieve the objective of the study, the experiments were conducted according to the following steps:

(i) The required datasets that matched the proposed method were gathered. In this step, open source data that contained multiple Arabic dialects were sought. These data were collected from their original sources. Iraqi Arabic dialect data were collected from Facebook. Facebook allows collecting comments of pages or users according to privacy policies. Iraqi dialect data were divided into two parts. The first part contained unlabeled data and was used for Doc2Vec training task. The second part contained 2000 comments classified manually by language native experts. Three date sets were obtained as a result of this step: first, a labelled dataset that contained MSA and different dialects. Second, an unlabeled dataset that contained IAD. Third, 2000 labelled sentences that contained IAD.

(ii) Cleaning data from noise and unwanted components. Data available online often contain noise. The preprocessing phase involves cleaning and structuring this data. All three types of datasets were preprocessed in the same tasks.

(iii) Extract features and represent words as vectors. It is difficult to deal with ML algorithms with words directly. Words should be represented in a way that is understandable and easy for ML. Doc2Vec was used to represent words and sentences in multi-dimensional vectors. Doc2Vec is an effective model in the sentiment analysis studies and other areas. The three types of data sets are used in this task. The data are trained together to get the best representation of the words.

(iv) Creating a predictive model of sentiment analysis. Training vectors with four ML algorithms: LR, DT, SVM, and NB. In this task, two types of dataset vectors were used. First, vectors of the pre-labelled dataset. These vectors were used in the training of ML algorithms. Second, vectors of 2000-labelled sentences that contained IAD. These vectors were used in the testing of ML algorithms.

(v) Evaluating the performance of the ML model. The binary class model can be evaluated based on the confusion matrix. Performance evaluation was based on terms: P, R, and F1.

In addition to the main tasks proposed in the methodology, the effect of Doc2Vec variable parameters (Alnawas and Arıcı, 2018b) and background corpora were also studied. Three of Doc2Vec variable parameters were studied. The first and most important parameter is Windows Size (W), which indicates the maximum distance between the current and the expected word within the sentence. The second parameter is Dimension (D) of the vector, which represents the dimensionality of the feature vectors. The third parameter is Negative Samples (NS), which assigns how many "noise words" should be drawn. NS parameter is related to the Softmax function to make it computationally possible. With a large vocabulary size Softmax deals with a small output layer. It contains the correct terms and only a handful of randomly sampled incorrect terms. Those parameters can affect classification performance.

Effects of background datasets on performance were also studied. Each dataset was trained individually with 2000 labelled IAD sentence in Doc2Vec training task.

All experiments were performed using Python 3.6; it provides many libraries for data science such as statistical modeling, data visualization, machine learning, and multi-dimensional array representation. A Jupyter notebook web application was used as an interface to handle the code. Jupyter notebook was accessed using Anaconda navigator desktop GUI.

5.1. Datasets Gathering

Three different datasets were used in this study; publicly available datasets, U-IAD dataset and labelled IAD dataset.

5.1.1. Publicly available datasets

These datasets were used for training task in both Doc2Vec and LM. Six publicly available Arabic datasets were used. The datasets consisted of MSA and Arabic dialects language. The first dataset was "Large Arabic Multi-domain Resources for Sentiment Analysis" (ElSahar

and El-Beltagy, 2015). Five domains were covered in these datasets. Attraction (ATT), Hotel (HTL), Restaurant (RES), Movie (MOV), and Product (PROD) reviews were included.

The second dataset was the Twitter Data set (Abdulla et al., 2013). 2000 tweets were labelled as positive and negative tweets. Multiple topics were covered such as politics and arts. MSA and the Jordanian dialect were used to write these reviews.

The third dataset was OCA (Rushdi-Saleh et al., 2011). 500 comments were collected from different movie blogs (positive and negative) .

The fourth dataset was “Arabic sentiment tweets dataset (ASTD)” (Nabil, Aly, and Atiya, 2015).

The fifth dataset was “Large Scale Arabic Book Reviews (LABR)” (Aly and Atiya, 2013). 51000 book reviews were collected and labelled as positive and negative.

The sixth dataset is “Multi-Perspective Question Answering (MPQA)” (Banea, Mihalcea, and Wiebe, 2010).

Table 5.1. Corpus collections and sources

Dataset	Word count	Unique Word count	Positive Reviews	Negative Reviews
ATT	610 275	15 041	1 939	80
HTL	894 848	80 454	10 049	2 470
RES	295 790	43 022	7 568	2 513
MOV	154 978	36 016	399	135
PROD	35 241	9 975	2 759	786
Twitter	18 404	8 045	1 000	1 000
OCA	130 495	33 964	250	250

Table 5.1. (continued) Corpus collections and sources

ASTD	21 060	10 622	799	1 684
LABR	694 440	96 489	42 832	8 224
MPQA	163 613	24 600	2 718	4 911
Sum	2 253 891	138 620*	70 313	22 053

* The summation of the unique word represents total unique vocabulary.

The datasets are available in CSV format with sentiment polarity labels (positive and negative). There are some differences in the way of representing polarity among the datasets. OCA and Twitter datasets are collocated in two directories. The negative directory contains negative comments and the positive directory contains positive comments. Other datasets are created in one CSV file for each dataset. These files contain positive and negative comments. The CSV files have multi comments separated by (tabs), the structure of files could be like (comments, sentiment, tab). “1” and “0” are used to present the labels of comment’s sentiment (“1” for positive and "0" for negative). "1” was used in some datasets for positive and "-1" for negative to assign labels.

Figure 5.1. Samples of a dataset of comments before unifying the format

In this thesis, a simple python program was used to unify the format of all files. At the end of processing, each dataset file is turned into two files (positive and negative). The datasets files are referred to as raw comments. Further preprocessing is applied to generate clean and unified datasets.

Figure 5.2. Samples of a dataset of comments after unifying the format

Table 5.2. Examples of comments in datasets

Dataset	Positive comments	Negative comments
ATT	المجمع راقي وجميع سبل الترفية والماركات موجودة فيه. /Al-mujammaʿ rāqī wa jamīʿ subul al-tarfīh wa al-mārkāt mawjūda fīh/	الجو حار كثيروالاقسام سيئة جداً والمبنى صغير ولا انصح بالذهاب إليه. /Al-jaww ḥār kathīr wa al-āqsām sayyiʾa jiddan wa al-mabnā ṣaghīr wala ānṣaḥ bi al-dhahāb ilayhu./
	The mall is sophisticated and there are many entertainment and brands.	It is so hot, the sections are very bad, the building is small, and I do not recommend going there.
HTL	الموقع جميل جداً فندق بعيد عن الزحمه وقريب على كل الاماكن السياحيه وسعر معقول. /Al-mawqiʿ jamīl jidan funduq baʿīd ʿan al-zaḥma wa qarīb ʿalā kull al-āmākin al-siyāḥiyya wa siʿir maʿqūl./	في الواقع، إنه لا ينبغي أن يكون حتى فندق. إنه متهالك. /Fiy alwaqiʿ, innahu lā yanbaghy an yakūna ḥattā funduq. Iinnaeu mutahālik./
	The location is very nice and the hotel is far from crowded and close to tourist destinations and a reasonable price.	In fact, it should not even be a hotel. It is worn out.
RES	المنيو وجدته أسعاره في متناول الجميع والطعام جيد جدا. /Al-minyū wajadtahu āsʿārhu fī mutanāwal al-jamīʿ wa alṭaʿām jayyid jiddan./	سيء جداً لا انصح به اطلاقاً، أكل سيء الطعم، طريقه التقديم أسوأ. /sayyiʾ jiddan lā ānṣaḥ bihi iṭlāqan, akil sayyiʾ al-ṭaʿm, ṭarīqa al-taqdīm aswāʾ./
	Affordable prices and a very good food.	Very bad, I do not recommend it at all, tasteless food, The service worse.
MOV	موسيقى هانز زيمر ممتازة و مناسبة لأجواء الفيلم. /Mūsīqā Hanz Zymr mumtāza wa munāsiba li āajwaʾ al-fīlm./	اخراج سئ وتمثيل مبالغ فيه من أسوأ الافلام الي شفتهم في حياتي. /Iikhrāj sayyiʾ wa tamthīl mubālagh fīhi min aswaʾ al-āflām illy shifithum fī ḥayātī./
	Hans Zimmer's music is good and suitable for the film.	Bad directing, bad acting, one of the worst films I have ever seen.
PROD	تم توصيل السلعة الى المنزل بحالة ممتازه في وقت قصير /Tama tawṣil al-silʿa alā al-manzil fiy waqt qaṣiyr/	المنتج غير جيدة ولا انصح به /Al-muntaj ghēr jayyida walā anṣaḥ bihi/
	The item was delivered to the home in excellent condition in a short time.	The product is not good and I would not recommend it.

Table 5.2. (contained) Examples of comments in datasets

Dataset	Positive comments	Negative comments
Twitter	كلمات رائعه و حلوه و لا احلى ان يكونو في النهايه رائعون. /Kalimāt rāʾiʿa wa ḥilwa wa la āḥlā an yakūnū fī al-nihāya raʾiʿūn./	برنامج فاشل جدا. /Barnamij fāshil jiddan./
	Wonderful words and sweet and it is great to be in the end a wonderful.	A very unsuccessful TV program.
OCA	الفلم رائع من حيث احداثه واساليب تصويره. /Al-film rāʾiʿ min ḥayth aḥdāthihi wa āsālīb taṣwīrihi./	اسوء افلام الموسم. /aswaʾ āflām al-mawsim./
	The film is wonderful in terms of events and methods of photography.	The worst movies of the season.
ASTD	فكرة المقالة حلوة وأسلوبك في توصيل الفكرة حلو. /Fikrat al-maqāla ḥulwa wa uslūbak fī tawṣīl al-fikra ḥulw/	الراتب مايكفي الحاجة. /Al-rātib mā yakfī al-ḥāja/
	The idea of the article is good and your writing style to deliver the idea is good	Salary is not enough.
LABR	رواية رائعة بما تحمله الكلمة من معنى . مراد تحسن اسلوبه كثيرا بعد فيرتيجو. /Riwāya rāʾiʿa bima taḥmilahu al-kalima min maʿnā. murād taḥasana āisluwbahu kathiran baʿda firtijū./	كتاب سي جدا .الاسلوب غير ممتع ,نهاية مفتوحة والكتاب بوجهة عام كئيب. /Kitab sayyiʾ jiddan. Al-āslūb ghēyr mumtiʿ, nihāya maftūwḥa wa al-kitāb biwajh ʿām kaʾīb./
	A wonderful novel, Murat improved his style a lot after Vertigo.	A very bad book. The writing style is not interesting and gloomy.
MPQA	الحكومة الاندونيسية قد اتخذ خطوات جادة لضمان السلامة الشخصية للمستثمرين. /Al-ḥukuma al-andūnīsiyya qad ittakhadh khaṭawāt jādda li ḍamān al-salāma al-shakhṣiyya lilmustathmirīn./	ان الوضع هو الأسوأ في زيمبابوي في الجنوب. /Inna al-waḍʿ huwa al-āswaʾ fī Zambābwī fī al-jenūb./
	The Indonesian government has taken serious steps to ensure the personal safety of investors.	The situation is the worst in the south of Zimbabwe.

5.1.2. Unlabeled IAD dataset

Doc2Vec embedding layer needs big data to train. For this task, more than 250000 comments were fetched from Facebook. These comments are publicly available and fetched using Facepager (Jünger and Keyling, 2018). The comments are collected from six pages that cover different topics such as home appliances company, Airways Company, News, Restaurant, Sport, and Communications Company. The dataset contains 19000 unique words.

5.1.3. Labelled IAD dataset

According to previous studies, there is a limitation in Arabic dialects resources. IAD has not been considered in previous studies. To overcome this limitation, 2000 comments were labelled manually in this thesis. Facebook was used to fetch comments using Facepager. The Arabic dialect used in selected pages is IAD. These pages cover different domains.

Three native experts tagged these comments manually. The sentiment classification in this thesis is binary, for that, the comments are classified as positive or negative comments. These comments are trained based on Doc2Vec model to generate the corresponding vectors. The vectors are used in ML test tasks to evaluate the proposed sentiments prediction model. Table 5.3 explains IAD dataset

Table 5.3. Negative and positive comments in each domain for test data

Facebook pages	Word count	Unique Word count	Positive comments	Negative comments
Home appliances company	4 912	2 090	306	170
Airways company	2 795	1 625	133	155
News page	1 754	1 189	71	170
Restaurant	1 743	884	216	16
Sport	1 675	972	145	147
Communications company	6 494	2886	129	342
Sum	19 373	7 436*	1 000	1 000

* The summation of the unique word represents total unique vocabulary.

Table 5.4. Examples of comments in labelled IAD dataset

Topic	Positive comments	Negative comments
Home appliances company	اني عندي جهازين وبرودة تخبل, شركه راقيه. /Āny ʿindī jihazēin wa brūda tkhabbil, sharika rāqya./	اتعس سبالت تحب اعطال ومجربها عندي واحد شوفني نجوم الظهر. /ātʿas sabālit tḥibb aʿṭāl wa mjarribhā ʿindī wāḥid shawwafnī njūm al-ẓuhr./
	I have two devices Good cooling Good company.	Bad cooling devices always break down. I tried it, it's show me stars in the daytime
Airways company	صدقة لالله اموت عالخطوط العراقية تخبل. /ṣadaqa li Allāh amūt ʿal-khuṭūṭ al-ʿirāqiyya tkhabbil./	محسيت لا بالامان ولا شفت الاكل الطيب. maḥasēt lā bi al-āmān wa lā shift al-akl al-ṭaiyyib./
	I love Iraqi Airways.	I did not feel safe and did not see the tasty food.
News page	خوش معلومات. /Khūsh maʿlūmāt/	التعينات مئيوس منها من زمان مو بس السنه الجايه. /Al-taʿyīnāt maʾyūs minhā min zaman mū bass al-sana al-jāya./
	A good information.	Getting a job is difficult at all times.
Restaurant	اتخبل الدولمة الي عدكم وآكلها دائما بمطعمكم. /Itkhabbal al-dūlma illy ʿudkum wa ākilhā dāʾiman bmaṭʿamkum./	العصير مو فريش صراحه معجبني. /Al-ʿaṣiyr mū frēsh ṣarāḥa maʿijabnī./
	The "Dolma" is very delicious and I always eat it at your restaurant.	The juice was not fresh and I did not like it.
Sport	فدوه لله شكد عمالقه. /Fidwa lāllah shəkad ʿamāliqa./	والله هذا المدرب مو شي للئسف. /WaAllāh hadhā al-mudarib mū shī lilasaf./
	They are supernatural.	Indeed, the coach is not good.
Communications company	والله خوش خطوه عاشت اديكم بس كون تخلوها مو تلغوها. /WaAllāh khōsh khuṭwa ʿāshat ādykum bas kūn tkhallūhā mū tlghūhā./	اني عندي جهاز بس النت بي زفت حتى صفحة مال فيس بوك ميفتح شسويله. /Āni ʿindiy jihaz bas al-nat bi zifit ḥattā ṣafḥat mal fays būk mayiftaḥ shasawiwīla./
	It's a good service and I hope it will not be cancelled.	I have a device but the internet is very weak and does not respond to Facebook.

5.2. Datasets Preprocessing

Formatting sentences and words are necessary before generating word embedding. Although the texts were carefully prepared in previous studies, some irregularities were noticed in the texts such as the punctuation marks, repetition of letters in the words and some non-Arabic characters. Therefore, further preprocessing was performed on the full-text of datasets. Python's Natural Language Toolkit (NLTK) was used as a tool for preprocessing. NLTK provides perfect assist in dealing with texts in encoding issues. The preprocessing was accomplished in the following steps:

- Delete non-Arabic words
  The objectives of the thesis consider the Arabic language, therefore contents of datasets that contained other languages were ignored.
- Delete stop words
  In NLP tasks, the stop words refer to words that are frequently used in sentences. Removing of these words from sentences did not alter its meaning.
- Delete punctuations
- Delete numbers
- Delete short vowels

Figures 5.3 and 5.4 show the difference before and after applies preprocessing.

```
د بهذا الفيلم ..يقدم فيلم مثل هذا .. عموما الفيلم يتكلم بشكل سخيف عن حرب تندلع بسبب امراة ليست امراة بالشكل الذي نعرف بل ان الزعيم تشاوتشاو قد اعجب بها منذ مدة طويله فترحل وتتزوج من نائب قائد في قرية مجاوره فيقرر ان يغزيها ليسترجعها Paycheck
 The Magnificent Seven ومثل ڤون في The Great Escape
Alien حين أخرج ريدلي سكوت
Back to the Future Part II
"Brain Donors, Big Daddy, Happy Gilmore , You Don't Mess with the Zohan"
Butch Cassidy and the Sundance Kid بوتش كاسيدي وسندانس كِد
CAST & CREDITS
Chaos Reign الفوضى تحكم
"Cop Land (1997) , Girl, Interrupted (1999), Walk the Line (2005)"
David McCallum
Dazed And Confused
Despair العنوان الثالث: يأس
Dramatic build-up
Eagle Eye Shot كتاب السيناريو
Envy الحسد
Genocide وبجانبها بين هلالين: إبادة جماعية
Get Out of here: لكل ذلك وجدت نفسي أصرخ
Gluttony النهم
Greed is Good نفسه: نعم قال عبارته الشهيرة تلك
Greed الجشع
Green Zone
Grief العنوان الأول هو: أسى
"Hitch النجاح لتتحول من مجرد لأفلام تسلية إلى دراسة شديدة التميز و الجدية لعلاقات الأمريكين العاطفية كما فى أفلام المخرجة و كاتبة السيناريو نورا إيفرون ، أو بإمكان هذه النوعية أيضاً أن تصبح أفضل وسيلة لإجهاض أفكار رائعة مثلما كان الحال مع أفلام مثل
"Hostel, Hostel II, Saw, Saw II"
Hot Fuzz يستدعي مثل هذه النقلا. المخرج الذي لديه فيلم رديء سابقا هو
http://www.mzyondubai.com/vb/showthread.php?t=97790اسم الفلم : مجنون أميرة
http://www.mzyondubai.com/vb/showthread.php?t=97790اسم الفلم : مجنون أميرة
http://www.mzyondubai.com/vb/showthread.php?t=97790اسم الفلم : مجنون أميرة
http://www.mzyondubai.com/vb/showthread.php?t=97790اسم الفلم : مجنون أميرة
http://www.mzyondubai.com/vb/showthread.php?t=97790اسم الفلم : مجنون أميرة
http://www.mzyondubai.com/vb/showthread.php?t=97790اسم الفلم : مجنون أميرة
http://www.mzyondubai.com/vb/showthread.php?t=97790اسم الفلم : مجنون أميرة
http://www.mzyondubai.com/vb/showthread.php?t=97790اسم الفلم : مجنون أميرة
http://www.mzyondubai.com/vb/showthread.php?t=97790اسم الفلم : مجنون أميرة
```

Figure 5.3. An example of a text dataset before preprocessing is applied.

```
ocaneg.txt - Notepad
File Edit Format View Help
 ابهر بشكل جيد بهذا فيلم يقدم فيلم مثل عمو فيلم يتكلم بشكل سخيف حرب تندلع بسب امراة امراة شكل نعرف زعيم تشاوتشاو اعجب مدة طويله فترحل وتزوج ناءب قاءد قرية مجاوره فيقر يغزي ليسترجع
ومثل ڤون
اخرج ريدلي سكوت
بوتش كاسيدي وسندانس كد
فوضي تحكم
عنو ثالث ياس
كتاب سيناريو
حسد
وبجانب هلال ابادة جماعية
لكل وجدت نفسي اصرخ
نهم
نفسه قال عبارته شهيرة
جشع
عنو اول اسي
، شديدة نجاح لتحول مجرد لافلام تسلية دراسة شديدة تميز و جدية لعلاق امريك عاطفية افلام مخرجة و كاتبة سيناريو نورا ايفر ، بامك نوعية ايضا تصبح افضل وسيلة لاجهاض افكار راءعة مثل حال افلام مثل ، و هيتش يمثل افضل عملية اجهاض قدمت هوليود سنو اخيرة
يستدعي مثل نقلا مخرج لديه فيلم رديء سابقا
اسم فلم مجن اميرة
اسم فلم مجن اميرة
اسم فلم مجن اميرة
اسم فلم مجن اميرة
اسم فلم مجن اميرة
اسم فلم مجن اميرة
اسم فلم مجن اميرة
اسم فلم مجن اميرة
اسم فلم مجن اميرة
اسم فلم مجن اميرة
اسم فلم مجن اميرة
اسم فلم مجن اميرة
اسم فلم مجن اميرة
اسم فلم مجن اميرة
اسم فلم حسن ومرقص
اخراج يان لومنيكي
اخراج جون فرانكنهايمر
اخرج عديد افلام متازه راقيه مثل لنيكولاس كيج و جون ترافولتا ويعد افضل افلام اكشن وايضا مهمه مستحيله لتوم كروز وايضا انسي انذكر فيلم خيال علمي
```

Figure 5.4. An example of a text dataset after apply preprocessing.

5.3. Building Word Embedding Models

Doc2vec is an unsupervised approach used to learn the document representation. The input of texts per document can vary while the output is fixed-length vectors. Sentence vectors are unique among all documents while word vectors are shared among all documents such that word vectors can be learned from a different document. During the training phase, word vectors are trained and the paragraph vectors are discarded following the process. During the prediction phase, paragraph vectors are initialized randomly and computed by word vectors.

In this thesis, python Gensim is used. It is capable of applying and implementing Doc2Vec. Doc2Vec model aggregates all the words in sentences into vectors. Every sentence is formatted as:

[['w_1', 'w_2', 'w_3',…, 'w_n'], ['$label_1$']]

LabeledLineSentence class in Doc2Vec model provides a useful process to do this task.
In simple words, LabeledLineSentence stores two things; a list of words and a label. The default constructor of LabeledLineSentence class can perform that for a single text file. In sentiment analysis tasks proposed, there are multiple files. Each file represents text data for a different task (training, testing, positive, negative etc.). Qiu (2015), modified the default constructor to avoid this limitation. The modified constructor defines the label prefixes of sentences based on the file name. Via the iterator, others can be read directly. For example, the sentences in the train positive file would be as follows:

[['w_1', 'w_2', 'w_3',…, 'w_n'], ['TRAIN_POS_1']]
[['w_1', 'w_2', 'w_3',…, 'w_n'], ['TRAIN_POS_2']]
.
.
.
[['w_1', 'w_2', 'w_3',…, 'w_n'], [' TRAIN_POS_m']]

Vocabulary table is required when deals with Doc2Vec. To build vocabulary table, all the words will be digested, filtered, and counted. The vocabulary table contains unique words.

In each training epoch, the sequence of sentences fed to the model is randomized for a better-trained model. Therefore this step is important to get better results. The data are trained with 100 epochs. Multi-dimensional vectors are presented using NumPy array. It is an efficient multi-dimensional container of generic data.

To measure the quality of embedding, multiple similarity queries are created to see how reasonable the embedding is. Table 5.5 shows examples of sentiment-related results of the embedding.

The sentiment classification model was considered as a binary classification. The vectors were trained with four typical classifiers: LR, DT, SVM and Bernoulli NB. All of the classifiers were run under the same training conditions. In term of N-fold cross-validation, our datasets were already split into training and testing datasets, for that no cross-validation was used. Learning library of Scikit-learn machine based on Python was used. Scikit-learn library includes the functionality of regression and classification and others. It is designed to interoperate with the numerical and scientific Python libraries such as NumPy and SciPy.

Table 5.5. Sample word similarity results for sentiment-related vocabulary

Query term (transliteration)		Top 5 results (transliteration)			Distance
Beautiful	حلو /ḥulw/	Beautiful Beautiful Magnificent Beautiful Excellent	/ḥulwa/ /ḥulwa/ /rawʿā/ /jamīl/ /mumtāz/	حلوه حلوة روعة جميل ممتاز	0,77 0,74 0,70 0,68 0,66
Thief	حرامي /ḥarāmī/	Thief Steal Terrorist Thief Steal	/Al-ḥarāmi/ /ybūk/ /irhābī/ /sāriq/ /ybūg/	الحرامي يبوك ارهابي سارق يبو گ	0,56 0,55 0,51 0,49 0,49

The predicted model is evaluated based on P, R, and F1. Scikit-learn library was used to generate a classification report. P, R, and F1 can be obtained from Scikit -learn library. The results derived from experimental studies are reviewed. The effect of Doc2Vec parameters (W, D, and N) is investigated. In addition, the effect of background training corpora is inspected.

5.4. Results Based on Doc2Vec Parameters

In this subsection, the effect of parameters (W, D, and NS) is addressed for IAD sentiment analysis task. Publicly available datasets were used in the training task and IAD for testing task. All the experiments in this section were run under same conditions.

5.4.1. The effect of window size and dimensionality

For the sentiment analysis task of IAD, three values of window size (W= 1, 2, 3) and three values of a dimension of the vector (D= 50, 100, 200) were used to generate word vectors. Tables 5.6-5.8 show the results of classifiers using word embedding as a feature selection technique. The representations of continuous word vectors were generated using PV-DBOW architecture. The tables show the results of windows sizes (W= 1, 2, 3) along each dimension sizes of (D= 50, 100, 200). The best overall scores were highlighted in bold.

In Table 5.6, three values of windows (W= 1, 2, 3) with dimension sizes of (D= 50) were used.

Actually, four ML algorithms were used (e.g. LR, DT, SVM, and NB) to show the difference in performances over these values of Doc2Vec parameters. SVM demonstrated the best performance compared to other classifiers.

With the window size of (W=1), SVM yields the results of preferences (P=0,81, R=0,76, F1=0,75). When the window size was fixed to (W=2), the performance of SVM reduced in terms of (P, R, F1), DT reduced in (R, F1). With window size (W=2), the performance of NB increased in (P, R, F1), as LR increased in (R, F1).

When the window size was fixed to (W=3). The performance of SVM and LR reduced in terms of (R, F1) and in terms of (P, R, F1) for NB. The performance of DT was increased in terms of (P, R, F1).

Table 5.6. Results of classifiers (W= 1, 2, 3, D= 50)

	LR			DT			SVM			NB		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
W=1	0,79	0,74	0,73	0,62	0,62	0,61	0,81	0,76	0,75	0,66	0,62	0,60
W=2	0,79	0,75	0,74	0,62	0,61	0,60	0,79	0,72	0,71	0,68	0,64	0,61
W=3	0,79	0,74	0,73	0,63	0,62	0,62	0,79	0,71	0,68	0,65	0,59	0,54

Table 5.7 shows three values of window sizes (W= 1, 2, 3) with a dimension size of (D= 100). The best performance was obtained by SVM in term of P and LR in terms of (R, F1) with windows size (W=1). When the window size was fixed to (W=2), the performance of classifiers LR, DT, SVM and NB were reduced in terms of (P, R, and F1). With window size (W=3), the performance of classifier LR increased in terms of (P, R, F1), SVM was increased in term of P, and NB was increased in terms of (R and F1).

Table 5.7. Results of classifiers (W= 1, 2, 3, D= 100)

	LR			DT			SVM			NB		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
W=1	0,79	0,74	0,73	0,64	0,63	0,62	0,80	0,73	0,72	0,67	0,64	0,62
W=2	0,77	0,72	0,71	0,60	0,59	0,59	0,78	0,69	0,67	0,66	0,61	0,58
W=3	0,78	0,73	0,72	0,59	0,58	0,58	0,79	0,69	0,66	0,64	0,62	0,60

In Table 5.8, three values of window sizes (W= 1, 2, 3) with a dimension size of (D= 200) are used. The performance of four ML algorithms (e.g. LR, DT, SVM, and NB) was studied

over these values of Doc2Vec parameters. With these settings, LR yielded the best performance when windows size (W=1). When the window size was fixed to (W=2), the performance of classifiers LR, DT, SVM and NB were reduced in terms of (P, R, and F1). If window size was fixed to (W=3), the performance of classifiers DT and NB was increased in terms of (R and F1), and SVM was decreased in terms of (R and F1).

Table 5.8. Results of classifiers (W= 1, 2, 3, D= 200)

	LR			DT			SVM			NB		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
W=1	0,80	0,77	0,76	0,61	0,61	0,60	0,80	0,73	0,71	0,71	0,66	0,65
W=2	0,79	0,75	0,74	0,58	0,58	0,57	0,79	0,69	0,66	0,67	0,64	0,63
W=3	0,79	0,75	0,74	0,60	0,60	0,59	0,79	0,68	0,65	0,67	0,65	0,64

Based on Tables 5.6-5.8, the best performance for LR was evaluated using windows size and dimension (W=1 and D=200) in terms of (P=0,80, R=0,77, F1=0,76). For DT, the best result is obtained with (W=3 and D=50) in terms of (P=0,63, R=0,62, F1=0,62).

When windows size and dimension (W=1 and D=50) are used, SVM yielded the best results in terms of (P=0,81, R=0,76, F1=0,75). NB achieved the best results with windows size and dimension (W=1 and D=200) in terms of (P=0,71, R=0,66, F1=0,65).

5.4.2. The effect of negative samples

Negative sample (NS) is another important parameter of word embedding that may have an effect on the classification performance. In word embedding models, NS is used to define the number of negative samples that are randomly selected for each data. By using NS, the word embedding model can distinguish the correct word relationships from noise. The literature studies that used word embedding model for English such as Mikolov et al. (2013a) observed the small number of NS is suitable for a large training corpus. For small training corpora, a large number of NS can be useful. For sentiment analysis problem, Doc2Vec model is trained using NS= (7, 10, 20, and 30). Here, the settings of window size and

dimension size parameters that obtain the best results of classifier performance (Tables 5.6-5.8) were used for experiments in this section.

Figure 5.5 shows the performance of the LR classifier. The best result was obtained using NS= 30 in terms of (P=0,81, R=0,78, F1=0,77). When NS=10 and NS=20, the results were equal to (P=0,79, R=0,76, F1=0,76). NS=7 obtained slightly better performances on P and R, the results was (P=0,80, R=0,77, F1=0,77).

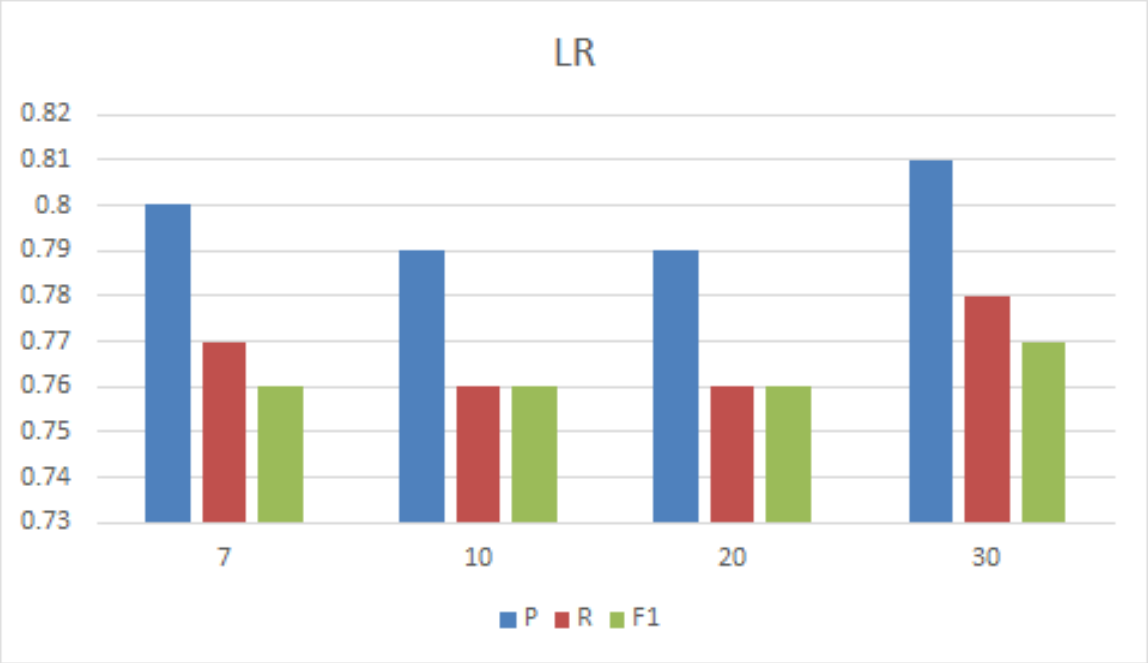

Figure 5.5. LR classifier scores using different Negative Sample sizes

The experiment results of DT classifier are shown in Figure 5.6. The best result was obtained using NS=10 in terms of (P=0,66, R=0,65, F1=0,65). The other values of NS (NS=7, NS=20 and NS=30) yielded equal results in terms of (P=0,62, R=0,61, and F1=0,60).

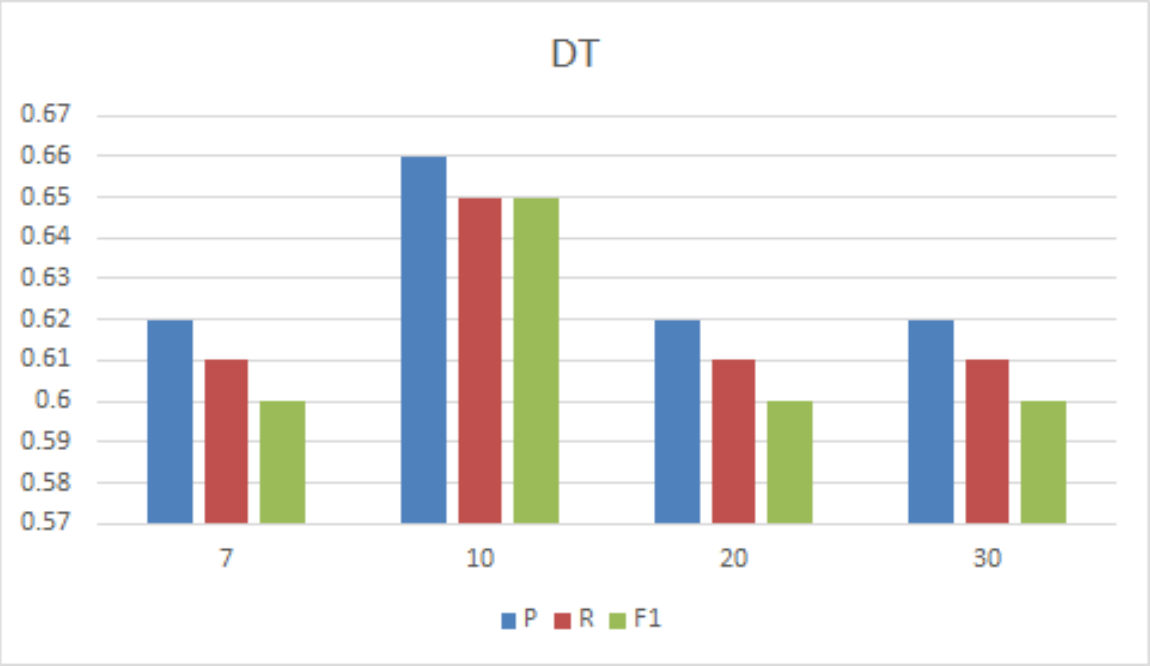

Figure 5.6. DT classifier scores using different Negative Sample sizes

SVM classifier yielded high results using NS=30 in terms of (P=0,82, R=0,79, F1=0,78) as shown in figure 5.7. Using NS=10 and NS=20 the result in terms of R and F1 were equal to (R=0,77, F1=0,76). The results were P=0,81 with NS=10 and P=0,80 with NS=20.

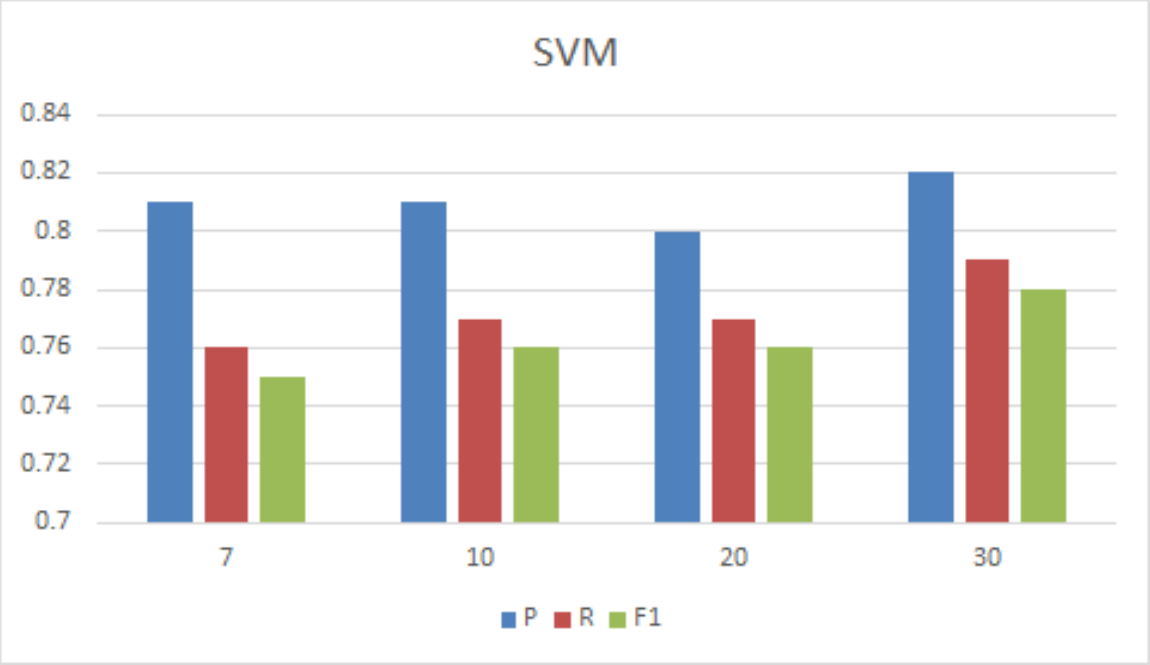

Figure 5.7. SVM classifier scores using different Negative Sample sizes

Figure 5.8 shows the performance of the NB classifier. The best result was obtained using NS= 7 in terms of (P=0,71, R=0,66, F1=0,65). When NS=10 and NS=20, the classifier performance in terms of P and R were equal to (R=0,66, F1=0,64). The classifier performance using NS= 30 was (P=0,63, R=0,62, F1=0,61).

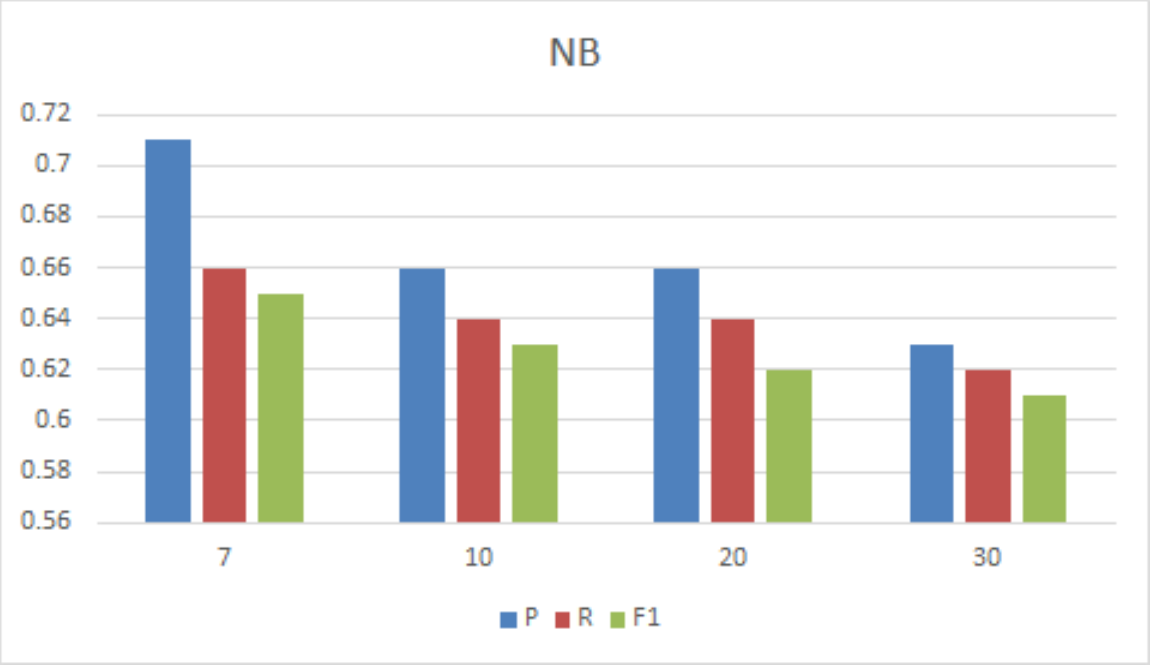

Figure 5.8. NB classifier scores using different Negative Sample sizes

5.5. The effect of Background Corpora

As mentioned in previous sections of this study, the Arabic language has many dialects. There is a convergence between some dialects. In this study, different corpora are used in Doc2Vec training task. The type of background corpus could affect the classification performance. To study the effect, two training Doc2Vec cases were studied. In the first case, six publicly available Arabic datasets were used as the corpus. In addition, U-IAD was used in Doc2Vec training task. For the second case, six publicly available Arabic datasets were used several times. In addition, U-IAD was used in Doc2Vec training task. The results of the two cases were used in ML tasks.

Table 5.9 shows the differences in performance of classifiers based on the first case of training Doc2Vec. The publicly available Arabic datasets were collected together to

establish a corpus. The corpus contained positive and negative comments. U-IAD was used in this task. Labelled IAD was also used in this task.

Table 5.9. The performance of classifiers using U-IAD in Doc2vec training task

Classifiers	Training model	P	R	F1	TP*	TN**
LR	Without U-IAD	0,71	0,68	0,66	710	665
	With U-IAD	0,81	0,78	0,77	810	771
DT	Without U-IAD	0,60	0,60	0,60	600	600
	With U-IAD	0,66	0,65	0,65	660	640
SVM	Without U-IAD	0,77	0,74	0,74	770	729
	With U-IAD	0,82	0,79	0,78	820	782
NB	Without U-IAD	0,68	0,63	0,62	680	600
	With U-IAD	0,71	0,66	0,65	710	634

* True Positive (total positive comment=1000)
** True Negative (total negative comment=1000)

Table 5.9 clearly shows the increase in performance of the classification of unlabeled IAD. This increase was due to the representation of words more closely in the vector space. The process of training with unlabeled IAD makes the representation of words belonging to other dialects easier. The words belonging to other dialects that have the closest meaning to the words in IAD were found based on the repetition of words in sentences. Since the numbers of sentences were large, the process for finding closest words achieved results that are more accurate.

Table 5.10 shows the performance of classifiers based on the second case of training Doc2Vec. U-IAD used as background corpus for Doc2vec training task. Labelled IAD used for training and testing ML model tasks.

Table 5.11 shows the differences in performance of classifiers based on the second case of training Doc2Vec. The publicly available Arabic datasets were used as background corpora in Doc2vec training task. Labelled IAD was also used as test dataset in this task.

Table 5.10. The performance of classifiers with/without U-IAD as background corpus

Dataset	LR			DT			SVM			NB		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
U-IAD	0,90	0,90	0,90	0,76	0,75	0,75	0,90	0,90	0,90	0,88	0,88	0,88
Without training datasets	0,85	0,84	0,85	0,72	0,71	0,72	0,85	0,84	0,84	0,83	0,82	0,83

Table 5.11. The performance of classifiers using the publicly available Arabic datasets as background corpora in Doc2vec training task.

Dataset	LR			DT			SVM			NB		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
ATT	0,64	0,59	0,55	0,67	0,52	0,39	0,64	0,59	0,55	0,64	0,61	0,58
HTL	0,71	0,74	0,73	0,62	0,58	0,60	0,80	0,75	0,77	0,72	0,74	0,73
RES	0,74	0,68	0,66	0,60	0,58	0,57	0,74	0,68	0,65	0,71	0,69	0,69
MOV	0,60	0,66	0,65	0,60	0,66	0,65	0,58	0,62	0,62	0,63	0,63	0,63
PROD	0,59	0,61	0,61	0,58	0,66	0,64	0,65	0,62	0,63	0,62	0,64	0,58
TWITTER	0,69	0,68	0,68	0,56	0,56	0,56	0,66	0,66	0,66	0,69	0,66	0,65
OCA	0,54	0,51	0,41	0,43	0,44	0,40	0,51	0,50	0,44	0,50	0,50	0,44
ASTD	0,70	0,61	0,56	0,62	0,61	0,60	0,69	0,62	0,57	0,68	0,59	0,53
LABR	0,65	0,62	0,63	0,71	0,74	0,73	0,59	0,61	0,61	0,60	0,66	0,65
MPQA	0,52	0,52	0,50	0,55	0,55	0,55	0,52	0,52	0,49	0,57	0,56	0,54

Tables 5.10 and 5.11 clearly show that the best performance of the classifiers was obtained using U-IAD as background corpus in Doc2Vec training task. The table also shows satisfying results without using any training datasets in Doc2Vec training task.

Also observed in Table 6.5, when some datasets were used as background in training task it achieved poor results. Poor results were due to the fact that the representation of words and sentences are far from each other. This indicates that the language used in these dataset differed from IAD.

6. CONCLUSION AND DISCUSSION

In this doctoral thesis, the sentiment analysis of Iraqi Arabic dialect was presented. Up to our knowledge, this is the first study that dealt with the Iraqi Arabic dialect. Through this study, the gaps in previous studies that discussed the Arabic language were determined. The framework of the thesis was designed as follows: i) Collecting datasets to build a corpus. The datasets consisted of three types; publicly available datasets, unlabeled-IAD and labelled IAD. ii) All dataset types were preprocessed to delete the noise and standardize the text format to be understood by the machine learning algorithms. iii) The Doc2Vec approach was used to extract features from text and represent words as vectors. This approach yielded a better performance compared to the approaches used in previous studies. Using Doc2Vec, words are represented by context of the text. In other words, a word that has more than one meaning is represented as different as the number of meanings it has. iv) Training vectors by machine learning algorithms to create a predictive model of Iraqi Arabic sentiment analysis. Four algorithms that are widely used in sentiment analysis were used in this thesis. v) Evaluating the sentiment predictive model based on precision, recall, and F1 score.

In addition, effects of three Doc2Vec variable parameters were also investigated, which are Windows Size context, Dimension of the vector, and Negative Samples. Effects of background datasets on the performance of the classifiers were studied. Each dataset was trained individually with 2000 labelled sentences of Iraqi Arabic dialect in Doc2Vec training task.

Six publicly available Arabic datasets were used. These datasets were labelled as positive and negative labels. First dataset was taken from (ElSahar and El-Beltagy, 2015) and consisted of five groups which are; Attraction, Hotel, Restaurant, Movie, and Product reviews. The second dataset was a Twitter Dataset taken from (Abdulla et al., 2013). The third dataset was OCA taken from (Rushdi-Saleh et al., 2011). The fourth dataset was “Arabic sentiment tweets dataset from (ASTD)” taken from (Nabil et al., 2015). The fifth dataset was “Large Scale Arabic Book Reviews (LABR)” taken from (Aly and Atiya, 2013). The sixth dataset was “Multi-Perspective Question Answering (MPQA)” taken from (Banea et al., 2010). These six datasets were used to improve the representation of positive and

negative words and sentences in doc2vec training task. Also, the vectors of these datasets were used to create a prediction model in the ML training task.

More than 250 000 comments in Iraqi Arabic dialect from Facebook were used. These comments are publicly available and fetched using Facepager. The comments were collected from six Iraqi web pages that cover different fields such as a home appliances company, an airways company, a news page, a restaurant, sports, and a communications company. The dataset contained 19 000 unique words. This dataset was used for Doc2Vec embedding layer training task.

Three native experts labelled manually 2000 comments that had a content in Iraqi Arabic dialect. Facepager was used to fetch the comments from Facebook. These comments were trained based on Doc2Vec model to generate the vectors. The vectors were used in machine learning test tasks to evaluate the sentiment prediction model proposed.

The main findings achieved by the results our experimental studies show that the best performance for the classification model proposed was achieved using the SVM classifier, which yielded a precision value of 82%, recall value of 79%, and an F-score value of 78%. The results were based on training six publicly available Arabic datasets and an un-labelled Iraqi Arabic dialect dataset.

The performance results of the classifiers based on variable parameters showed that the best results were obtained using windows size equaling to 1 for all classifiers. The best results were obtained using the dimension of the vector equaling to 50 for the SVM classifier, 100 for DT classifier, and 200 for LR and NB classifiers. The values of the Negative Samples that yielded satisfying results were equal to 7 for NB classifier, 10 for DT classifier, and 30 for SVM and LR classifiers.

Results obtained from training datasets individually showed that the best result was obtained using an un-labelled Iraqi Arabic dialect dataset. LR and SVM yielded a precision, recall, and F1 value of 90%.

For the future works, it is recommended that an Iraqi Arabic dialect corpus containing more vocabulary words and more than two classes be established. It is also recommended that more experiments be carried out using other approaches such as CNN or LSTM.

REFERENCES

Abdul-Mageed, M. (2017). Modeling Arabic subjectivity and sentiment in lexical space. *Information Processing & Management, 56*(2), 17.

Abdul-Mageed, M. and Diab, M. T. (2012). *AWATIF: A Multi-Genre Corpus for Modern Standard Arabic Subjectivity and Sentiment Analysis.* Paper presented at the The eighth international conference on Language Resources and Evaluation (LREC), Istanbul,Turkey.

Abdulla, N. A., Ahmed, N. A., Shehab, M. A., and Al-Ayyoub, M. (2013). *Arabic sentiment analysis: Lexicon-based and corpus-based.* Paper presented at the 2013 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT), Amman, Jordan.

Abu-Errub, A., Odeh, A., Shambour, Q., and Hassan, O. A.-H. (2014). Arabic roots extraction using morphological analysis. *International Journal of Computer Science Issues (IJCSI), 11*(2), 128.

Abuata, B. and Al-Omari, A. (2015). A rule-based stemmer for Arabic Gulf dialect. *Journal of King Saud University-Computer and Information Sciences, 27*(2), 104-112.

Acat, Y. (2015). *Dirasa muqārana fy al-ʿnāsir al-mushtaraka fy al-lahjāt al-ʿarabiyya alānāḍwlya almuʿasira [A Comparative Study of the Common Elements of Contemporary Anatolian Arab Dialects]*. Paper presented at the 11th International Conference of AIDA (Arabic varieties – Far and wide), Bucharest.

Al-Ayyoub, M., Khamaiseh, A. A., Jararweh, Y., and Al-Kabi, M. N. (2019). A comprehensive survey of arabic sentiment analysis. *Information Processing & Management, 56*(2), 320-342.

Al-Azani, S. and El-Alfy, E.-S. M. (2017). *Using word embedding and ensemble learning for highly imbalanced data sentiment analysis in short arabic text.* Paper presented at the The 8th International Conference on Ambient Systems, Networks and Technologies, ANT 2017, Madeira, Portugal.

Al-Bazi, M. P. K. (2005). *Iraqi Dialect Versus Standard Arabic*. United States: MATFL, 23-35.

Al-Rfou, R., Kulkarni, V., Perozzi, B., and Skiena, S. (2015). *Polyglot-NER: Massive multilingual named entity recognition.* Paper presented at the Proceedings of the 2015 SIAM International Conference on Data Mining, British Columbia, Canada.

Al-Rubaiee, H., Qiu, R., and Li, D. (2016). *Identifying Mubasher software products through sentiment analysis of Arabic tweets.* Paper presented at the 2016 International Conference on Industrial Informatics and Computer Systems (CIICS), Sharjah, United Arab Emirates.

Al-Sallab, A., Baly, R., Hajj, H., Shaban, K. B., El-Hajj, W., and Badaro, G. (2017). AROMA: A recursive deep learning model for opinion mining in arabic as a low resource language. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 16*(4), 25.

Al-Sharkawi, M. (2016). *History and Development of the Arabic Language*. UK: Routledge, 131-207.

AL-Smadi, M., Al-Ayyoub, M., Al-Sarhan, H., and Jararweh, Y. (2015a). *Using aspect-based sentiment analysis to evaluate arabic news affect on readers.* Paper presented at the 2015 IEEE/ACM 8th International Conference on Utility and Cloud Computing (UCC), Limassol, Cyprus.

Al-Smadi, M., Qawasmeh, O., Talafha, B., and Quwaider, M. (2015b). *Human annotated arabic dataset of book reviews for aspect based sentiment analysis.* Paper presented at the 3rd International Conference on Future Internet of Things and Cloud (FiCloud), Rome, Italy.

Alayba, A. M., Palade, V., England, M., and Iqbal, R. (2018). *Improving Sentiment Analysis in Arabic Using Word Representation.* Paper presented at the 2nd International Workshop on Arabic Script Analysis and Recognition, London, UK.

Alhumoud, S. O., Altuwaijri, M. I., Albuhairi, T. M., and Alohaideb, W. M. (2015). Survey on arabic sentiment analysis in twitter. *International Journal of Social, Behavioral, Educational, Economic, Business and Industrial Engineering, 9*(1), 364-368.

Alnawas, A. and Arıcı, N. (2018a). The Corpus Based Approach to Sentiment Analysis in Modern Standard Arabic and Arabic Dialects: A Literature Review. *Journal of Polytechnic, 21*(2), 461-470.

Alnawas, A. and Arıcı, N. (2018b). *Effect of word embedding variable parameters on Arabic sentiment analysis performance [abstract]*. Paper presented at the 5th International Conference on Computational and Experimental Science and Engineering, Antalya, Turkey.

Alnawas, A. and Arıcı, N. (2019). Sentiment Analysis of Iraqi Arabic Dialect on Facebook Based on Distributed Representations of Documents. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 18*(3), 20.

Althobaiti, M., Kruschwitz, U., and Poesio, M. (2014). *AraNLP: a Java-based Library for the processing of Arabic text.* Paper presented at the Ninth International Conference on Language Resources and Evaluation, Reykjavik, Iceland.

Altowayan, A. A. and Tao, L. (2016). *Word embeddings for Arabic sentiment analysis.* Paper presented at the 2016 IEEE International Conference on Big Data, Washington, DC, USA.

Aly, M. A. and Atiya, A. F. (2013). *LABR: A Large Scale Arabic Book Reviews Dataset.* Paper presented at the 51st Annual Meeting of the Association for Computational Linguistics, Sofia, Bulgaria.

Internet: Bagadiya, J. (2018). 171 Amazing Social Media Statistics You Should Know in 2018. Retrieved from URL: http://www.webcitation.org/query?url=https%3A%2F%2Fwww.socialpilot.co%2Fblog%2Fsocial-media-statistics&date=2019-03-02, Accessed: 2019.03.02.

Baly, R., Hajj, H., Habash, N., Shaban, K. B., and El-Hajj, W. (2017). A sentiment treebank and morphologically enriched recursive deep models for effective sentiment analysis in arabic. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 16*(4), 23.

Banea, C., Mihalcea, R., and Wiebe, J. (2010). *Multilingual subjectivity: Are more languages better?* Paper presented at the Proceedings of the 23rd international conference on computational linguistics, Beijing, China.

Bassiouney, R. (2009). *Arabic sociolinguistics*. UK: Edinburgh University Press, 10.

Bhadane, C., Dalal, H., and Doshi, H. (2015). Sentiment analysis: measuring opinions. *Procedia Computer Science, 45*, 808-814.

Blanc, H. (1964). *Communal dialects in Baghdad*. USA: Harvard UP, 12-17.

Boudad, N., Faizi, R., Thami, R. O. H., and Chiheb, R. (2017). Sentiment analysis in Arabic: A review of the literature. *Ain Shams Engineering Journal, 9*(4), 2479-2490.

Chen, X., Xue, Y., Zhao, H., Lu, X., Hu, X., and Ma, Z. (2018). A novel feature extraction methodology for sentiment analysis of product reviews. *Neural Computing and Applications*, 1-18.

Cherif, W., Madani, A., and Kissi, M. (2015a). *A new modeling approach for Arabic opinion mining recognition.* Paper presented at the 2015 Intelligent Systems and Computer Vision (ISCV), Fez, Morocco.

Cherif, W., Madani, A., and Kissi, M. (2015b). Towards an efficient opinion measurement in Arabic comments. *Procedia Computer Science, 73*, 122-129.

Collobert, R. and Weston, J. (2008). *A unified architecture for natural language processing: Deep neural networks with multitask learning.* Paper presented at the Proceedings of the 25th international conference on Machine learning.

Dahou, A., Xiong, S., Zhou, J., Haddoud, M. H., and Duan, P. (2016). *Word embeddings and convolutional neural network for arabic sentiment classification.* Paper presented at the Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, Osaka,Japan.

Duwairi, R., 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, 29*(1), 107-117.

Duwairi, R., Marji, R., Sha'ban, N., and Rushaidat, S. (2014). *Sentiment analysis in arabic tweets.* Paper presented at the 2014 5th international conference on Information and communication systems (ICICS), Irbid, Jordan.

Duwairi, R., Marji, R., Shaban, N., and Ershaidat, S. (2012). *Sentiment Analysis.* B.S. Thesis, Jordan University of Science and Technology, Jordan

Duwairi, R. and Qarqaz, I. (2014). *Arabic sentiment analysis using supervised classification.* Paper presented at the 2014 International Conference on Future Internet of Things and Cloud (FiCloud), Barcelona, Spain.

ElSahar, H. and El-Beltagy, S. R. (2015). *Building large arabic multi-domain resources for sentiment analysis.* Paper presented at the International Conference on Intelligent Text Processing and Computational Linguistics, Cairo, Egypt.

Farghaly, A. and Shaalan, K. (2009). Arabic natural language processing: Challenges and solutions. *ACM Transactions on Asian Language Information Processing (TALIP), 8*(4), 14.

Frome, A., Corrado, G. S., Shlens, J., Bengio, S., Dean, J., and Mikolov, T. (2013). *Devise: A deep visual-semantic embedding model.* Paper presented at the Advances in neural information processing systems, Nevada, USA.

García-Pablos, A., Cuadros, M., and Rigau, G. (2018). W2VLDA: Almost unsupervised system for aspect based sentiment analysis. *Expert systems with Applications, 91*, 127-137.

Grigore, G. (2005). Conditional Structures in Baghdadi Arabic. *Revue roumaine de linguistique, 3*(4), 273-281.

Grigore, G. (2014). The Verb of Perception šāf "to see" in Baghdadi Arabic. *Romano-Arabica, 14*, 139.

Haeri, N. (2003). *Sacred language, ordinary people: Dilemmas of culture and politics in Egypt*. UK: Palgrave Macmillan, 23.

Hassan, Q. (2015). *Concerning some negative markers in South Iraqi Arabic.* Paper presented at the 11th International Conference of AIDA (Arabic Varieties: Far and Wide), Bucharest, Romania.

Hathlian, N. F. B. and Hafezs, A. M. (2016). *Sentiment-subjective analysis framework for arabic social media posts.* Paper presented at the Saudi International Conference on Information Technology (Big Data Analysis) (KACSTIT), Riyadh, Saudi Arabia.

Hayran, A. and Sert, M. (2017). *Sentiment analysis on microblog data based on word embedding and fusion techniques.* Paper presented at the 2017 25th Signal Processing and Communications Applications Conference (SIU), Antalya, Turkey.

Heikal, M., Torki, M., and El-Makky, N. (2018). Sentiment Analysis of Arabic Tweets using Deep Learning. *Procedia Computer Science, 142*, 114-122.

Holes, C. (2004). *Modern Arabic: Structures, functions, and varieties*. USA: Georgetown University Press, 16.

Ibrahim, H. S., Abdou, S. M., and Gheith, M. (2015a). *MIKA: A tagged corpus for modern standard Arabic and colloquial sentiment analysis.* Paper presented at the 2015 IEEE 2nd International Conference on Recent Trends in Information Systems (ReTIS), , Kolkata, India.

Ibrahim, H. S., Abdou, S. M., and Gheith, M. (2015b). Sentiment analysis for modern standard Arabic and colloquial. *International Journal on Natural Language Computing, 4*(2), 95-109.

Ingham, B. (2009). The dialect of the euphrates bedouin, a fringe mesopotamian dialect. In E. Al-Wer & R. d. Jong (Eds.), *Arabic Dialectology*. BOSTON, USA: Brill, pp. 99-108.

Internet: InternetWorldStats. (2019). INTERNET WORLD USERS BY LANGUAGE. Retrieved from URL: http://www.webcitation.org/query?url=https%3A%2F%2Fwww.internetworldstats.com%2Fstats7.htm%3Futm_source%3Dlasindias.info%2Fblog&date=2019-05-06, Accessed: 2019.05.06.

Itani, M., Roast, C., and Al-Khayatt, S. (2017). *Corpora for sentiment analysis of Arabic text in social media.* Paper presented at the 2017 8th International Conference on Information and Communication Systems (ICICS), Irbid, Jordan.

Internet: IWS. (2017). Iraq Internet Usage and Marketing Report. Retrieved from URL: http://www.webcitation.org/query?url=https%3A%2F%2Fwww.internetworldstats.com%2Fme%2Fiq.htm&date=2019-03-02, Accessed: 2019.03.02.

Jarrar, M., Habash, N., Alrimawi, F., Akra, D., and Zalmout, N. (2017). Curras: an annotated corpus for the Palestinian Arabic dialect. *Language Resources and Evaluation, 51*(3), 745-775.

Internet: Jünger, J. and Keyling, T. (2018). Facepager: An application for generic data retrieval through APIs. Retrieved from URL: http://www.webcitation.org/query?url=https%3A%2F%2Fgithub.com%2Fstrohne%2FFacepager&date=2019-03-13, Accessed: 2019.03.13.

Le, Q. and Mikolov, T. (2014). *Distributed representations of sentences and documents.* Paper presented at the International Conference on Machine Learning, Beijing, China.

Maamouri, M., Bies, A., Buckwalter, T., and Mekki, W. (2004). *The penn arabic treebank: Building a large-scale annotated arabic corpus.* Paper presented at the NEMLAR conference on Arabic language resources and tools, Cairo, Egypt.

Mikolov, T. (2012). *Statistical language models based on neural networks.* PhD, Brno University, Brno, Czech Republic.

Internet: Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *arXiv*. Retrieved from https://arxiv.org/abs/1301.3781, Accessed: 2019.03.12

Internet: Mikolov, T., Le, Q. V., and Sutskever, I. (2013b). Exploiting similarities among languages for machine translation. *arXiv*. Retrieved from https://arxiv.org/abs/1309.4168, Accessed: 2019.01.014

Mikolov, T., Yih, W.-t., and Zweig, G. (2013c). *Linguistic regularities in continuous space word representations.* Paper presented at the Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Atlanta, Georgia.

Mnih, A. and Hinton, G. E. (2009). *A scalable hierarchical distributed language model.* Paper presented at the Advances in neural information processing systems, Nevada, USA.

Mnih, A. and Teh, Y. W. (2012). *A fast and simple algorithm for training neural probabilistic language models.* Paper presented at the 29th International Conference on Machine Learning, Edinburgh, Scotland, UK.

Nabil, M., Aly, M., and Atiya, A. (2015). *ASTD: Arabic sentiment tweets dataset.* Paper presented at the Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portuga.

Naili, M., Chaibi, A. H., and Ghezala, H. H. B. (2017). Comparative study of word embedding methods in topic segmentation. *Procedia Computer Science, 112*, 340-349.

Nasser, A. (2018). *Large-Scale Arabic Sentiment Corpus and Lexicon Building for Concept-Based Sentiment Analysis Systems.* PhD Thesis, Hacettepe University, Ankara, Turkey.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*(Oct), 2825-2830.

Pennington, J., Socher, R., and Manning, C. (2014). *Glove: Global vectors for word representation.* Paper presented at the Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), Doha, Qatar.

Ponomareva, N. (2014). *Graph-based approaches for semi-supervised and crossdomain sentiment analysis.* PhD Thesis, University of Wolverhampton.

Internet: Qiu, L. (2015). Sentiment Analysis using Doc2Vec in gensim. Retrieved from URL: http://www.webcitation.org/query?url=https%3A%2F%2Fgithub.com%2Flinanqiu%2Fword2vec-sentiments&date=2019-03-02, Accessed: 2019.03.02.

Refaee, E. and Rieser, V. (2014). *An Arabic Twitter Corpus for Subjectivity and Sentiment Analysis.* Paper presented at the The International Conference on Language Resources and Evaluation (LREC), Reykjavik, Iceland.

Rushdi-Saleh, M., Martín-Valdivia, M. T., Ureña-López, L. A., and Perea-Ortega, J. M. (2011). OCA: Opinion corpus for Arabic. *Journal of the Association for Information Science and Technology, 62*(10), 2045-2054.

Sghaier, M. A. and Zrigui, M. (2016). *Sentiment Analysis for Arabic e-commerce websites.* Paper presented at the International Conference on Engineering and MIS (ICEMIS), Agadir, Morocco.

Shoukry, A. and Rafea, A. (2012). *Sentence-level Arabic sentiment analysis.* Paper presented at the 2012 International Conference on Collaboration Technologies and Systems (CTS), Denver, CO, USA.

Shoukry, A. and Rafea, A. (2015). *A hybrid approach for sentiment classification of Egyptian Dialect Tweets.* Paper presented at the 2015 First International Conference on Arabic Computational Linguistics (ACLing), Cairo, Egypt.

Socher, R., Huang, E. H., Pennin, J., Manning, C. D., and Ng, A. Y. (2011). *Dynamic pooling and unfolding recursive autoencoders for paraphrase detection.* Paper presented at the Advances in neural information processing systems.

Soliman, A. B., Eissa, K., and El-Beltagy, S. R. (2017). AraVec: A set of Arabic Word Embedding Models for use in Arabic NLP. *Procedia Computer Science, 117*, 256-265.

Internet: StatCounter. (2017). Social Media Stats Iraq. Retrieved from URL: http://www.webcitation.org/query?url=http%3A%2F%2Fgs.statcounter.com%2Fsocial-media-stats%2Fall%2Firaq%2F%23yearly-2017-2018-bar&date=2019-03-02, Accessed: 2019.03.02.

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., and Tsujii, J. i. (2012). *BRAT: a web-based tool for NLP-assisted text annotation.* Paper presented at the 13th Conference of the European Chapter of the Association for Computational Linguistics, Avignon, France.

Suçin, M. H. (2015). *mawāqif asatidhat al-lugha al-ʿarebiyya wa ṭullābihā min tadrīs al-lahjāt al-ʿarabiyya ʿalā al-mustawā al-jāmiʿī fī Turkiyā [Views of Arabic language teachers and their students about the teaching of Arabic dialects at the university level in Turkey]*. Paper presented at the 11th International Conference of AIDA (Arabic varieties – Far and wide), Bucharest.

Suleiman, D., Awajan, A., and Al-Madi, N. (2017). *Deep Learning Based Technique for Plagiarism Detection in Arabic Texts.* Paper presented at the International Conference on New Trends in Computing Sciences (ICTCS).

Tubishat, M., Idris, N., and Abushariah, M. A. (2018). Implicit aspect extraction in sentiment analysis: Review, taxonomy, oppportunities, and open challenges. *Information Processing & Management, 54*(4), 545-563.

Turian, J., Ratinov, L., and Bengio, Y. (2010). *Word representations: a simple and general method for semi-supervised learning.* Paper presented at the Proceedings of the 48th annual meeting of the association for computational linguistics.

Internet: UNESCO. (2017). Unesco World Arabic Language Day. Retrieved from URL: http://www.webcitation.org/query?url=https%3A%2F%2Fen.unesco.org%2Fcommemorations%2Fworldarabiclanguageday%2F2017&date=2019-03-02, Accessed: 2019.03.02.

Versteegh, K. (2014). *Arabic Language* (Second ed.). UK: Edinburgh University Press, 34-61.

Internet: WorldAtlas. (2018). The World's Most Spoken Languages. Retrieved from URL: http://www.webcitation.org/query?url=https%3A%2F%2Fwww.worldatlas.com%2Farticles%2Fmost-popular-languages-in-the-world.html&date=2019-05-06, Accessed: 2019.05.06.

Zhila, A., Yih, W.-t., Meek, C., Zweig, G., and Mikolov, T. (2013). *Combining heterogeneous models for measuring relational similarity.* Paper presented at the Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.

Zou, W. Y., Socher, R., Cer, D., and Manning, C. D. (2013). *Bilingual word embeddings for phrase-based machine translation.* Paper presented at the Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing.

APPENDICES

Appendix-1. IJMES transliteration system for Arabic

Consonants

ء	ʾ	ز	z	ك	k
ب	b	ژ	zh	گ	g
پ	p	س	s	ل	l
ت	t	ش	sh	م	m
ث	th	ص	ṣ	ن	n
ج	j	ض	ḍ	ه	h
چ	ch	ط	ṭ	و	w
ح	ḥ	ظ	ẓ	ي	y
خ	kh	ع	ʿ	ة	a
د	d	غ	gh	ال	al
ذ	dh	ف	f		
ر	r	ق	q		

Vowels

Long ا or	ى	ā
	و	ū/ō
	ي	ī
Doubled	ـِيّ	iyy (final form ī)
	ـُوّ	uww (final form ū)
Diphthongs	ـَو	au
	ـَي	ai
Short	ـَ	a
	ـُ	u/o
	ـِ	i

CURRICULUM VITAE

Personal Information

Surname, Name : ALNAWAS, Anwar

Nationality : IRAQI

Date and Place of Birth : 29.06.1983, ThiQar

Marital status : Married

Phone number : +9647816101727

E-mail : anwar.alnawas@stu.edu.iq

Education

Degree	School/ Program	Graduation Date
PhD	Gazi University / Computer Engineering	Ongoing
MSc	UUM / Information Technology	2010
Undergraduate	Thi-Qar University / Computer Sciences	2005
High School	Souk Al-Shuyukh High School	2001

Professional Experience

Year	Place of Work	Position
2011-Ongoing	Technical Institute in Nasiriyah, STU	Lecturer
2006-2008	Ministry of Education	A high school lecturer

Foreign Language

Arabic (Native), Turkish, English

Publications

Alnawas, A. and Arıcı, N. (2018). The corpus based approach to sentiment analysis in modern standard Arabic and Arabic dialects: a literature review. *Journal of Polytechnic, 21*(2), 461-470.

Alnawas, A. and Arıcı, N. (2019). Sentiment analysis of Iraqi Arabic dialect on Facebook based on distributed representations of documents. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 18*(3), 20.

Alnawas, A. and Arıcı, N. (2018). *Effect of word embedding variable parameters on Arabic sentiment analysis performance* [abstract]. Paper presented at the 5th International Conference on Computational and Experimental Science and Engineering, Antalya, Turkey.

Hobbies

Internet, Reading, Travel

GAZİ GELECEKTİR..