DEEP LEARNING FOR SENTIMENT AND EMOTION DETECTION IN MULTILINGUAL CONTEXTS

by

Malak Abdel-Ghani Abdullah

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Computing and Information Systems

Charlotte

2018

Approved by:

____
Dr. Mirsad Hadzikadic

____
Dr. Samira Shaikh

____
Dr. Mohamed Shehab

____
Dr. Richard Lambert

ABSTRACT

MALAK ABDEL-GHANI ABDULLAH. Deep Learning for Sentiment and Emotion Detection in Multilingual Contexts. (Under the direction of DR. MIRSAD HADZIKADIC)

Social media is growing as a communication medium where people can express online their feelings and opinions on a variety of topics in ways they rarely do in person. Detecting sentiments and emotions in texts have gained a considerable amount of attention in the last few years. Thus, the terms sentiment analysis and emotion detection have taken their own path to become essential elements of computational linguistics and text analytics. These terms are designed to detect peoples' opinions and emotions that consist of subjective expressions across a variety of products or political decisions. Recently, the Arab region has played a significant role in international politics and in the global economy, which has grasped the attention of political and social scientists. Yet, the Arabic language has not received proper attention from modern computational linguists.

This dissertation provides a comprehensive study of sentiment analysis and emotion detection on Twitter data and analyzes the existing work that has been accomplished to detect and analyze English and Arabic tweets. It also examines a case study where a random sample of tweets has been extracted that reflect people's sentiments regarding a political event. In this case study, an R package "Sentiment" has been applied to detect sentiments and emotions in the extracted tweets. The results demonstrate a need for more investigation towards improving the effectiveness and efficiency of sentiment and emotion detection systems. Therefore, the main contribution of this dissertation is to propose a system that automatically determines the intensity of sentiments and emotions in both languages. Emotion detection for Arabic text is relatively new; to the best of our knowledge, the proposed system is the first system to detect the intensity of emotions for Arabic text using deep learning approaches. The main data inputs to the system are a combination of word and document embeddings and a set of psycholinguistic features (e.g., AffectiveTweets Weka-package, Deepmoji, Unsupervised Sentiment Neurons). Our approach is novel in using and applying CNN-LSTM with fully connected neural network architecture to obtain performance results that show substantial improvements in Spearman correlation scores over the baseline models. In addition to the aforementioned contributions, this dissertation aims to optimize the model performance for both languages by constructing and selecting informative features. It illustrates the contribution of deep learning in sentiment and emotion detection and highlights the role of using the extracted features from raw Arabic tweets and Arabic tweets translated into English during this process.

DEDICATIONS

In the name of ALLAH, Most Gracious, Most Merciful

"Allah will raise up to (suitable) ranks (and degrees), those of you who believe and who have been granted knowledge. And Allah is Well-Acquainted with all ye do"

AYAH al-Mujadilah 58:11

This dissertation is dedicated to:

My beloved father Abdel-Ghani Abdullah

And my dearest mother Nadia Ali

For their endless love, prayers, encouragement, and sacrifices.

You have successfully made me the person that I am now.

and

The loving memory of my brother, Azhar.

You will always be remembered.

ACKNOWLEDGMENTS

At the very outset, Alhamdulillah, all praises to ALLAH, the source of wisdom and knowledge, for His blessing and giving me the strength, knowledge, ability, and opportunity to complete this work successfully.

Pursuing a Ph.D. is a life-changing personal journey. Although it may seem as one person's effort, the truth is pursuing this path comes with the effort and support of remarkable individuals. With great pleasure, I would like to acknowledge the support, assistance, and contributions made by individuals in the course of my doctoral program from the beginning of the fieldwork to the writing process until the completion of this dissertation.

With great honor and sense of gratitude, I would like to express my deepest appreciation to my advisor, Dr. Mirsad Hadzikadic, whose constant encouragement was vital in making this dissertation a reality. I have been extremely lucky to have a supervisor who provided insightful discussions about the research and who responded to my questions and queries so promptly despite his busy agenda. Thanks to his supportive attitude, amazing guidance, perfect time management, endless encouragement, sense of humor, and consistently wise advice, he was a true source of inspiration. It was a real privilege and a great pleasure for me not only to learn from his exceptional knowledge but also from his extraordinary human qualities.

Besides my advisor, I would like to thank the rest of my respected dissertation committee members, Dr. Samira Shaikh, Dr. Mohamed Shehab, and Dr. Richard Lambert, for generously offering their time, support, guidance, and feedback through-out the preparation of this work. Their advice on both research as well as my career has been invaluable. I am deeply indebted to Dr. Shaikh for her crucial remarks that shaped my final dissertation. Without her constructive feedback and continuous encouragement, I would not have been able to complete the dissertation.

I gratefully acknowledge the funding sources that made my Ph.D. work possible. I was awarded a full scholarship from Jordan University of Science and Technology. I also was honored with a Graduate Assistantship in Software and Information Systems department at UNC Charlotte.

I am grateful to all of those with whom I have had the pleasure to work during this and other related work. I would like to thank my lab mates at Complex System Institute for their continuous support and for making my experience in the lab and graduate school exciting and fun. I also spent a summer at the World Well Being Project at the University of Pennsylvania, where I had the chance to collaborate with fantastic researchers. Thanks to all of them. More specifically, I would like to thank Dr. Muhammad Abdul-Mageed and Hassan AlHuzali for the fruitful discussions and for making that summer an eye-opening experience. Moreover, I must express my gratitude to the winning team that I was a member of in the Data Science Challenge at UNC Charlotte, led by Dr. Lixia Yao and Dr. Wenwen Dou.

My time at UNC Charlotte was made enjoyable in large part due to the many friends (too many to list here but you know who you are!) and groups that became a part of my life. I am grateful for what I have learned from them by being exposed to their cultures and for our memorable travels to conferences, to the mountains and remarkable places, and for the great doses of fun that we have had together. The hazards of Ph.D. life would be insufferable without their constant support and love.

For writing the dissertation, I am grateful to Dr. Lisa Russell-Pinson. With her writing class and her directions, I was able to explore my ability to write this work. Moreover, I want to fondly mention my friend, Yousra Javed, with whom I kept exchanging ideas and good-spirited discussions relating to research and the Ph.D. world. Thanks, Yousra, for always being so supportive, proofreading my research paper drafts and posters, and creating the incredible memories that we have had together. I also appreciate the efforts of Chris Harrington, Senior Tutor in the Writing Resources Center at UNC Charlotte, who provided me with the best advice and the knowledge to improve my writing skills. Moreover, my sincere thanks go to UNC Charlotte staff who all helped me in numerous ways during various stages of my PhD. Thanks to Sandra Krause and Julie Fulton for caring and supporting me wholeheartedly. The role they played in my Ph.D. life was invaluable. I'm grateful for all of your help and continued support.

Moreover, I have special friends, students, and colleagues in Jordan that I want to thank for motivating me and inspiring me to complete this work. My immense thanks to Dr. Rehab Duwairi, a former dean of Computer and Information Technology College at Jordan University of Science and Technology (JUST), for all of the support she has given me, for being a wonderful teacher, and for her can-do spirit that helped me get where I am today. I want also to express my gratitude to all friends and colleagues from Jordan, who I cherish for all the lovely memories we have created together over the years and for the great value of our genuine friendships. Thank you for the words of encouragement and motivation to pursue my Ph.D.; you have been truly a blessing in my life. Talking about Jordan brings up the fond memories of my students, who are by far the biggest reason to embark upon this journey. Thanks to my students for inspiring me with their life stories, for respecting and loving me, and for showing me that teaching is the perfect career for me.

Finally, I would like to express my deepest gratitude and most heartfelt thanks to my family. This dissertation would not have been possible without their warm love, continued patience, and endless support. Words can not express how grateful I am to my parents for all of the sacrifices that they have made on my behalf. Their prayers for me are what have sustained me thus far. Thanks for encouraging me to be an independent thinker and having confidence in my ability to go after new things that inspired me. Thanks for everything that has helped me get to this day.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION

"And suddenly you know: It's time to start something new and trust the magic of beginnings."

–Meister Eckhart

1.1 Overview

The rise and diversity of social microblogging channels encourage people to express their feelings and opinions on a daily basis. Consequently, sentiment analysis and emotion detection have gained the interest of researchers in natural language processing and other fields that include political science, marketing, communication, social sciences, and psychology [8, 24, 74]. Twitter plays a vital role in spreading information and influencing people's opinions in a specific direction. As an easy-to-use platform, Twitter motivates people to share their thoughts and express their opinions. It has been shown by researchers that tracking and analyzing public opinions from social media can help to predict certain events.

In general, sentiment analysis refers to classifying a subjective text as positive, neutral, or negative, whereas emotion detection recognizes types of feelings through the expression of texts, such as anger, joy, fear, and sadness [8, 31]. Studying people's emotions attracted psychologists and behavioral scientists for a long time [30, 111].

Several theories have formed a list of basic emotions [83]. Ekman [32] identified the six basic emotions as anger, disgust, fear, happiness, sadness, and surprise. Plutchik [83] added two more emotions to Ekman's list: trust and anticipation. Arnold [16] listed eleven emotions: anger, aversion, courage, dejection, desire, despair, fear, hate, hope, love, and sadness. Numerous sentiment analysis studies annotated texts as positive or negative. However, only a few studies exist that built corpus for emotion labeling [14, 79]. An interest in studying and building models for sentiment analysis and emotion detection for social microblogging platforms has increased significantly in recent years [62, 85, 84, 51]. Going beyond the task of mainly classifying tweets as positive or negative, several approaches to detect emotions were presented in previous research papers [76, 112, 73]. Researchers [74, 75] introduced shared tasks of detecting the intensity of emotion felt by the speaker of a tweet.

Recently, the Arab region has played a significant role in international politics and in the global economy that has grasped the attention of political and social scientists. Detecting Arabic tweets will be helpful for politicians in predicting global events based on the popular news and people's comments. However, the Arabic language has not received a proper attention from modern computational linguists. The decision to study Arabic sentiment analysis is motivated by several factors. The main factor is realizing that Arabic is a resource-poor language relative to other languages. Although Arabic is the fifth most widely spoken language in the world[1], there is a shortage of language resources and minimal support to analyze Arabic sentiments. It is believed that the main reasons behind the lack of studying and

[1] https://www.redlinels.com/most-widely-spoken-languages

analyzing the Arabic language is its complex morphology and structure in addition to the limited research funding in this area [12, 106, 18]. Although recent research has been dedicated to detect emotions for English content, to our knowledge, there are few studies for Arabic content. Researchers [90] collected and annotated data and applied different preprocessing steps related to the Arabic language. They also used a simplification of the SVM (known as SMO) and the NaiveBayes classifiers. Another two related works [59, 94] shared different tasks to identify the overall sentiments of the tweets or phrases taken from tweets in both English and Arabic. Our work is using the state-of-the-art approaches of deep learning and word/doc embeddings. To the best of our knowledge, there is no emotion detection system for Arabic tweets that uses these approaches.

Sentiment analysis usually goes through the following phases to predict sentiments or emotions, see Figure 1:

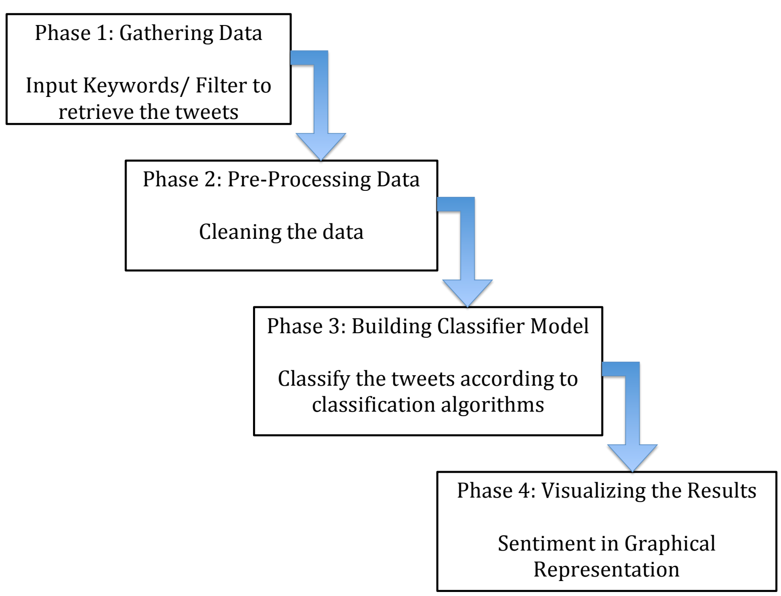


Figure 1: Work-Flow chart for sentiment analysis process

Crawling data from social media highly relies on Application Programming Interfaces (API) that are provided by social media platforms themselves. The Twitter microblogging platform provides a streaming API to extract tweets in real-time. Twitter uses OAuth to provide authorized access to its API [2]. This helps researchers to retrieve the required tweets by using filtering strategies on specific keywords or phrases. It also helps to retrieve all the public tweets or re-tweets for specific users. The list below is glossary of key terms that are used in the Twitter social network[3]:

1. Tweets: The messages that are posted on Twitter are publicly visible by default, but some users restrict their messages to be seen only by their followers.

2. Hashtag: A hash character (#) that is placed with a word or phrase on social network to make it easier for users to find all the information regarding a specific topic or content.

3. Tags or mention: The symbol @ is assigned to a user name to mention the user in a tweet or to reply to his/her tweet.

4. Trending topic: A topic that is mentioned at a greater rate than other topics in a specific time.

5. Follow: A link that gives the user the ability to follow other users' posts.

6. Emoticon: Emotion faces that indicate the user's feelings about the tweet.

7. Retweet or (RT): This function is used to forward a tweet that someone else has initiated to the recipient's followers.

---

[2] REST API resource, https://dev.twitter.comldocs/api

[3] https://help.twitter.com/en/glossary

Data-preprocessing, aka text normalization, involves converting sentences from the retrieved data into words. Moreover, it removes any irrelevant and redundant information and eliminates noisy and unreliable data by deleting white-spaces, punctuation marks, numbers, and URL links. It also filters stop words, tags, hashtags, and common words that have little meaning. In addition, this phase performs stemming by reducing words to their roots, and all uppercase letters will be converted to lowercase to make it easier to compare words in the next phase. Furthermore, this phase computes term frequency and inverse document frequency for each word in the corpus.

Choosing a classifier with extracting the features comprises half the sentiment analysis process. There are several techniques that perform sentiment analysis on text data. According to Boiy [22], Symbolic and Machine Learning techniques are the two basic methodologies used in sentiment analysis for text [106] (Figure 2).

The algorithms that have been conducted with features are divided into two main groups:

1. Extraction (Transformation): A process through which a new set of features is created from ”raw” data. It also includes scaling, converting, or modifying features. [67]. Vectorization is the general process of turning a collection of text documents into numerical feature vectors. One example of vectorization is Bag of Words or Bag of n-grams representation [52]. In this representation, a corpus of documents can be represented by a matrix with one row per document and one column per word, disregarding grammar or word order. Other examples of

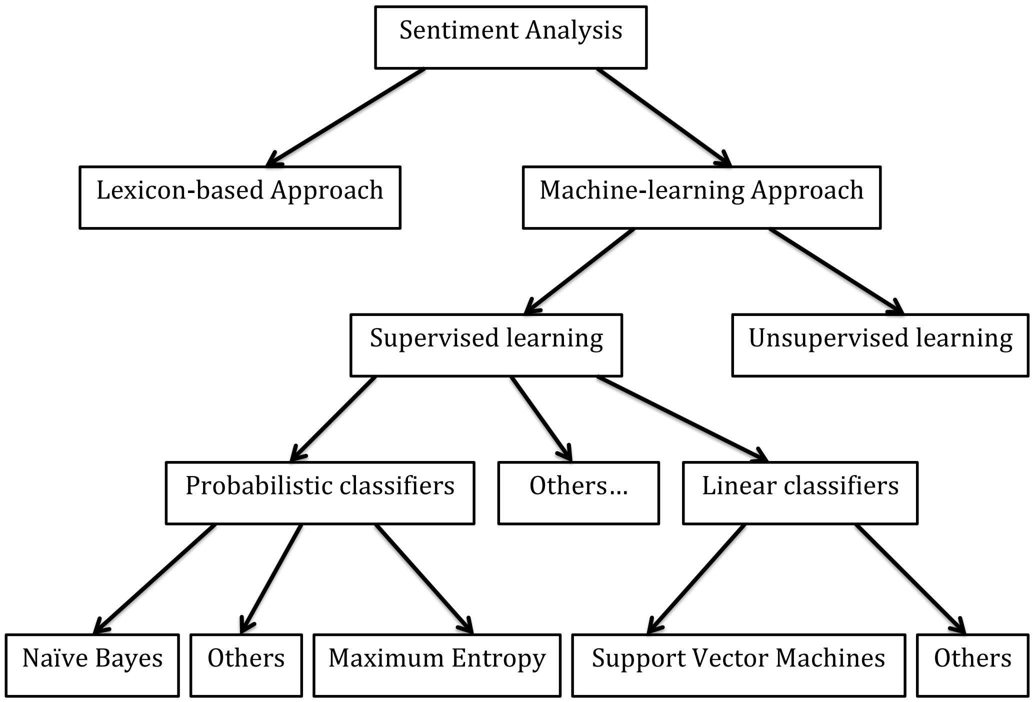


Figure 2: Sentiment analysis classification algorithms

feature vectorization methods include:

(a) TF-IDF (term frequency-inverse document frequency): a numerical statistic intended to reflect the importance of a term to a document in the corpus [115]. Denote a term by t, a document by d, and the corpus by D. Term frequency $TF(t,d)$ is the number of times that term t appears in document d. IDF is a numerical measure of how much information a term provides. The TF-IDF measure is simply the product of TF and IDF.

(b) Word embeddings: An improvement over Bag of Words, it is an approach to provide a dense vector representation of words that capture something about their meaning. It takes as its input a substantial corpus of text and produces a vector space. The word2vec model was created by a team of researchers led by Tomas Mikolov [72] to produce word embeddings.

2. Selection: the most relevant attributes or features as a subset of a larger set of features. Knowing that some features are simply not correlated to other features, they can have a negative effect on the performance of a model. The problem of feature selection is trying to find the optimal subset of features, which can lead to NP-hard [116].

Neural networks (NN) is becoming gradually popular in language modeling tasks. Deep Neural Networks (DNN) have recently shown significant improvements over traditional machine learningbased approaches on classification tasks[41]. This dissertation study aims to use DNNs for sentiment and emotion detection. Therefore, it will be convenient to provide an overview for this subject. Neural networks are inspired by biological nervous systems in animal brains. They are composed of collection of connected units or nodes called neurons. As in nature, each connection between neurons can transmit a signal from one to another. The network function is determined by these connection. Training a neural network depends on adjusting the values of connections, which called weights, between neurons. The inputs to neurons are real-valued numbers and the output of each neuron is calculated by a non-linear function of the sum of its inputs. The weights of the connections are adjusted to lead to a specific target output. Commonly, neurons are organized in layers, so the output of first layer is the input for the second layer, and so on. Therefore, increasing the number of training examples (input/target pairs), the network can learn faster and improve its accuracy [26]. Figure 3 shows an example of neural network system.

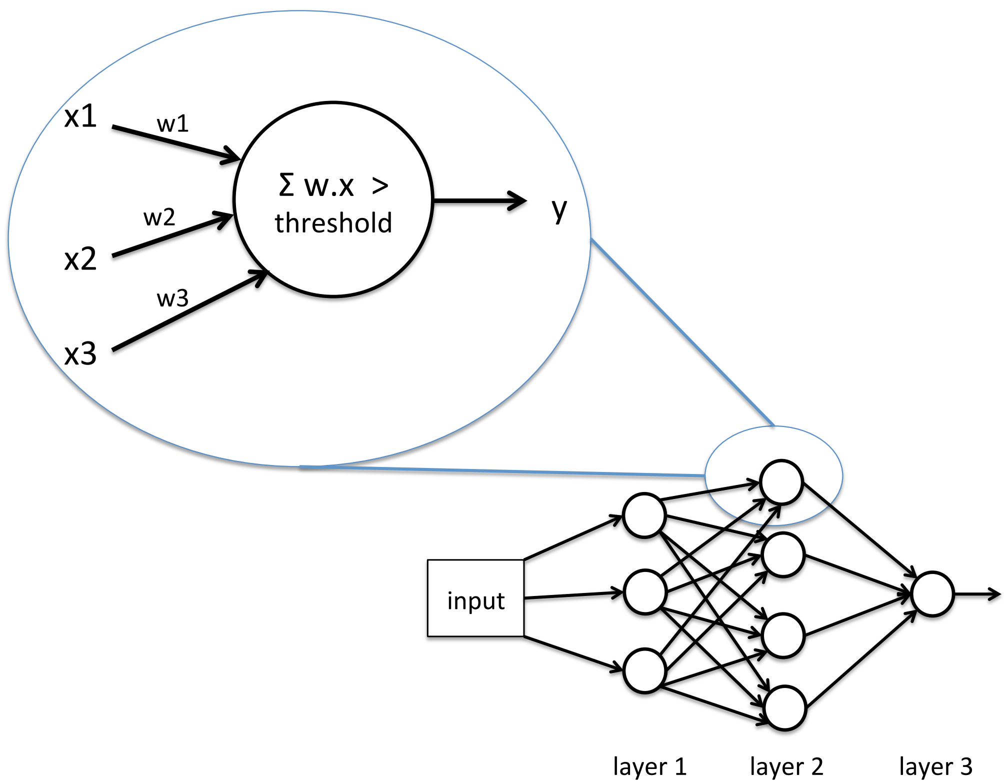


Figure 3: Neural Network system

With feed-forward neural network, the signals travel from input to output in one way only. Whereas with Recurrent Neural Networks (RNN), the signals travel in both directions by introducing loops in the network. Computations derived from earlier input are fed back into the network as input of the next time step, which gives them a kind of memory and makes RNN aware of time. The Convolutional Neural Network (CNN) [66] is a type of feed-forward artificial neural network, and Long Short Term Memory network (LSTM) [48] is a special kind of RNN. In recent years, both networks have become the state-of-the-art models for a variety of machine learning problems. A common architecture for LSTM is composed of a memory cell, an input gate, an output gate, and a forget gate. The cell stores a value (or state), for either long or short time periods. This is achieved by using activation function for the memory cell. CNN makes an efficient use of layers with convolving filters that are applied to local features. In [41, 19, 58], researchers show that CNNs and LSTMs outperform the traditional machine learning approaches on text classifications, such as sentiment, emotion, and stance detections. This dissertation uses feed-forward, LSTM, and CNN to predict the sentiment and emotion in a tweet.

1.2 Problem Statement

This dissertation primarily reviews the efforts of building sentiment analysis systems for English and Arabic languages and proposes a system to detect the intensity of sentiments and emotions in both languages that is comparable to the state-of-the-art systems.

The motivations of this dissertation are the following: 1- Explore and examine the existing techniques that have been used to analyze Arabic tweets. 2- Design a system to automatically determine the intensity of emotions and sentiments in English and Arabic languages that competes the state-of-the-art models. 3- Obtain performance results that show substantial improvements in different measurement scores and metrics by selecting the informative features for both languages. 4- Shed light on the extracted features from the original Arabic tweets and those extracted from translating the Arabic into English languages.

1.3 Dissertation Structure

The remainder of this dissertation is organized as follows: Chapter 2 provides a comprehensive study of sentiment analysis on Twitter data and analyzes the existing work that has been done in order to detect and analyze Arabic tweets. It also summarizes the key findings of recent research in this field and concludes with future directions of research. Chapter 3 gives a case study on Tweets' sentiment analysis by collecting and analyzing American originated tweets that mentioned then-US Republican presidential candidate Donald Trump after his primary debates. Chapter 4 proposes a system to detect the intensity of emotions in English and Arabic tweets that is comparable to the state-of-the-art systems. The main input to the system is a combination of word2vec and doc2vec embeddings and a set of psycholinguistic features (e.g. from AffectiveTweets Weka-package). We apply a fully connected neural network architecture and obtain performance results that show substantial improvements in Spearman correlation scores over the baseline models. Chapter 5 focuses on Arabic language's system to detect and determine the intensity of sentiments and emotions in tweets. A CNN-LSTM architecture is added as a model to improve the results of the previous model. Chapter 6 concludes this dissertation and discusses potential techniques and plans to improve the performance of sentiment and emotion detection to extend this work.

CHAPTER 2: BACKGROUND AND LITERATURE REVIEW ON TWITTER SENTIMENT ANALYSIS

"If you don't know history, then you don't know anything. You are a leaf that doesn't know it is part of a tree."

–Michael Crichton

2.1 Introduction

Sentiment analysis and opinion mining are considered hot topics where researchers are extracting information regarding emotions and viewpoints. It is believed that these concepts consist of subjective expressions across a variety of products or political decisions [8, 102]. The terms sentiment analysis and opinion mining are not exactly the same. The meaning of the term "opinion" is broader than the term "sentiment". Prior researchers have used these two terms interchangeably. In this literature review, the term sentiment analysis has been used to refer to both of them. Sentiment analysis is used to track attitudes and opinions on the web and determines if the audience positively or negatively receives these ideas. This helps companies determine strategies for improving the quality of their products or to assist decision makers. Sentiment analysis of data involves building a system by using natural language processing, statistics, and machine learning methods to examine opinions or sentiments in a text unit [8].

Microblogging services, such as Twitter and Facebook, are considered important communication tools for people to share their opinions or spread information. The nature of these microblogs encourages people in their daily lives to post real-time messages about their opinions on current events. People are sharing their daily life activities on these microblogging tools [62].

The Twitter microblog was launched in July 2006, and since then it has gained worldwide popularity. Many scholars hold the view that the use of Twitter is playing a vital role in spreading information and influencing people's opinions in a specific direction. Statistics from Statista website show that the Twitter in 2016 has more than 317 million active users. Many users tweet their opinions on a variety of subjects, discuss many political topics or marketing issues, and express their views on many aspects of their lives. Every tweet has a maximum length of 140 characters. Due to the shortness of the messages, people convey their opinions and thoughts openly most of the time. Therefore, Twitter is considered a rich data bank and one of the largest platforms that is full of sentiments [78].

According to recent reports, the fastest growing language on Twitter between 2010 and 2011 was Arabic [46]. While there is a great need for natural language analysis of large amounts of Arabic language text, the reality shows little work has been done in this area. Most of the sentiment analysis resources and systems built so far are tailored to English and other Indo-European languages. Reasons for the lack of research in this area include the complexity and the variety of dialects of the Arabic language that make it harder to build one system that is applicable to all of its dialects [12] [106].

The Arabic language belongs to the Semitic language family. It is recognized as the fifth most widely spoken language in the world and is considered the official or native language for 22 countries (approximately more than 300 million people) [50][61]. The Arab region has a large, growing population and has become an important player in international politics and the global economy. Furthermore, the Arabic language is in the top ten of the most used languages to create Internet content [106].

This study primarily aims to review the efforts of building sentiment analysis systems for the Arabic language and lists some applications and systems that have been built to analyze Arabic Twitter data. This research also presents a general study of sentiment analysis and explores some of the machine learning algorithms and natural language processing classification techniques.

The remainder of this literature review is organized as follows. Section 2.2 gives the reader general background material on Twitter sentiment analysis by describing techniques and vital features that have been used in this area. Section 2.3 examines the techniques that have been used to analyze Arabic tweets and to summarize the key findings of recent research in this field. This literature review concludes with future directions of research in Section 2.4.

2.2 Background on Sentiment Analysis

In general, tweets generated by users can be categorized as objective or subjective tweets. Objective tweets contain facts that refer to the nature of entities, events, and attributes [107]. An example of an objective tweet is: *Election Day in the United States of America is the Tuesday following the first Monday in November.* While subjective tweets express users' opinions regarding entities, events, and attributes. Subjectivity classification seeks tweets that contain users opinions. Some examples of subjective tweets include:

- *I'm happy election day is almost here.* (positive tweet)
- *I hate this election. Everything about it makes me miserable.* (negative tweet)
- *I don't care who wins the upcoming presidential election.* (neutral tweet)

Sentiment analysis is considered a part of the Natural Language Processing (NLP) field. It was first explored in 2003 by Nasukawa and Yi [82]. In sentiment analysis of Twitter data, the researchers focus their studies on subjective, not objective, tweets. They are interested mainly in classifying tweets as positive and negative [56]. The researchers studied sentiment analysis through three levels. The first level is the document level that classify and analyze sentiments for the whole document [114, 86]. Analyzing sentences is considered a second level. And finally, the phrase level is when the researchers are analyzing sentiments in phrases [117, 7]. They also investigated the utility of linguistic features for detecting the sentiment of Twitter posts.

2.2.1 Sentiment Analysis Work-Flow

The process of performing sentiment analysis for a micro-blogging tool usually goes through multiple phases [12, 18]:

- Phase 1: Data-gathering (crawling data). In this phase, the required amount of tweets that are related to a specific topic are retrieved. This data is filtered according to a particular time frame and keywords/users.

- Phase 2: Data-preprocessing (text normalization). This is an important step in the data mining field. The retrieved data from the first phase will be tokenized by converting the sentences into words. These words will be cleaned to remove any irrelevant and redundant information.

- Phase 3: Building-a-classifier. In this phase, a classifier model will be selected. Subsection 2.1.1.2 discusses the classification techniques that can be used to analyze peoples' sentiments more deeply.

- Phase 4: Visualization. This phase focuses on visualizing the results of sentiments attached to a particular topic and follows opinion changes over time. This can be performed by a graphical representation in several forms.

2.2.2 Sentiments Classification Algorithms

There are many techniques that perform sentiment analysis on Twitter data. According to Boiy [22], Symbolic techniques and Machine Learning techniques are the two basic methodologies used in sentiment analysis for text[106]. Symbolic technique, which is also called Semantic Orientation, uses sentiment lexicons which are lists of words or phrases associated with positive and negative sentiments. Some of these lexicons add other features and provide a score to specify the strength of its class. This approach works to extract the score of its words and sum them up to show an overall positive or negative sentiment. Turney [114] used bag-of-words approach in which the document is treated as a collection of words regardless of the relationship between the words. Turney gave every word a value and combined all the values by using aggregation functions. Turney's technique is used to figure out the overall value for the whole document. On the other hand, Kamps [53] developed a distance metric on wordNet which is a database consisting of words and their relative synonyms. Another simple classifier model is the k-nearest neighbor algorithm that uses distance measure to assign a class label y to x if y is the nearest label to x [108].

Many classifier models have been built to classify tweets as positive, negative, or neutral according to their training data sets. These are grouped under the machine learning umbrella. The term machine learning was first coined by Samuel in the 1950s and was meant to encompass many intelligent activities that could be transferred from human to machine. The research in this field focuses on finding relationships in data [42]. Machine learning modeling methods can be supervised or unsupervised. In the supervised learning classification model, a training labeled set of data are used to predict the class of a search query. While in the unsupervised learning classification models, there is no labeled training data and the model will classify the corpus to specific classes based on some clustering computations. Labeling data in many applications is an expensive process and sometimes it may be labeled with errors and that may reflect the classification results. The unsupervised learning model is used frequently to predict the topic for a page or a text. Out of the sentiment analysis models that are using the supervised modeling in this survey, most of them have been built by using one of three standard algorithms: Naive Bayes classification, Maximum Entropy classification, and Support Vector Machines classification [8].

The efficiency of a classifier depends on the type of engineering feature associated with it. Feature extraction is the process of creating a representation for, or a transformation from, the original data. Numerous feature extraction algorithms have been proposed and successfully applied in many classifier models. Features can be binary, categorical, or continuous. Some of these powerful features are [18, 8]:

1. Term Presence vs. Term Frequency: It has been proven experimentally that the presence of a term is more important than counting the term frequencies.

2. Term Position: The term position can determine the sentiment for a tweet which plays an important role in sentiment analysis.

3. Part-of-Speech: Many articles show that this feature plays an important role in all Natural Language Processing tasks. This feature concentrates on the adjective and adverb words in the text.

4. Unigram: In this feature, a single word can be considered as a feature by itself. The results showed that unigram presence taken as feature turns out to be the most efficient.

Results show that n-grams features are the most widely used features for Twitter sentiments analysis [8].

The performance of sentiment classification system can be evaluated by using a well-known table called (Error Matrix) or (Confusion Matrix) [110]. Each column of the matrix represents predicted classifications and each row represents actual defined classifications. Based on the Confusion Matrix, four indexes can be computed to reflect the performance. These are Accuracy, Precision, Recall and F1-score.

2.3 Arabic Sentiment Analysis

2.3.1 Arabic Language Aspects and Challenges

Arabic is the mother tongue of 22 countries with more than 300 million people speaking that language[50]. It is also the language of more than 1.4 billion Muslims around the world. It has been used for more than 2000 years [57]. The Arabic alphabet consists of 28 letters with no upper or lower cases and the orientation of writing is from right to left. Its letters can be written with different shapes according to their position in the word. According to [50] [44], the Arabic language is classified into two main categories: Standard Arabic (SA) and Dialectical Arabic (DA). SA consists of two forms: Classical Arabic (CA), and Modern Standard Arabic (MSA). CA is the standard poetic language and the language of the Qur'an (Holy Islamic Book). While MSA language, which is a simplified form of CA, is used in most current printed Arabic publications such as books, newspapers, and also used in news broadcasts or formal speeches [104]. Although MSA is the primary language of the media and education in Arab countries, it is not spoken as a native language in people's informal daily communication. In contrast to MSA, DA is spoken but not written in books or taught in schools. DA has a strong presence in texting SMS on cellular phones, commenting on microblogging networks or in emails, blogs, discussion forums, and chats. Each dialect is spoken by a specified geographical area for daily verbal communication. Therefore, there is only one MSA language for all Arabic speakers but several dialects with no formal written form[36]. According to [57], the dialects are affected by many factors such as: which Arab tribe has lived in this geographical area and which foreign language was the source of loanwords. Also, if the geographical area is a village or countryside, or if the people are bedouin or sedentary. Arabic Dialects are greatly varied, and are classified into five main groups according to [118]: Egyptian, Levantine, Iraqi, Gulf, and Maghribi.

Arabic words are classified into three categories: verb, noun, and particle. Most of the words are derived from a root of three up to six letters. Each Arabic word follows a pattern that is inherited from a specific root. The same three-letter root can generate several words with different meanings. For example the Arabic word كتب (Write) is pronounced as (kataba) and has a root of three letters (ك ت ب) (k t b). From this root, a lot of words can be formed using a combination of the root and other letters. These combinations form patterns of words that follow specific rules. Table 1 shows an example of some patterns inherited from this root.

Arabic has a complex and rich morphological structure that makes the number of vocabularies very large. A full meaningful statement can only contain one word. As an example, أنلزمكموها, is one word that forms a complete sentence containing: verb, subject, two objects, and starts with question letter. If we want to rewrite this Qura'anic word in another simple statement, it will be as: هل نحن نلزم انتم هذه (Shall we compel you to accept it).

The sentence structure in Arabic has two types according to what kind of phrases the sentence starts with: (1) verbal sentence and (2) nominal sentence. As an example, the statement: *The boy is eating in the restaurant*, can be written in Arabic by

Table 1: Patterns of some words derived from the root (k t b)

Arabic word	Spelling	Meaning
كَتَبَ	kataba	wrote
مَكتوب	maktoob	written
كاتب	kateb	writer
يَكتُبُ	yaktobo	he is writing
كتاب	ketab	book
مَكتَبَة	maktaba	library
مَكتَب	maktab	office

using both structure types as the following:

1. Verbal sentence يأكل الولد في المطعم
2. Nominal sentence الولد يأكل في المطعم

Where (الولد) means (the boy), (يأكل) means (is eating), and (في المطعم) means (in the restaurant).

2.3.2 Classification Techniques for Arabic Tweets

Minimal work has been done in Arabic sentiment analysis area. Several reasons may have explained the lack of studies in this area. Assiri in [18] mentioned two main reasons: 1- limited research funding in this area, 2- Arabic has a very complex morphology relative to the morphology of other languages. The complexity and variety of Arabic dialects require advanced pre-processing and lexicon-building procedures [12, 106, 18].

Working in this area needs a full understanding of Arabic standard layer-based structure of linguistic phenomena such as phonology, morphology, syntax and semantics [103]. Arabic is a highly inflectional and derivational language with many word forms and diacritics. Several suffixes, affixes, and prefixes in Arabic words make it harder for lexicon or morphological analyzers to extract the root of words correctly[37].

MSA has more studies and analysis as compared to DA. Numerous tools for detecting sentiments on short or long texts in MSA have been built. Knowing that applying NLP tools designed for MSA directly to DA yields significantly lower performance. This led a group of researchers to build other resources and tools for analyzing DA [106, 29, 98].

Many researchers have applied Machine Translation (MT) in their studies by translating Arabic statements into English and then applying sentiment analysis tools on the translated materials[80][99]. This approach has been explored widely for other foreign languages by performing sentiment analysis on the English translation [21] [11]. The problem of this approach was the loss of nuance after translating the source to English. It is shown in [13] that finding an Arabic MT that meets human requirements is a difficult task. This field still needs more efforts to be improved. Most of the previous work focused on the translation of news and official texts. Much work has been done on MSA; however, research on DA is still lacking in MT [100].

A prior important step for analyzing sentiments in any language is Building Resources (BR). This step aims at creating lexica, corpora with annotated expressions or opinions. There is a need for large scale of annotated resources for the Arabic

Table 2: Building resources for Arabic

Ref	Name	Year	Description
[95]	OCA	2011	Arabic corpus contains 500 movie reviews collected from different web pages and blogs in Arabic with 250 positive reviews and 250 negative reviews. It has limited size and for specific domain (movies).
[1]	AWATIF	2012	Multi-genre corpus of Modern Standard Arabic labeled for subjectivity and Sentiment Analysis.It is only dedicated for MSA and not available for public.
[81]	LABR	2014	Over 63,000 book reviews rated on a scale of 1 to 5 stars. It is only for specific domain (books).
[2]	SANA	2014	A large scale multi-genre sentiment lexicon (more than 200K) of MSA and some Arab dialects. It is also not public.
[92]	NA	2014	A dataset of 8,868 multi-dialectal Arabic annotated tweets.
[35]	NA	2015	A large multi-domain dataset (33K annotated reviews for movies, hotels, restaurants and products) for sentiment analysis in Arabic.

language in order to do sentiment analysis. Some efforts have been paid to build Arabic Treebanks that contain collections of manually-annotated syntactic analyses of sentences [69] [45] [43]. Researchers focus mainly on building corpus/corpora that contain annotated data for MSA and less attention is paid towards DA. Most of these resources are either of limited size or not available for public. Recently, a study [35] addressed this problem and generated a large multi-domain dataset for sentiment analysis in Arabic. The study scraped 33K annotated reviews for movies, hotels, restaurants and products. Then, the researchers built multi-domain lexicons from the generated datasets and tested the classifier models on this data. Another research published in 2014 [92] with a dataset of 8,868 multi-dialectal Arabic annotated tweets. They employed morphological features, simple syntactic features, such as n-grams, as well as semantic features. Other research studies can be found in Table 2 which summarizes the recent work on building resources for Arabic language.

Most of the sentiment analysis tools perform three main data pre-processing steps before applying the classification techniques in order to prepare the Arabic texts, which are:

1. Normalization: the process of transforming the text in order to be consistent by converting all the various forms of a word/letter to a common form. The normalization conditions for Arabic includes the following:

    (a) Remove punctuation marks from the word.

    (b) Remove any diacritics (short vowels) from the word.

    (c) Remove non-letters or symbols from the word.

    (d) Replace similar letters that are used interchangeably by one of them. (example: the letter (A) in Arabic (أ) can be written as (أ), (إ), (آ),(ء). Replace all to (ا)).

2. Stemming: the process of reducing derived or inflected words to their stem, base, or root form. This is accomplished by removing suffixes, prefixes, and infixes.

3. Stop words removal: the process of removing words that have very little meaning. (example: من , حتى, مثل)

Once data pre-processing has been applied to the text, it will be ready for the feature extraction step. Several text features are considered for the Arabic sentiment analysis: n-grams, term presence or its frequency, part-of-speech, or emoticon symbols. The goal of feature extraction step is to select which text features are best to be applied in sentiment analysis tool. Most of the features in Arabic sentiment analysis are classified into three types[9]: (1) Syntactic, which includes: word/POS tag n-grams, phrase patterns, punctuation, (2) Semantic, this type includes: polarity tags, appraisal groups, semantic orientation, and (3) Stylistic, which is concerned with lexical and structural measures of style.

A considerable amount of previous work has been published on Arbic sentiment analysis. This literature review is focused on the studies that categorized the Arabic tweets into specific domains using different classification techniques. Table 3 presents and summarizes the latest work in this area according to the classification techniques and extracted features. It also states whether the study has been applied to MSA or DA.

In 2012, a study [106] proposed a model that used two machine learning approaches, NB and SVM. The researchers used a list of stop words from Egyptian dialect in the preprocessing step. They selected 1000 tweets that hold only one opinion, not sarcastic, subjective and from different topics. SAMAR, is another proposed tool in the same year [3]. It is also a machine learning system for Arabic social media texts. The researchers tested their tool in four different genres: chat, Twitter, Web forums, and Wikipedia talk pages. For Twitter, a corpus of 3015 Arabic tweets has been collected that has a mixture of MSA and DA.

Table 3: Analysis of previous work on Arabic sentiment analysis for Twitter data

Ref	Year	Tweets	Features	Classification Techniques	Performance	MSA or DA
[106]	2012	1000	Unigrams and Bigrams	NB and SVM	SVM Acc 72.6%	Egyptian DA
[3]	2012	3015	Morphological, POS tags, and adjective polarity lexicon	SVM-light	Acc of 71.85%	MSA and DA
[9]	2013	4000	N-Grams and unigrams	SVM, NB, MaxEnt, Bayes Net, and J48 D-tree	SVM Acc 86.38%	-
[80]	2013	2300	Stem-level, Sentence-level, and tweet specific	NB	Acc 80.6%	MSA and DA
[4]	2013	2000	Unigram	supervised (SVM, NB, KNN, and D-tree) and unsupervised ML	Acc 87.5%	MSA and Jordanian DA
[33]	2013	500	POS tag with weight	Unsupervised approach	Acc 83.8%	Egyptian DA
[29]	2014	25000	-	NB, k-NN, SVM	NB (76.78%)	MSA and Jordanian DA
[98]	2014	340,000	Opinions-Oriented words extraction	D-tree and SVM	Prec 76%, recall 61%	Kuwaiti DA
[10]	2015	900	Giving weights to words	Unsupervised	Acc 86.89%	MSA
[50]	2015	1000	linguistically and syntactically motivated	Semi-supervised with SVM	Acc 95%	MSA and DA

Next year, 2013, a new study [9] annotated 4000 tweets from different popular topics: technology, politics, religion, and sports, respectively. The study found that it is better to use unigrams with tweets. Another study has also been presented in 2013 [80] that built a baseline system for performing subjectivity and sentiment analysis for Arabic news and tweets. MT has been employed to translate an existing English subjectivity lexicon to build large coverage lexicons in Arabic. Another study in 2013 [4] addressed both approaches; supervised and unsupervised, for sentiment analysis for Arabic twitter data. The researchers in this study collected and labeled 2000 tweets in both MSA and Jordanian dialect. One of the key finding of this study was that the unsupervised approach gives much lower accuracy compared to the supervised approach. A group of researchers constructed a lexicon-based tool to analyze sentiments of egyptian dialectical tweets in 2013 [33]. Every word in the lexicon has been assigned weights that determined semantic orientation based on the sentiment lexicon.

In 2014, an Arabic sentiment analysis tool was presented in [29] which contains a lexicon that maps Jordanian Dialect to MSA, a lexicon that maps Arabizi words to MSA, and a lexicon of emoticons. In the same year, SVM classifier has also been tested on a corpus of 340,000 tweets in Kuwait[98]. This system handeled Kuwaiti dialect which used Opinions-Oriented words extraction features to extract the opinion-oriented words through language resources that they have been developed for the Kuwaiti dialect.

Recently in 2015, another tool that used an unsupervised (lexicon-based) approach has been introduced [10]. This tool has access to a sentiment lexicon that contains a set of words along with their sentiment values. A sentiment lexicon of about 120,000 Arabic terms has been constructed through three steps: collect Arabic stems, translate them into English, and use online English sentiment lexicons to determine the sentiment value of each word. They stated that the proposed tool performed better than the keyword-based approach.

Finally, a research [50] studied an Arabic idioms/saying phrases lexicon to improve the sentiment polarity in Arabic sentences has gained a high accuracy around 95%. This study used semi-supervised approach with using SVM classifier to analyze MSA and Egyptian dialectal Arabic tweets and microblogs, such as hotel reservation, and product reviews.

2.4 Conclusion

This chapter has presented the challenging task of sentiment analysis and opinion mining on Twitter data in the domain of the Arabic language. We reviewed numerous studies that analyzed people's opinions in English and other Indo-European languages. However, we found few studies that analyzed people's opinions in the Arabic language.

This current investigation examined the prior studies to determine how the sentiment analysis was applied to a high volume of Arabic tweets. This study aimed to help newcomers to this field understand the different aspects posed by the research within the past few years. A sophisticated categorization of a large number of recent articles has been reviewed in this study to cover a wide variety of sentiment analysis in the Arabic language.

One of the main findings of this review shows that there is still a great need for extensive research to gain a better understanding of Arabic dialects in addition to further MSA studies. Up to the time of writing this literature review, no single system existed that could handle all Arabic dialects and MSA with high accuracy. This has created a wide gap in this field for researchers to address in subsequent investigations.

This study demonstrated a need for building and publishing additional lexicon Arabic resources with different genres and various dialects for both the public and research community. Assembling all lexicons for Arabic dialects from different geographical areas in the Middle East in one lexicon repository is a worthy goal.

Recently, growing Internet usage has produced a new written form called Arabizi. This type of the Arabic language is derived from the spoken Arabic dialects and is written using Latin letters and numbers. Detecting and analyzing tweets written in Arabizi has not been thoroughly studied. Knowing that Arabizi has been used widely by teen-agers, it is important to conduct future studies on this type of language and to include young researchers and annotators to bridge the gap.

CHAPTER 3: CASE STUDY ON TWITTER SENTIMENT AND EMOTION DETECTION

"Sentiments, as I have found, can be harvested from places where our memories are fondest."

–Fennel Hudson

3.1 Introduction

The American political system is commonly called a two-party system because most of the candidates who compete for offices come from the two major parties that dominate the system. The Democratic Party has liberal values, while the Republican Party has conservative values [54]. The election of the president of the United States occurs every four years. Therefore, it has become traditional for the two parties to engage in debates during the presidential election as a formal contest of argumentation. These debates are broadcasted live on television and radio with a broad audience not only in the US, but also worldwide [25, 89]. There were a series of scheduled debates for the 2016 presidential election. Then-candidate Donald Trump was involved in eleven Republican primary debates.

Social media played a vital role in revealing the pulse of the US after every primary debate. People and the media were absolutely obsessed with the two parties' debates and the primary campaigns. Twitter was one of the most powerful media platforms that was used by news stations and politicians alike. Many users posted their opinions and discussed political issues on Twitter. Users expressed positive or negative tweets in order to reflect their satisfaction or complaints [8].

Tweets are a unique source of information regarding the election. It has been noticed that, among the longest running Democratic and Republican candidates, Trump was the most mentioned on Twitter. Trump's business career, branding efforts, lifestyle and use of the media helped make him a celebrity, a status strengthened as superstar in the reality TV-show, "The Apprentice" [55].

In this chapter, we aim to perform sentiment analysis on people's tweets to study their reactions and emotions regarding Trump's primary debates. To achieve this goal, the study has mined a large number of people's tweets for every Republican primary debate that Trump was engaged in. The crawling process took almost three hours, starting from the beginning of each debate. The total number of the retrieved tweets is 85,000. To be unbiased, the study has used different hashtags for/or against Trump's campaign, such as #voteTrump and #antiTrump. The study shows the polarity and emotional results present in the tweets, and it also identifies the most frequently used keywords in people's tweets. One of the key findings of this study is identifying the percentage of the expressed emotions in the tweets that support Trump.

The remainder of this chapter is organized as follows: Section 3.2 gives the reader general background material on previous works that talk about Twitter and political sentiment analyses; section 3.3 examines the technique used to analyze the tweets in this study; section 3.4 summarizes the key findings of the study; and section 3.5 concludes with future directions for this research.

3.2 Related Work

Social media platforms, such as Facebook or Twitter, are easy to use, which drives people to share their thoughts and express their opinions. Many large companies and political offices poll tweets from Twitter microblogs to know the attitude of people towards certain decisions or mainstream issues. Moreover, researchers and scientists in different fields considered this a promising topic. They work towards improving their tools for analyzing people's reactions to their respective fields [102, 8]. Whenever people need to make a decision, they want to know others' opinions. For example, individual consumers want to know the opinions of existing users before purchasing a product. Likewise, voters want to know the opinions of other voters when deciding how to cast a ballot in any election. Similarly, it has been shown that tracking and analyzing public opinions from social media may help in predicting certain political events [96]. Strong studies have been published concerning this topic under the name of sentiment analysis or opinion mining.

Sentiment analysis is considered a part of the Natural Language Processing (NLP) field. It was first explored in 2003 by Nasukawa and Yi [82]. In sentiment analysis of Twitter data, the researchers focus their studies on subjective, not objective, tweets. They are interested mainly in classifying tweets as positive or negative [56] [5]. The researchers investigated the usability of linguistic features to detect the sentiment of Twitter posts. Some researchers label the tweet as positive if it ends with positive emoticons and as negative if it ends with negative emoticons. Several approaches to detect emotions have been presented in previous research papers [87, 68, 20]. One of these is a web-based text mining approach for detecting the emotions of an individual event embedded in English sentences. In this approach, the researchers proved that the emotion-sensing problem was context sensitive.

Several studies have performed sentiment analysis using different techniques. According to Boiy [22], there are two basic methodologies used in sentiment analysis for text: symbolic and machine learning techniques. The symbolic approach (which is also called lexicon-based) uses manually crafted rules and lexicons to determine the sentiment of every word and to combine these values with some aggregation function; the machine learning approach uses supervised or unsupervised learning techniques to build a model from training data. The current study uses sentiment lexicons, which are lists of words or phrases associated with positive and negative sentiments, then applies the Bayes model to determine the overall sentiment for each tweet.

Many researchers and media have investigated the prediction of public opinion and consequently have predicted the results of political events, such as the US presidential election. A study in 2012 [105] indicated that social media could be used to predict public opinions regarding the election and may replace traditional polling. They analyzed millions of tweets from September 2011 leading up to the Republican primary elections. Another study by Tumasjan et al. [113] on the German federal election has reported positively about the use of microblogging message content as a valid indicator of political sentiment. They analyzed 10,000 messages that contained a reference to politicians and found that the great amount of messages mentioning a party reflects the election result. On the other side, Gayo-Avello et al. [39] argued not to accept the social media predictions about political events unless if it is accompanied by a model explaining the predictive power of social media. Another study in 2016 [24] aimed to analyze how the public views the top presidentially candidates, namely Trump, Hillary Clinton, Ben Carson, and Bernie Sanders. They classified the tweets that targeted these candidates into five emotions: happy, sad, fear, laughter, and anger. They only considered tweets with emojis to classify the tweets. Anuta et al. [15] found that both the polls and Twitter were biased in the 2016 US presidential election. The interesting part is they found that the polls had a bias towards Clinton, while Twitter had a bias towards Trump.

3.3 Methodology

The process of performing sentiment analysis for social media platform tools usually proceeds through multiple phases [96, 12]:

3.3.1 Phase 1: Data-gathering.

Crawling data from social media highly relies on Application Programming Interfaces (API) that are provided by the social media platform itself. The Twitter microblogging platform provides a streaming API to extract tweets in realtime. Twitter uses OAuth to provide authorized access to its API. This helps researchers to retrieve the required tweets using filtering strategies on specific keywords or phrases. In this phase, the required amounts of tweets that are related to the current study are retrieved. The data was filtered according to a particular time frame and specific keywords, and the events zone and the language were also specified.

Table 4: Crawling tweets talking about Donald Trump in Republican debates

Debates	Date	No.Tweets
Debate 1	August 6, 2015	9000
Debate 2	September 16, 2015	9000
Debate 3	October 28, 2015	4000
Debate 4	November 10, 2015	4000
Debate 5	December 15, 2015	8000
Debate 6	January 14, 2016	5000
Debate 7	February 6, 2016	6000
Debate 8	February 13, 2016	8000
Debate 9	February 25, 2016	11000
Debate 10	March 3, 2016	12000
Debate 11	March 10, 2016	9000

In order to perform this study, we extracted a random sample of tweets from the Twitter database that reflected people's sentiments regarding Trump. In order to filter the tweets, we used the following keywords: trump2016, #antitrump, #trump2016, #DonaldTrump, donaldtrump, donald_trump, #voteTrump, #trump, and any other tweet that contained donaldtrump. The crawling data started at the beginning of each debate and ended one hour after each debate's conclusion. Only Republican primary debates in which Trump participated were involved in this process (We disregarded the Republican debate on January 28, 2016 since Trump did not participate.). Also, only English tweets that originated from the US were mined. It is worth mentioning that the tweet required a hashtag, and we avoided retweets. Table 4 shows the number of tweets that were collected for the Republican primary debates.

3.3.2 Phase 2: Data-preprocessing (text normalization).

Data-preprocessing is an important step in the data mining field. The retrieved data from the first phase were tokenized by converting the sentences into words. These words were cleaned to remove any irrelevant and redundant information. Also, this phase eliminated all noisy and unreliable data by deleting white-spaces, punctuation marks, numbers, and URL links; furthermore, all uppercase letters were converted to lowercase to make it easier to compare words in the next phase (building classifier model).

Example of Tweet before cleaning:

"@user we have such powerful talented people on stage who genuinely love america. They should all have cabinet positions. #trump2016"

Example of Tweet after cleaning:

"we have such powerful talented people on stage who genuinely love america they should all have cabinet positions trump"

3.3.3 Phase 3: Building a Classifier Model

In building a classifier model, the cleaned data from the original tweets were scored and classified by polarity (positive or negative) and by Ekman emotions [32] (joy, sadness, fear, anger, surprise, or disgust). To achieve this, we used the R package titled "Sentiment." Two lists of negative and positive opinion words (around 6800 words)[49] had been used to score the tweets for detecting polarity. Also, a list of 1542 emotional words had been used from the same package to classify the tweets to the six listed emotions. More than 78% of the tweets had been labeled as unknow emotion or neutral polarity, or both. This study used only 22% of the tweets that had been labeled with both emotion and polarity. Figure 4 and Figure 5 show the results of polarity and emotion classifications, respectively.

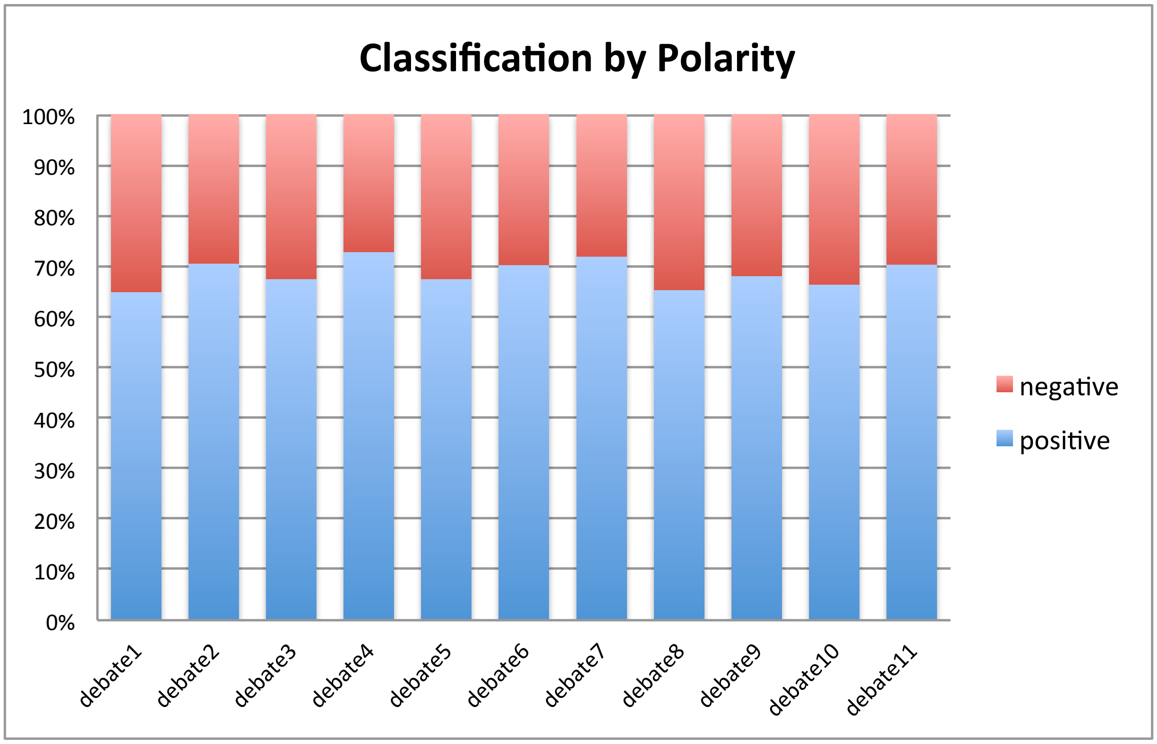


Figure 4: Polarity classification for tweets

3.3.4 Phase 4: Visualization.

The goal of the visualization phase is to represent the results of the sentiment analysis graphically. In this section, we showed the preliminary results for the sentiment analysis of Twitter data that had been extracted for people's tweets about Trump after every debate. Figure 4 shows the percentage of people's emotions for every debate. We could tell from the figure that more than half of the tweets in every debate expressed joy. Specifically, debate 5 had the highest joy tweets (62%), while debate 11 had the lowest (52%). The second largest emotion measured was sadness.

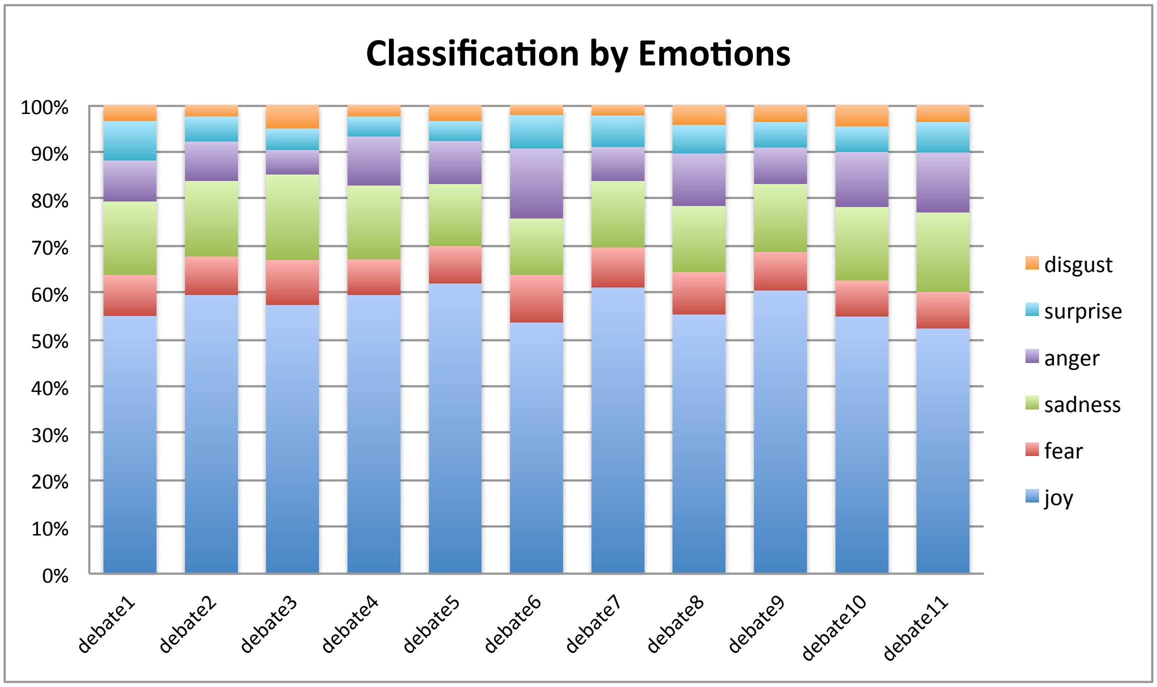


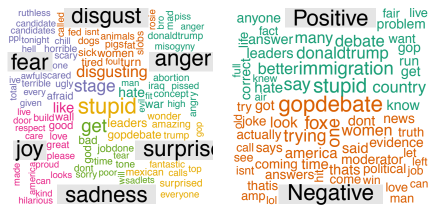
Figure 5: Emotion classification for tweets

The study showed that the highest percentage of sadness tweets occurred in debate 3 (18%), while the lowest occurred in debate 6 (12%). The figure showed the other emotions and the percentages per debate.

Word clouds give a clear perceptive about what people are tweeting. In this study, we plotted word clouds to compare the frequencies of each emotion for each debate. Figure 6 shows the highest word frequencies for each group of emotions and polarities for each debate.

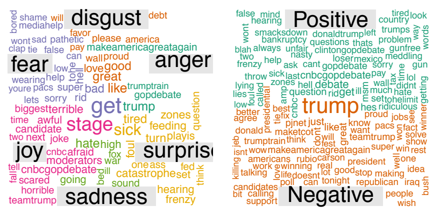
3.4 Discussion

This study is not a comparison of emotional tweets from supporters of the presidential candidates. Rather, it investigates people's emotions regarding then-candidate Trump after his debates. The model that has been used to detect emotions and polarities has labeled only 22% of the tweets with the corresponding emotions. One

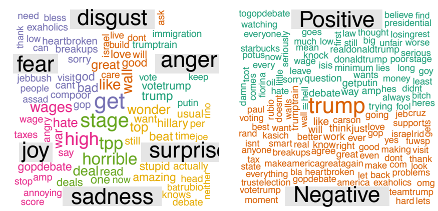


Emotional and Polarity WordClouds for Debate 1

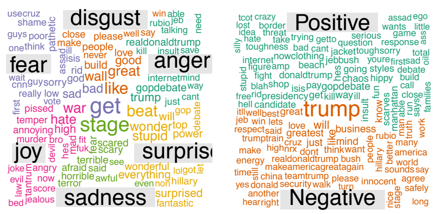
Emotional and Polarity WordClouds for Debate 2



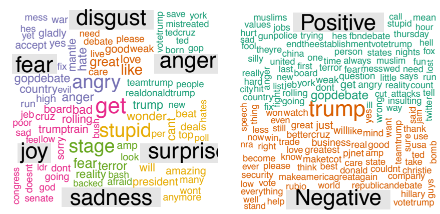
Emotional and Polarity WordClouds for Debate 3



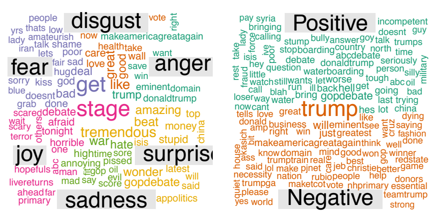
Emotional and Polarity WordClouds for Debate 4



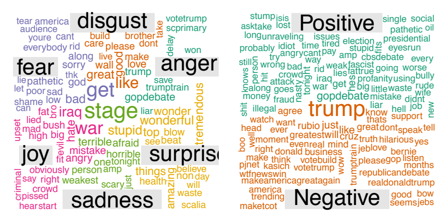
Emotional and Polarity WordClouds for Debate 5



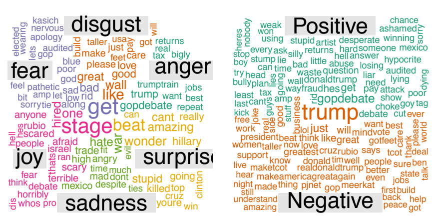
Emotional and Polarity WordClouds for Debate 6



Emotional and Polarity WordClouds for Debate 7

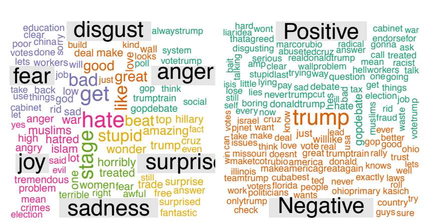


Emotional and Polarity WordClouds for Debate 8



Emotional and Polarity WordClouds for Debate 9

Emotional and Polarity WordClouds for Debate 10



Emotional and Polarity WordClouds for Debate 11

Figure 6: WordClouds for each group of emotions and polarities for each debate

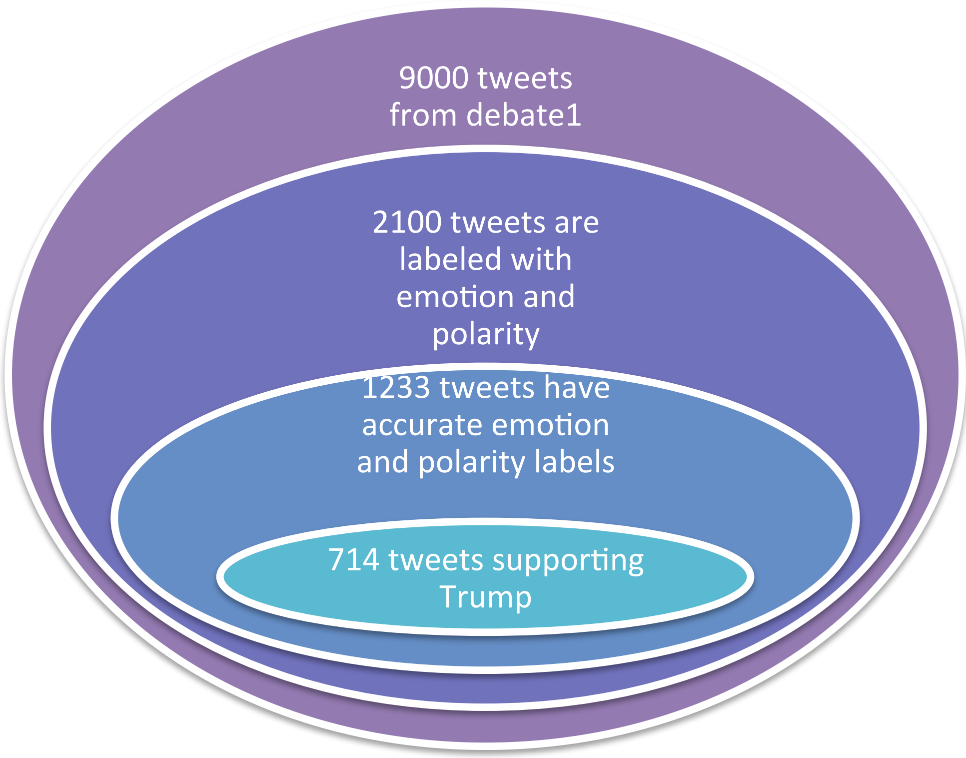


Figure 7: Stacked Venn diagram for Debate1 tweets in validation process

of the reasons behind this is the small list of "emotion" words that is used by this selected model. Unfortunately, 78% of the tweets have been labeled as an "unknow" emotion, a "neutral" polarity, or both. Of the 9000 tweets from debate1, 2100 that have emotion and polarity labels have been annotated to validate the model. Figure 7 shows a stacked Venn diagram of the tweets that have been included in the validation process. The study found that the accuracy of emotion labels was 0.68, and the accuracy of polarity labels was 0.77. For a complete depiction of the accuracy of all emotion labels, see Figure 8.

Also, one of the key findings of this study was that 48% of the 2100 tweets supported Trump. However, it was found that, of the accurate emotion/polarity detected tweets (1233 tweets), 58% (714) supported Trump, as shown in Figure 9. Upon fur-

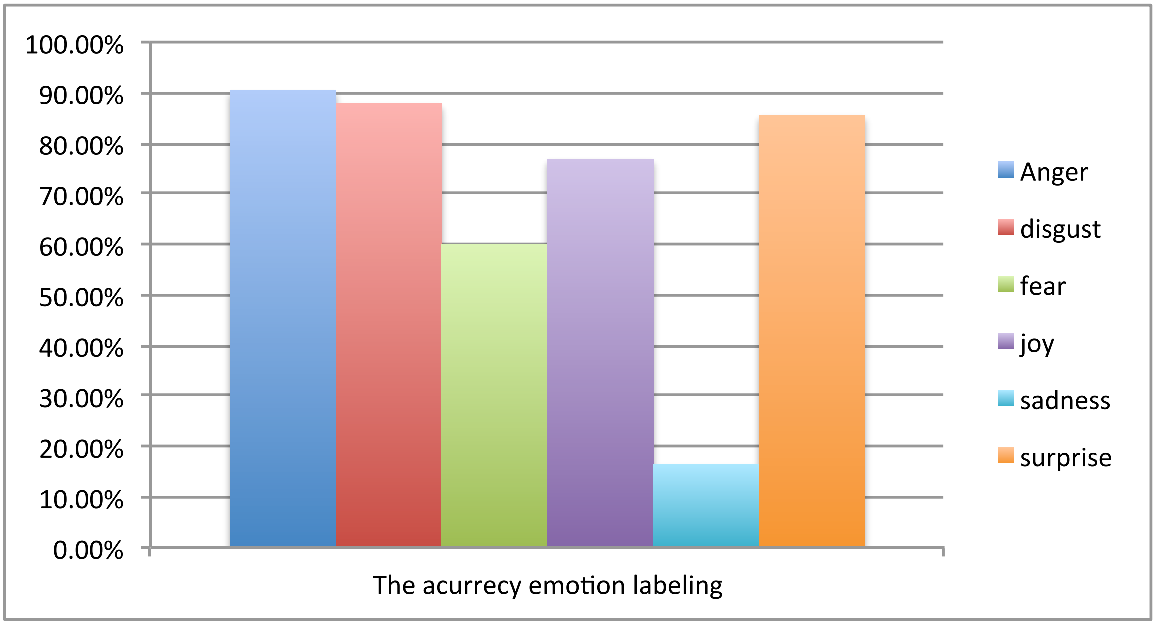


Figure 8: The accuracy percentage of detecting people's emotions

ther investigation, it was clear that Joy was the most dominant emotion in the tweets that supported Trump. Figure 10 shows more details about the percentage of emotions in the tweets that supported Trump. Figure 11 shows the percentage of the expressed emotions in the tweets that supported Trump. It is worth mentioning that about 64% of tweets that expressed Fear supported Trump. To make it clear for the reader, we can tell from this Anger tweet, "*we all hate you @realdonaldtrump #gopdebate*" the tweet held Anger for Trump. However, in this tweet "*@foxnews i'm very angry all questions put to #trump were framed as attacks,not to get his positions on the issues.very unprofessional of fox*", it supported Trump even if it held an Anger emotion. Therefore, a tweet with anger emotion that mentioned Trump could not be considered against this candidate.

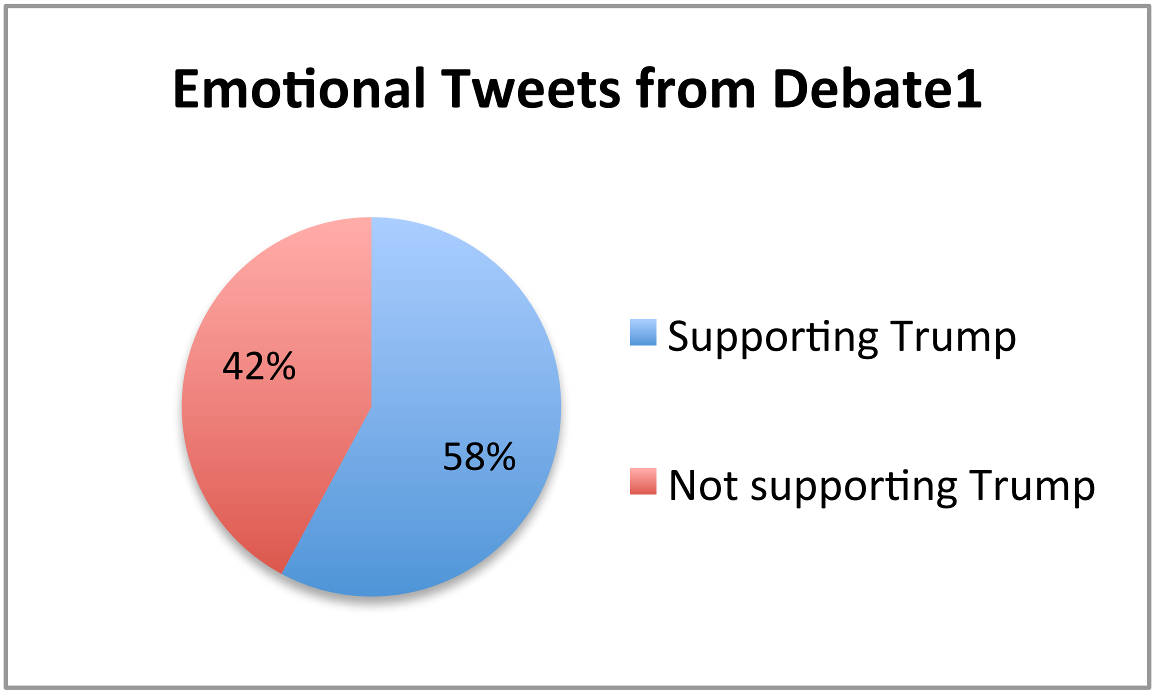


Figure 9: Emotional tweets with support or non support of Trump in debate1

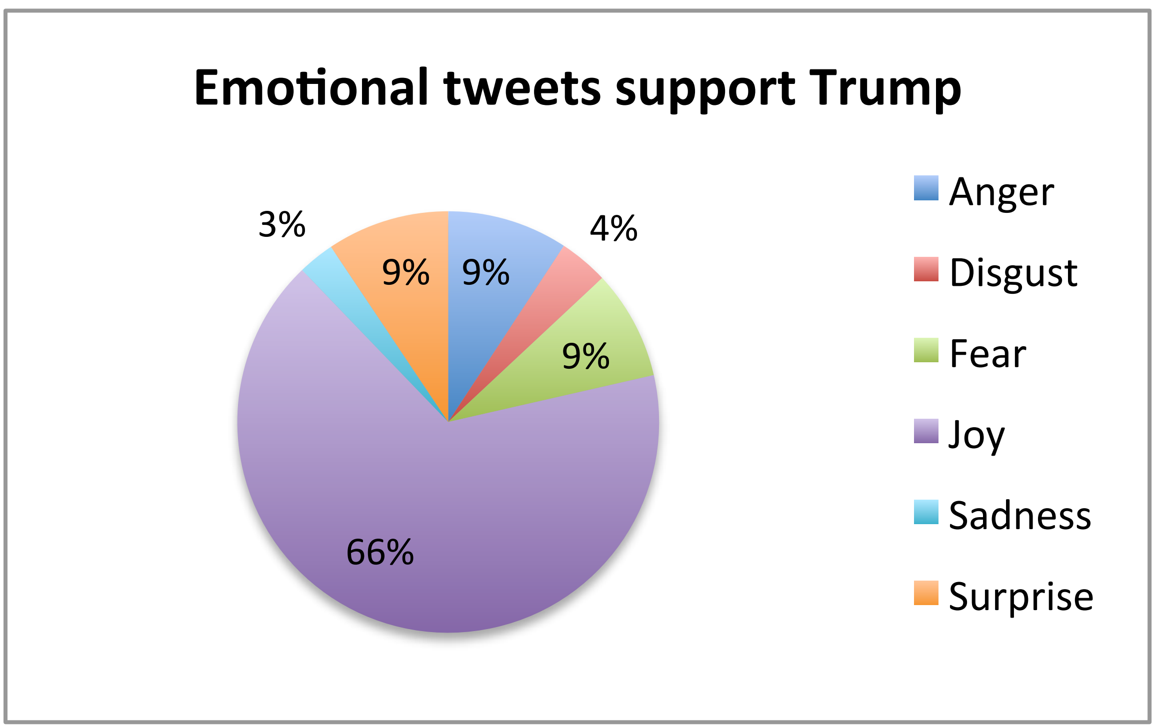


Figure 10: Percentage of emotions for Trump supporters (N=714)

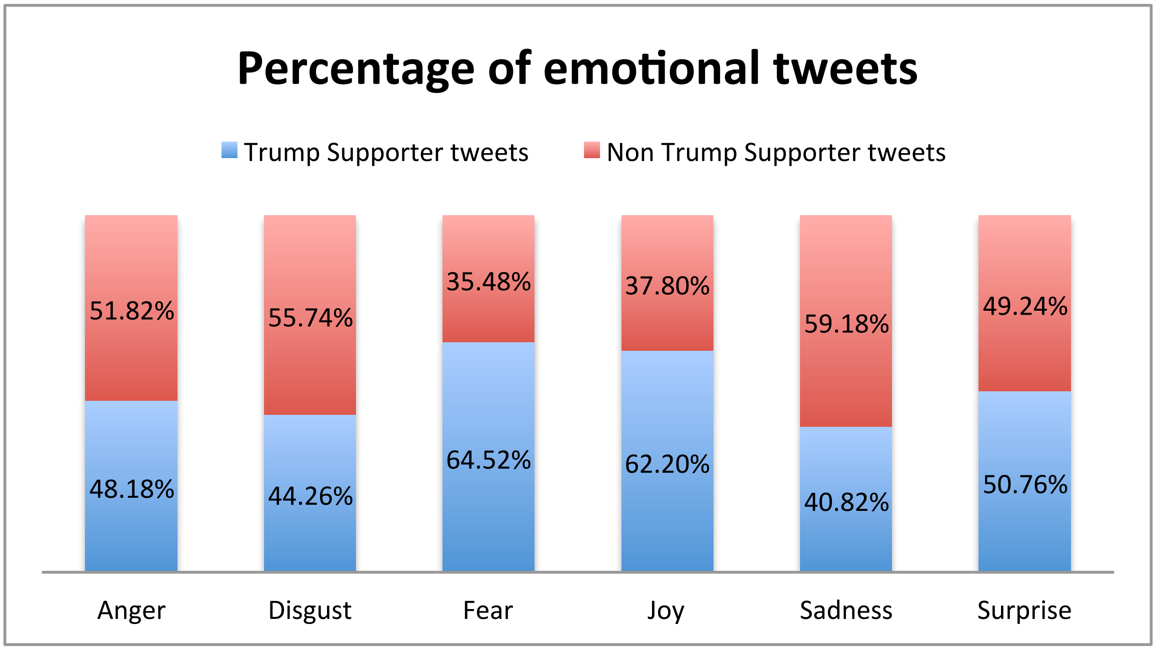


Figure 11: Percentage of each emotion for entire dataset (N=1233)

3.5 Conclusion

This study collected a dataset of political tweets in which Donald Trump was mentioned in his primary debates. Overall, 85,000 tweets were preprocessed and cleaned in order to detect polarities and emotions. To validate the results, a sample of 2100 tweets related to debate1 were annotated to detect polarity, emotions, and support of Trump. It was found that mentioning a candidate in a tweet did not imply that the tweet supported this candidate. Moreover, negative or positive polarities in the tweets also was not a good indicator to determine support for this candidate. Of the tweets that supported Trump, we found about 39% of the tweets had a negative polarity. In addition, detecting emotions in the tweets was not enough to predict the number of voters or supporters. In other words, the anger emotion in the tweet did not always imply that the tweet was against this candidate. Anger can apply to a feeling against Trump himself or against other people who opposed this candidate.

This demonstrates a need for detecting emotions and tagging the entity that lead to this emotion. In a future work, more investigation on the collected tweets should be applied to predict whether or not an emotional tweet supports any candidate. Also, we need to direct our research toward building a more accurate model for detecting emotions in political tweets and for classifying whether the emotion is related to the candidate or to the opponents.

CHAPTER 4: SENTIMENT AND EMOTION DETECTION IN ENGLISH AND ARABIC TWEETS USING DEEP LEARNING

"Your emotions are the slaves to your thoughts,

and you are the slave to your emotions."

–Elizabeth Gilbert

4.1 Introduction

SemEval is the International Workshop on Semantic Evaluation that has evolved from SensEval. The purpose of this workshop is to evaluate semantic analysis systems, the SemEval-2018 being the 12^{th} workshop on semantic evaluation. Task 1 [75] in this workshop presents five subtasks with annotated datasets for English, Arabic, and Spanish tweets. The task for participating teams is to determine the intensity of emotions in text. Further details about Task 1 and the datasets appear in Section 4.3.

This chapter proposes our system "TeamUNCC" to detect sentiments and emotions in English and Arabic tweets. Our system covers five subtasks for both English and Arabic. The input to the system are word embedding vectors [71], which are applied to fully connected neural network architecture to obtain the results. In addition, all subtasks except the last one, use document-level embeddings doc2vec [65] that are concatenated with different feature vectors. The models built for detecting emotions related to Arabic tweets ranked third in subtask El-oc and fourth in the other subtasks. We use both the original Arabic tweets as well as translated tweets (to English) as input. The performance of the system for all subtasks in both languages shows substantial improvements in Spearman correlation scores over the baseline models provided by Task 1 organizers, ranging from 0.03 to 0.23.

The remainder of this chapter is organized as follows: Section 4.2 gives a brief overview of existing work on social media emotion and sentiment analyses, including for English and Arabic languages. Section 4.3 presents the requirements of SemEval-2018 Task1 and the provided datasets. Section 4.4 examines the proposed system to determine the presence and intensity of emotion in text. Section 4.5 summarizes the key findings of the study and the evaluations. Section 4.6 concludes with future directions for this research.

4.2 Related work

Sentiment and Emotion Analysis: Sentiment analysis was first explored in 2003 by Nasukawa and Yi [82]. An interest in studying and building models for sentiment analysis and emotion detection for social microblogging platforms has increased significantly in recent years [62, 85, 84, 51]. Going beyond the task of mainly classifying tweets as positive or negative, several approaches to detect emotions were presented in previous research papers [76, 112, 73]. Researchers [74] introduced the WASSA-2017 shared task of detecting the intensity of emotion felt by the speaker of a tweet. The state-of-the-art system in that competition [40] used an approach of ensembling three different deep neural network-based models, representing tweets as word2vec embedding vectors. In our system, we add doc2vec embedding vectors and classify tweets to ordinal classes of emotions as well as multi-class labeling of emotions.

Arabic Emotion Analysis: The growth of the Arabic language on social microblogging platforms, especially on Twitter, and the significant role of the Arab region in international politics and in the global economy have led researchers to investigate the area of mining and analyzing sentiments and emotions of Arabic tweets [5, 34, 17]. The challenges that face researchers in this area can be classified under two main areas: a lack of annotated resources and the challenges of the Arabic language's complex morphology relative to other languages [18]. Although recent research has been dedicated to detect emotions for English content, to our knowledge, there are few studies for Arabic content. Researchers [90] collected and annotated data and applied different preprocessing steps related to the Arabic language. They also used a simplification of the SVM (known as SMO) and the NaiveBayes classifiers. Another two related works [59, 94] shared different tasks to identify the overall sentiments of the tweets or phrases taken from tweets in both English and Arabic. Our work uses the state-of-the-art approaches of deep learning and word/doc embedding.

4.3 Datasets Description

SemEval-2018 Task 1, Affect in Tweets, presents five subtasks (El-reg, El-oc, V-reg, V-oc, and E-c.) The subtasks provide training and testing for Twitter datasets in the English, Arabic, and Spanish languages [77]. Task 1 mainly asks the participants to predict the intensity of emotions and sentiments in the testing datasets. It also includes a multi-label emotion classification subtask for tweets. Our work focuses on

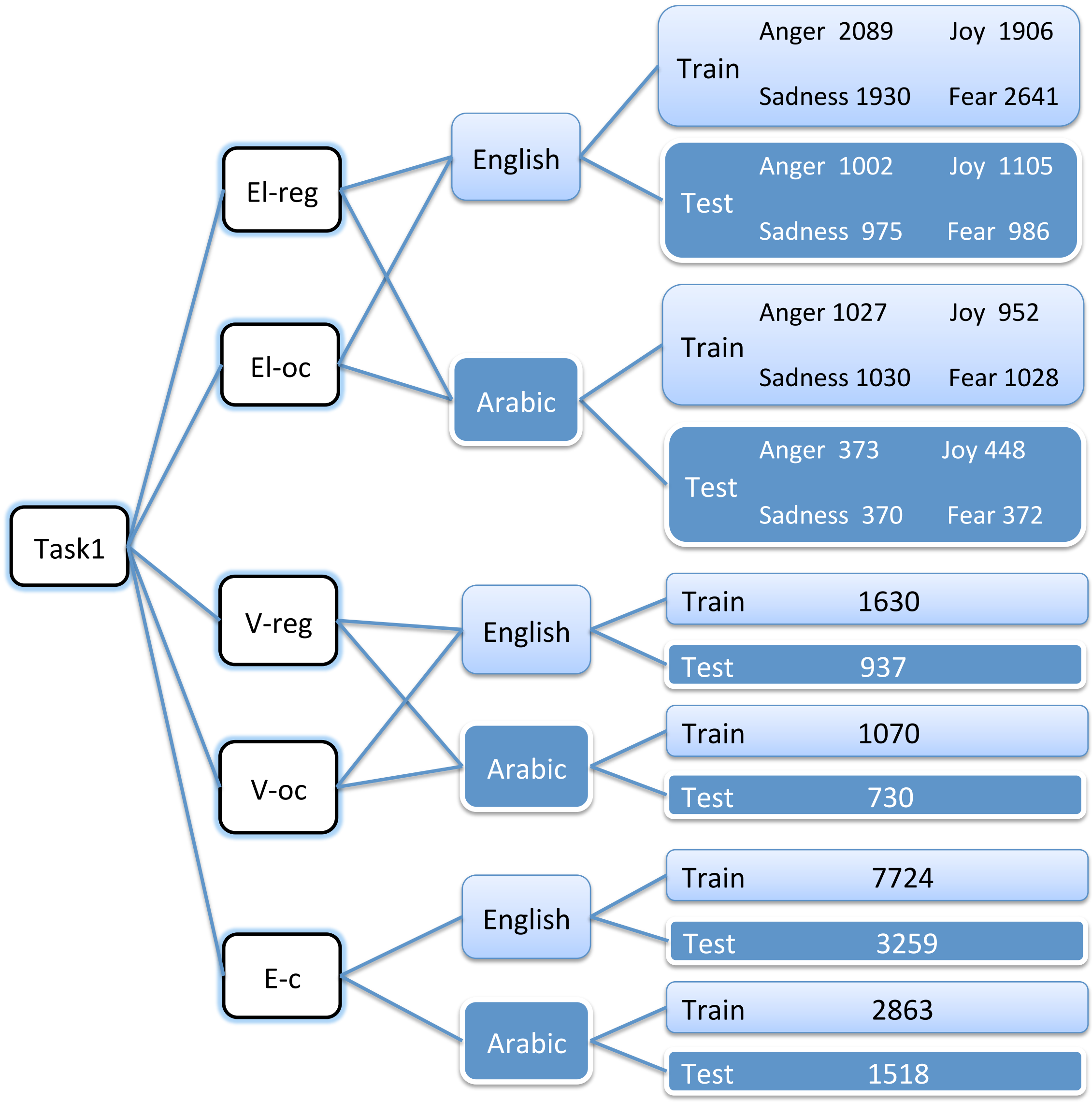


Figure 12: Datasets of SemEval-2018 Task 1

determining emotions in English and Arabic tweets. Figure 12 shows the number of tweets for both training and testing datasets for individual subtasks. We note that subtasks *El-reg* and *El-oc* share the same datasets with different annotations, and the same for subtasks *V-reg* and *V-oc*.

The description of each subtask is:

EI-reg: Determine the intensity of an emotion in a tweet as a real-valued score between 0 (least emotion intensity) and 1 (most emotion intensity).

EI-oc: Classify the intensity of emotion (anger, joy, fear, or sadness) in the tweet into one of four ordinal classes (0: no emotion, 1, 2, and 3 high emotion).

V-reg: Determine the intensity of sentiment or valence (V) in a tweet as a real-valued score between 0 (most negative) and 1 (most positive).

V-oc: Classify the sentiment intensity of a tweet into one of seven ordinal classes, corresponding to various levels of positive and negative sentiment intensity (3: very positive mental state can be inferred, 2, 1, 0, -1, -2, and -3: very negative mental state can be inferred)

E-c: Classify the tweet as 'neutral or no emotion' or as one, or more, of eleven given emotions (anger, anticipation, disgust, fear,joy, love, optimism, pessimism, sadness, surprise, and trust).

4.4 Proposed System

Our system is the only system that participated in all subtasks of Task 1 of SemEval-2018 for both English and Arabic tweets. Subtasks *El-reg* and *V-reg* are considered similar because they determine the intensity of an emotion or a sentiment (respectively) in a tweet as a real-valued score. While subtasks *El-oc* and *V-oc* classify the intensity of the emotion or the sentiment (respectively) to ordinal classes. Our system, designed for these subtasks, shares most features and components; however, the fifth subtask, *E-c*, uses fewer of these elements. Figure 16 shows the general structure of the system. More details for the system's components are shown in the following subsections: Section 4.4.1 describes the system's input and prepocessing.

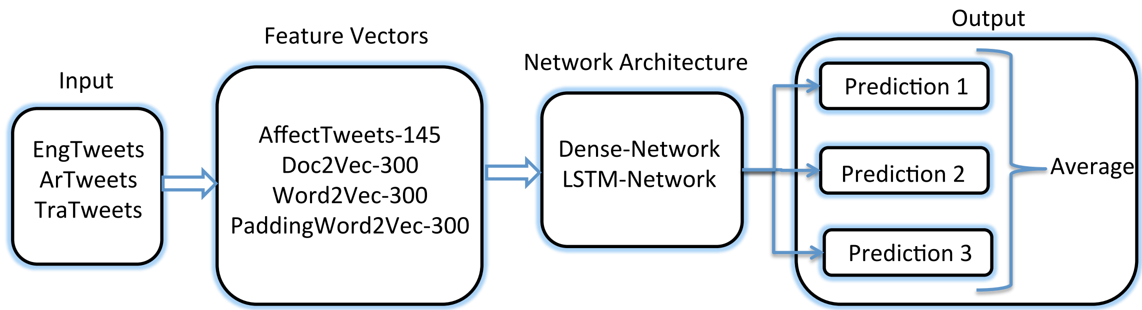


Figure 13: The general structure for the proposed system

Section 4.4.2 lists the feature vectors, and Section 4.4.3 details the architecture of neural network. Section 4.4.4 discusses the output details.

4.4.1 Input

EngTweets: The original English tweets in training and testing datasets have been tokenized by converting the sentences into words, and all uppercase letters have been converted to lowercase. The preprocessing step also includes stemming the words and removal of extraneous white spaces. Punctuation have been treated as individual words (".,?!:;()[]#@'), while contractions (wasn't, aren't) were left untreated.

ArTweets: The original Arabic tweets in training and testing datasets have been tokenized, white spaces have been removed, and the punctuation marks have been treated as individual words (".,?!:;()[]#@').

TraTweets: The Arabic tweets have been translated using a powerful translation tool written in python (translate 3.5.0)[4]. Next, the preprocessing steps that are applied to EngTweets are also applied on TraTweets.

[4]https://pypi.python.org/pypi/translate

Feature Vectors

AffectTweets-145: Each tweet, in either EngTweets or TraTweets, is represented as 145 dimensional vectors by concatenating three vectors obtained from the AffectiveTweets Weka-package [74, 23], 43 features have been extracted using the TweetToLexiconFeatureVector attribute that calculates attributes for a tweet using a variety of lexical resources; two-dimensional vector using the Sentiment strength feature from the same package, and the final 100 dimensional vector is obtained by vectorizing the tweets to embeddings attribute also from the same package.

Doc2Vec-300: Each tweet is represented as a 300 dimensional vector by concatenating two vectors of 150 dimensions each, using the document-level embeddings ('doc2vec') [65, 64]. The vector for each word in the tweet has been averaged to attain a 150 dimensional representation of the tweet.

Word2Vec-300: Each tweet is represented as a 300 dimensional vector using the pretrained word2vec embedding model that is trained on Google News [72], and for Arabic tweets, we use the pretrained embedding model that is trained on Arabic tweets (Twt-SG) [109].

PaddingWord2Vec-300: Each word in a tweet is represented as a 300 dimensional vector. The same pretraind word2vec embedding models that are used in Word2Vec-300 are also used in this feature vector. Each tweet is represented as a vector with a fixed number of rows that equals the maximum length of dataset tweets and a standard 300 columns using padding of zero vectors.

4.4.2 Network Architecture

Dense-Network: The input 445 dimensional vector feeds into a fully connected neural network with three dense hidden layers. The activation function for each layer is RELU [70], with 256, 256, and 80 neurons for each layer, respectively. The output layer consists of one sigmoid neuron, which predicts the intensity of the emotion or the sentiment between 0 and 1. Two dropouts are used in this network (0.3, 0.5) after the first and second layers, respectively. For optimization, we use SGD (Stochastic Gradient Descent) optimizer (lr=0.01, decay=1×10^{-6}, and momentum=0.9) [5], optimizing for 'mse' loss function and 'accuracy' metrics. Early stopping is also applied to obtain best results.

LSTM-Network: The input vector feeds an LSTM of 256 neurons that passes the vector to a fully connected neural network of two hidden layers and two dropouts (0.3, 0.5). The first hidden layer has 256 neurons, while the second layer has 80 neurons. Both layers use the RELU activation function. The output layer consists of one sigmoid neuron, which predicts the intensity of the emotion or the sentiment between 0 and 1. For optimization, we use SGD optimizer (lr=0.01, decay=1×10^{-6}, and momentum=0.9), optimizing for 'mse' loss function and 'accuracy' metrics as well as early stopping to obtain the best results.

4.4.3 Output

Subtasks El-reg, El-oc, V-reg, and V-oc: These four subtasks for each language (English and Arabic) share the same structure as shown in Figure 16, the only difference

[5] https://keras.io/optimizers/

Table 5: The architecture details for English subtasks El-reg, El-oc, V-reg, and V-oc

-	**Prediction 1**	**Prediction2**	**Prediction3**
Input	EngTweets	EngTweets	EngTweets
Feature Vectors	AffectTweets-145 Doc2Vec-300	AffectTweets-145 Word2Vec-300	PaddingWord2Vec-300
Neural Network	Dense-Network	Dense-Network	LSTM-Network

Table 6: The architecture details for Arabic subtasks El-reg, El-oc, V-reg, and V-oc

-	**Prediction 1**	**Prediction2**	**Prediction3**
Input	TraTweets	ArTweets TraTweets	ArTweets
Feature Vectors	AffectTweets-145 Doc2Vec-300	AffectTweets-145 Word2Vec-300	PaddingWord2Vec-300
Neural Network	Dense-Network	Dense-Network	LSTM-Network

is in the output stage. Each subtask passes the tweets to three different models that produces three predictions. See Table 5 and Table 6 for more comprehensive details on how each prediction with English and Arabic language is produced, respectively. The average of the predictions for each tweet is a real-valued number between 0 and 1. This output is considered the final output for both subtasks *El-reg* and *V-reg*, while subtasks *El-oc* and *V-oc* classify this real-valued number to one of the ordinal classes that are shown in Section 3. We note that *El-reg* and *El-oc* shares the same datasets. We also noticed that V-reg and V-oc shares the same dataset. Therefore, we found the ranges of values for each ordinal class by comparing the datasets. Table 7 shows the range of values to obtain the ordinal classes for *El-oc* subtask in English, Table 15 shows the same for *El-oc* subtask in Arabic, and Table 16 shows the for *V-oc* in both English and Arabic.

Table 7: Classify the output to ordinal classes for English El-oc subtask

Output class	**Angry**	**Joy**	**Fear**	**Sadness**
0: no emotion can be inferred	0-0.42	0-0.36	0-0.57	0-0.44
1: low amount of emotion can be inferred	0.42-0.52	0.36-0.53	0.57-0.69	0.44-0.54
2: moderate amount of emotion can be inferred	0.52-0.7	0.53-0.69	0.66-0.79	0.54-0.7
3: high amount of emotion can be inferred	0.7-1	0.69-1	0.79-1	0.7-1

Table 8: Classify the output to ordinal classes for Arabic El-oc subtask

Output class	**Angry**	**Joy**	**Fear**	**Sadness**
0: no emotion can be inferred	0-0.40	0-0.31	0-0.45	0-0.47
1: low amount of emotion can be inferred	0.40-0.55	0.31-0.51	0.45-0.56	0.47-0.54
2: moderate amount of emotion can be inferred	0.55-0.64	0.51-0.75	0.56-0.76	0.54-0.67
3: high amount of emotion can be inferred	0.64-1	0.75-1	0.76-1	0.67-1

Table 9: Classify the output to ordinal classes for English and Arabic V-oc subtasks

Output class	**English Sentiment**	**Arabic Sentiment**
-3: very negative emotional state can be inferred	0-0.23	0-0.20
-2: moderately negative emotional state can be inferred	0.23-0.38	0.20-0.37
-1: slightly negative emotional state can be inferred	0.38-0.43	0.37-0.43
0: neutral or mixed emotional state can be inferred	0.43-0.61	0.43-0.56
1: slightly positive emotional state can be inferred	0.61-0.70	0.56-0.69
2: moderately positive emotional state can be inferred	0.70-0.78	0.69-0.81
3: very positive emotional state can be inferred	0.78-1	0.81-1

Subtask E-c: In this subtask, our system makes only one prediction. See Figure14 for more details on the process of predicting the results. The input is *EngTweets* for English language and ArTweets for Arabic language. We use *Word2Vec-300* as the feature vector with *GoogleNews* for English tweets and *Twt-SG* for Arabic tweets. The network architecture is Dense-Network. This process is applied for each emotion of the eleven emotions: anger, anticipation, disgust, fear, joy, love, optimism, pessimism, sadness, surprise, and trust. The output of each individual tweet is a real-valued number between 0 and 1. This output is normalized to either 1 (contains an emotion) if it is greater than 0.5 or 0 (no emotion) if it is less than 0.5.

4.5 Evaluations and Results

Each participating system in the subtasks *El-reg*, *El-oc*, *V-reg*, and *V-oc*, has been scored by using Spearman correlation score. The subtask *E-c* has been scored by using accuracy metric. Table 14 shows the performance of our system in *E-reg* and *El-oc* with each emotion and the average score for both English and Arabic. Table 7 shows the results for subtasks *V-reg*, *V-oc*, and *E-c*. The performance of our system

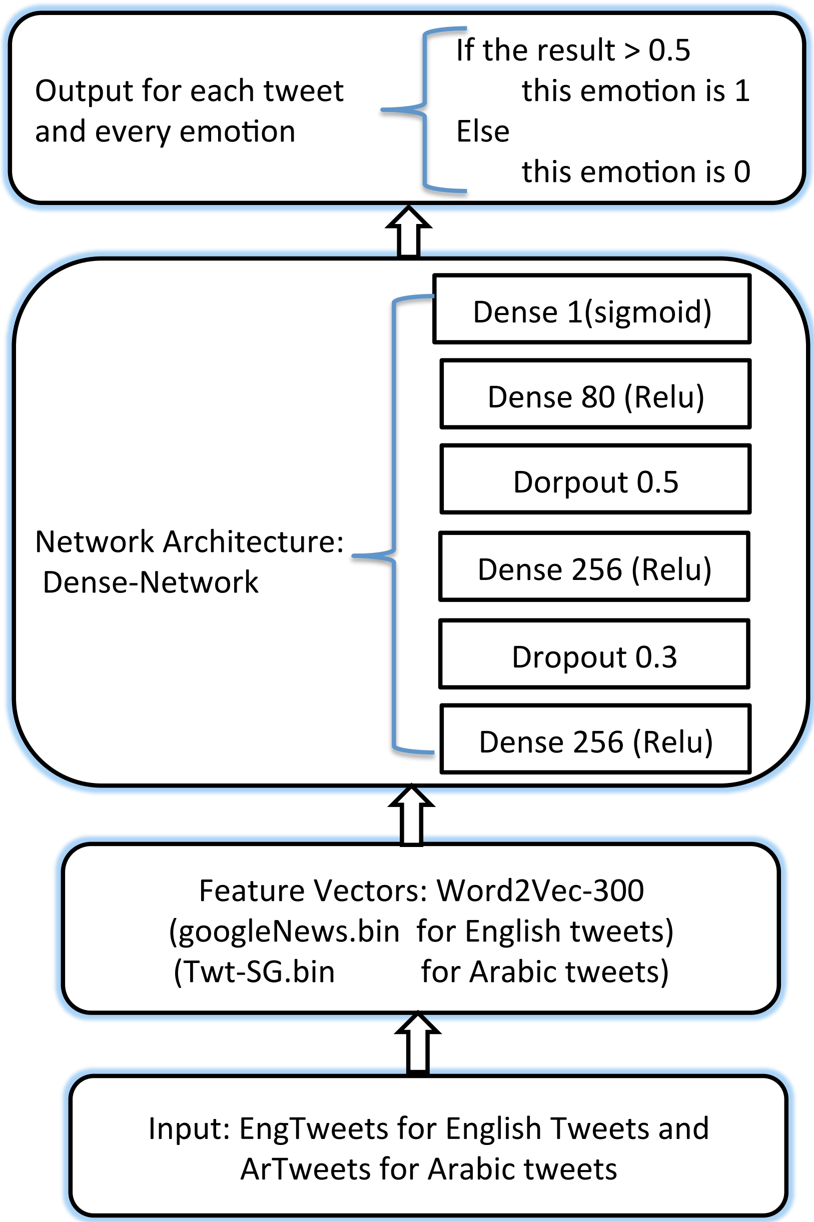


Figure 14: The detailed structure for the proposed system related to subtask E-c

beats the baseline model's performance, which is provided by the Task's organizers, see Figure 17 to capture the difference between the two performances.

The proposed system ranks third in the subtask *El-oc* for Arabic language, and Fourth in the subtasks *El-reg*, *V-reg*, *V-oc*, and *E-c* for Arabic language too. It is worth mentioning that these results have been obtained by using the Task's datasets without using any external data.

Table 10: The Spearman correlation scores for subtasks El-reg and El-oc

Task	Angry	Joy	Fear	Sadness	Average
El-reg (English)	0.722	0.698	0.692	0.666	**0.695**
El-reg (Arabic)	0.524	0.657	0.576	0.631	**0.597**
El-oc (English)	0.604	0.638	0.544	0.610	**0.599**
El-oc (Arabic)	0.459	0.538	0.483	0.587	**0.517**

Table 11: The results for subtasks V-reg, V-oc, and E-c

Task	Spearman score
V-reg (English)	**0.787**
V-reg (Arabic)	**0.773**
V-oc (English)	**0.736**
V-oc (Arabic)	**0.748**

Task	Accuracy score
E-c (English)	**0.471**
E-c (Arabic)	**0.446**

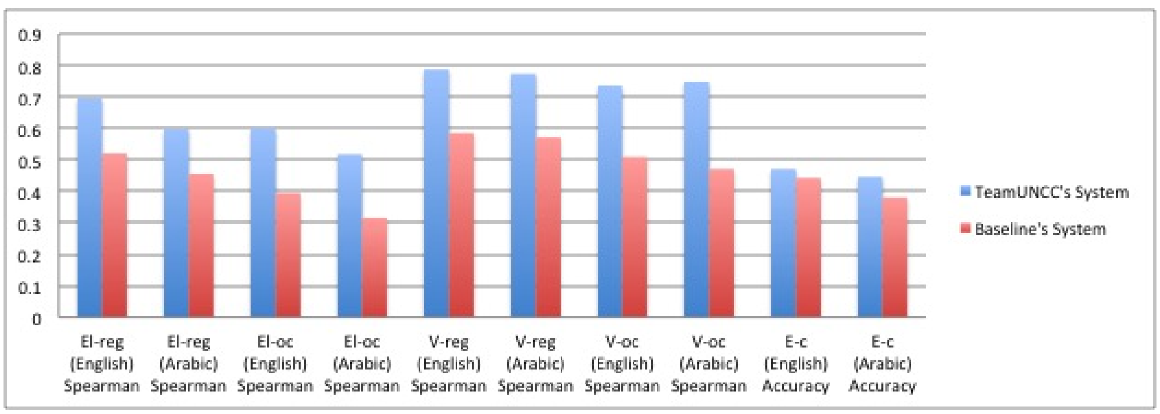


Figure 15: Comparing performances of the proposed system and the baseline systems

4.6 Conclusion

In this chapter, we have presented our system that participated in Task 1 of Semeval-2018. Our system is unique in that we use the same underlying architecture for all subtasks for both languages - English and Arabic to detect the intensity of emotions and sentiments in tweets. The performance of the system for each subtask beats the performance of the baseline's model, indicating that our approach is promising. The system ranked third in El-oc for Arabic language and fourth in the other subtasks for Arabic language too.

In this system, we used word2vec and doc2vec embedding models with feature vectors extracted from the tweets by using the AffectTweets Weka-package, these vectors feed the deep neural network layers to obtain the predictions.

In future work, we will add emotion and valence detection in Spanish language to our system by applying the same approaches that have been used with Arabic. We also want to investigate the Arabic feature attributes in order to enhance the performance in this language.

CHAPTER 5: SEDAT SYSTEM FOR SENTIMENT AND EMOTION DETECTION IN ARABIC TEXT USING CNN LSTM DEEP LEARNING

"Emotions are essential part of human intelligence.

Without emotional intelligence, AI is incomplete."

–Amit Ray

5.1 Introduction

Social microblogging channels, such as Twitter, have become popular communication tools that encourage individuals to express their feelings and opinions on a wide variety of topics. Twitter today plays a vital role in spreading information and influencing people's opinions. It was launched in July 2006 and has since gained worldwide popularity. Statistics from Statista website show that Twitter had 336 million active users in the first quarter of 2018[6]. Due to the shortness of the tweets (Twitter messages), people share their daily activities and thoughts openly most of the time. Therefore, Twitter is considered a rich data bank full of sentiments, emotions, and opinions. Sentiment analysis and emotion detection is the area where researchers extract valuable information regarding people's viewpoints and moods across a variety of products or political decisions [6, 8]. Sentiment analysis refers to classifying a

[6] https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/

subjective text as positive, neutral, or negative; emotion detection recognizes types of feelings through the expression of texts, such as anger, joy, fear, and sadness [8, 31].

A study on Staista website[7] that shows the most common languages on the Internet as of June 2017, by share of Internet users, declares that the Arabic language was ranked fourth with a 4.8 percent share. Although the Arabic language is considered one of the fastest growing languages on Twitter [46], the reality shows little work has been accomplished regarding analysis of Arabic tweets. Most sentiment analysis systems are tailored to English and other Indo-European languages. The mentioned reasons for the lack of research in the Arabic language declared that the complexity and the variety of dialects in Arabic language make it hard to build a single system to detect sentiments and emotions for such a language [12, 5].

Arabic is the official language for 22 countries (approximately 300 million people), and it belongs to the Semitic language family. The Arabic alphabet consists of 28 letters with no upper or lower cases, and the orientation of writing is from right to left. The language is classified into two main categories: Standard Arabic (SA) and Dialectical Arabic (DA). SA consists of two forms: Classical Arabic (CA) and Modern Standard Arabic (MSA) [104]. The current research paper studies the MSA and DA types of the Arabic language, which are used widely on Twitter. Although MSA is the primary language of Arab countries and is written in books and taught in schools, DA is spoken as a native language in people's informal daily communication and has a strong presence in texting and commenting on microblogging networks or in emails.

[7] https://www.statista.com/statistics/262946/share-of-the-most-common-languages-on-the-internet/

On the other hand, these dialects are greatly varied, and can be classified into five main groups according to [118]: Egyptian, Levantine, Iraqi, Gulf, and Maghribi. The problem that researchers face in the natural language processing area is the fact that each dialect is spoken by a specified geographical area for daily verbal communication. Therefore, there is only one MSA language for all Arabic speakers but several dialects with no formal written form[36]. This leads to a lack of lexicon resources for these dialects, and the official grammar rules do not work as efficiently with DA as with MSA.

Deep Neural Networks (DNN) have recently shown significant improvements over traditional Machine Learning (ML) based approaches on classification tasks [41]. In [41, 19, 58], researchers show that Convolution Neural Networks (CNN) and Long-Short Term Memory Networks (LSTM) outperform the traditional machine learning approaches on text classifications, such as sentiment, emotion, and stance detections. Our system, SEDAT (Sentiment and Emotion Detection in Arabic Text), is the first system designed to detect and to predict the intensity of sentiments and emotions in Arabic Tweets using Deep Neural Networks (DNN). The data that is used as an input to our system is obtained from the public Twitter datasets of SemEval Task-1, Affect in Tweets [77]. The extracted features are mainly word embedding vectors [71] and semantic features acquired from the AffectiveTweets package [74, 23]. Our system applies these feature vectors to CNN-LSTM [48, 66] and a fully connected neural network architecture to classify sentiment, emotion as well as intensity of emotion in Arabic language tweets. The performance of our system shows substantial improvements in Spearman correlation scores over the baseline models, with 0.01-0.02

points difference between the state-of-the-art model and our proposed model.

The remainder of this Chapter is organized as follows. Section 5.2 gives a brief description of existing works in detecting sentiments and emotions in social media for the Arabic language. Section 5.3 provides detailed SEDAT system architecture to determine the presence and the intensity of sentiments and emotions in tweets. Section 5.4 presents the evaluations and the results. Section 5.5 describes how the system can handle different dialects and compares the results with different systems. Finally, section 5.6 concludes with future directions for this research.

5.2 Related work

Sentiment and Emotion Detection: There is great body of work from psychology that theorizes about emotions [30, 111]. Ekman [30] identified the six basic emotions as anger, disgust, fear, happiness, sadness, and surprise. Plutchik [88] added two more emotions to Ekman's list: trust and anticipation. while others have listed [16] eleven basic emotions: anger, aversion, courage, dejection, desire, despair, fear, hate, hope, love, and sadness. In this work, we investigate how anger, joy, fear, and sadness are expressed in Arabic text.

Very few corpora exist for emotion labeling [14, 79]. Prior research collected and annotated tweets to analyze how emotions can be distributed in the annotated tweets [93]. They also trained a classifier that automatically discovers the emotions in tweets. EmoTex [47] applied supervised learning methods to detect emotions. A group of researchers [74, 75] introduced shared tasks of detecting the intensity of emotion felt by the speaker of a tweet. State-of-the-art systems in these competitions [40, 28] used approaches of ensembling different models and applied feature vectors including word embeddings, semantic, and syntactic features to represent tweets.

Table 12: Examples of annotated Arabic tweets with translations

Arabic Tweets with English Translations	**Task**	**Annotation**
صباح الخير لإبتسامة أمي لصديقتي المفضله لكل تلك الأشياء الصغيرة التي تزرع بدواخلنا سعادة عظيمة صباح الخير للعالم أجمع Good morning to my mothers's smile, to my best friend, to all those little things that grow inside us with great happiness. Good morning to the whole world.	Sentiment	0.828 3: very positive
وجود الاشخاص بغرفتي وانا مو فيها يزعجني Having people in my room when I am not there annoys me.	Anger Emotion	0.484 1: low anger
ايه ده طب هنحضر خطوبتك امتى طيب انا عاوز ارقص What is this? So, when are we going to attend your engagement? I want to dance!	Joy Emotion	0.672 2: moderate joy
فراق الأم ألم لا يعرفه إلا من تجرع مرارة فقدها رحم الله والدتك ووالدتي وأسكنهما فسيح جناته وجميع موتى المسلمين The parting of the mother is a pain that only the person who lost his mother knows. May Allah have mercy on my mother and your mothers and keep them and all the dead Muslims in paradise.	Sadness Emotion	0.913 3: high sadness

Arabic Sentiment and Emotion Detection: To the best of our knowledge, sentiment and emotion detection for Arabic text is relatively new [5, 34, 17]. The main challenges that most researchers face in analyzing sentiment and emotion in Arabic text can be classified under two main areas [18]: a lack of annotated resources and more complex morphology relative to other languages. Researchers in [90] had collected and annotated data and applied different preprocessing steps related to the Arabic language. They also used a simplification of the SVM (known as SMO) and the NaiveBayes classifiers. Another two related works [59, 94] shared different tasks to identify the overall sentiments of the tweets or phrases taken from tweets in both English and Arabic. This current work investigates different neural network architectures with selecting best features to construct SEDAT system. Moreover, we are able to detect emotions in different dialects of Arabic language. To the best of our knowledge, SEDAT system is the first system to detect the intensity of emotions for both MSA and DA with using deep learning approaches.

5.3 SEDAT System

Our system, SEDAT, has the ability to determine the existence and the intensity of an emotion (Anger, Joy, Fear, or Sadness) in an Arabic tweet as a real-valued score between 0 (least intensity) and 1 (most intensity). It also classifies the intensity of emotion into one of four ordinal classes (0: no emotion, 1: low emotion, 2: moderate emotion, and 3: high emotion). Furthermore, the system is able to to determine the sentiment or valence in a tweet as a real-valued score between 0 (most negative) and 1 (most positive). It can also classify the sentiment intensity of a tweet into one of seven ordinal classes, corresponding to various levels of positive and negative sentiment intensity, starting with 3: very positive and ending with -3: very negative. Table12 shows some examples of Arabic tweets with the English translations and the intensity of sentiment or emotion in the original Arabic tweets.

SEDAT system consists mainly of two sub-models. Figure 16 shows the structure of our system. More details about the system's components are provided in the following subsections: Section 3.1 describes the system's input and preprocessing step. Section 3.2 lists the feature vectors that are used in both sub-models, and Section 3.3 presents the different architectures of neural networks and deep learning that are used in both sub-models. Section 3.4 discusses the output results.

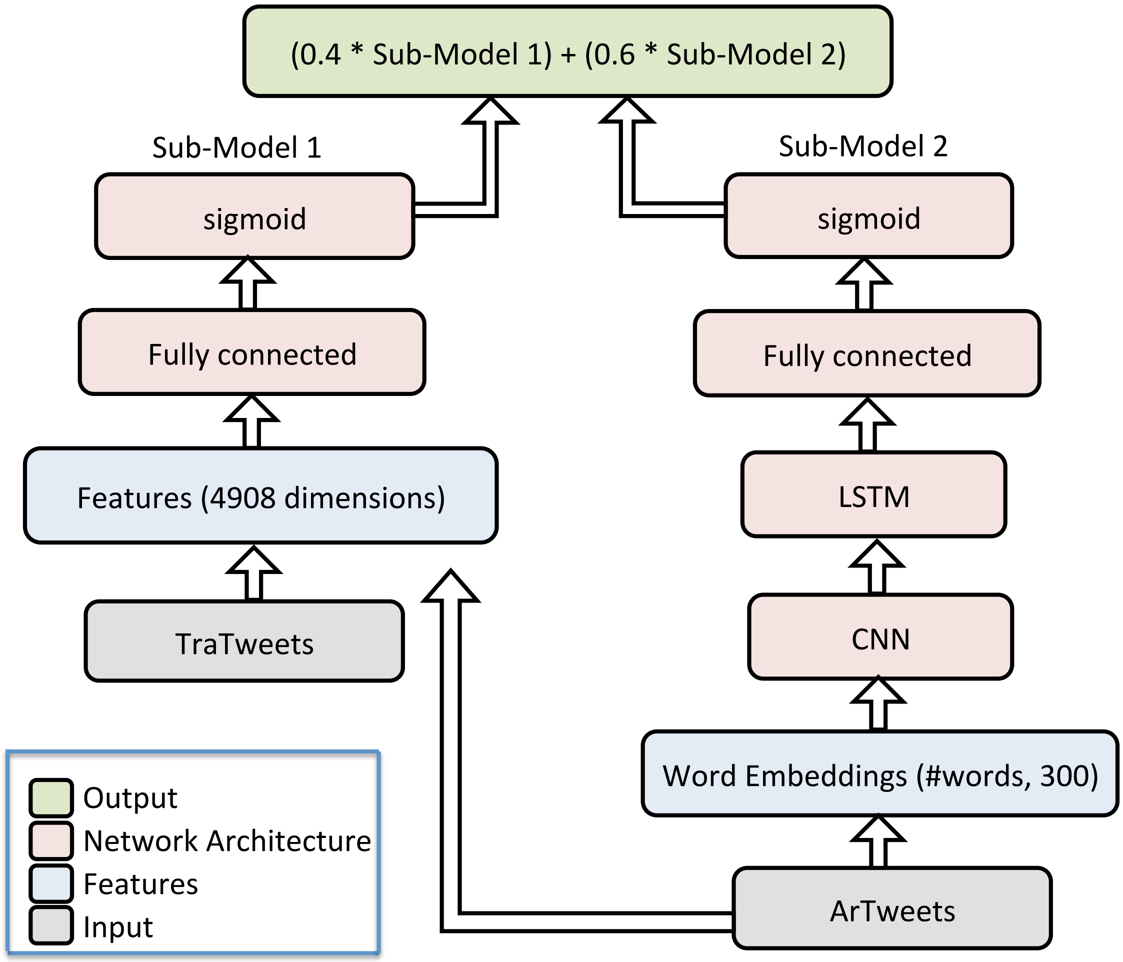


Figure 16: The structure of SEDAT system

5.3.1 Input and Preprocessing

The data inputs for our system have been attained from the public datasets of SemEval-2018 (Task 1: Affect in Tweets) [77]. The number of Arabic training and testing datasets for both sentiment and emotion tasks are illustrated in Table 13. The Arabic tweets in our system have been used in two forms: as original raw Arabic tweets (ArTweets) or as translated into English (TraTweets) since English language has more

Table 13: Number of Arabic tweets as input datasets

–	**Train Data**	**Test Data**
Anger	1027	373
Joy	952	448
Sadness	1030	370
Fear	1028	372
Sentiment	1070	730

preprocessing and feature extraction methods than Arabic. Several preprocessing steps have been applied to both ArTweets and TraTweets.

ArTweets: The original Arabic tweets in training and testing datasets have been tokenized, white spaces have been removed, and the punctuation marks have been treated as individual words (".,?!:;()[]#@'). It is worth mentioning that the preprocessing methods that were not included in our system are normalizing Arabic characters, removing diacritics, removing punctuations, and removing repeating characters. We have tried them but removed them after noticing that there is no improvement in the classification or regression results.

TraTweets: The Arabic tweets have been translated into English using a powerful translation tool written in Python (translate 3.5.0)[8]. The translated tweets then have been tokenized by converting the sentences into words, and all uppercase letters have been converted to lowercase. The preprocessing step also includes stemming the words and removing extraneous white spaces. Punctuation marks have been treated as individual words (".,?!:;()[]#@'), while contractions (wasn't, aren't) were left untreated. We need to highlight that applying preprocessing step for Arabic

[8]https://pypi.python.org/pypi/translate

language before the translation was not appropriate for our system.

5.3.2 Feature Vectors

We explore different features to represent both ArTweets and TraTweets and select the best configurations that result in high performance. SEDAT system consists of two sub-models. The first sub-model uses ArTweets with a set of Arabic lexicons to produce ArabicFeature vector with 5 dimensions and TraTweets to produce other vectors with a total of 4903 dimensions (More details about the extracted features for the first sub-model are listed below). The second sub-model uses only ArTweets. Each word in a tweet in the second sub-model is represented as a 300 dimensional vector using the pretrained word embedding model AraVec (Twt-SG)[109] that is trained on Arabic tweets. Then, each tweet is represented as a vector with a fixed number of rows that equals the maximum length of dataset tweets and a standard 300 columns using padding of zero vectors. The following are the set of features for First sub-model:

AffectiveTweets-142: Each tweet in TraTweets is represented as a vector with 142 dimensions by concatenating three vectors obtained from the AffectiveTweets Weka-package [74, 23], 100 dimensional vector is obtained by vectorizing the tweets to embeddings attribute; two-dimensional vector using the Sentiment Strength feature; and finally 40 features have been extracted using the TweetToLexiconFeatureVector attribute that calculates attributes for a tweet using a variety of lexical resources. TweetToLexiconFeatureVector produces 43 dimensional vector, but after applying feature selection (LinearRegression and RandomForestRegressor) to these features we delete the least three significant features, this step shows 0.005 improvement in the results.

Doc2Vec-600: Each tweet in TraTweets is represented as a 600 dimensional vector using the document-level embeddings (doc2vec) [65, 64]. The 600 dimensions are acquired by concatenating two vectors of 300 dimensions each (dm and dbow). Averaging method has been applied to the vectors for each word in the tweet to attain 300 dimensions that best represent the tweet.

ArabicFeatures-5: This vector has been built using different features from [97, 60] and [63]. We start with 10 features then apply feature selection (LinearRegression and RandomForestRegressor) that helps to rank features and choose the configuration of 5 features, this step improves the performance of SEDAT model. The 5 best selected features are: Arabic Emoticon Lexicon, Arabic Hashtag Lexicon, Arabic Hashtag Lexicon (dialectal), Arabic translation of Bing Lius Lexicon, and one feature that represents the emoji's in the tweet from [63].

DeepEmoji-64: Each TraTweet is represented as a 64 dimensional vector using deepMoji model [38], which is a model trained on 1.2 billion tweets with emojis to understand how language is used to express emotions. DeepMoji model predicts the sentiment of a tweet and produces different representation for a tweet. We extract the embeddings from the softmax layer with 64 dimensional vector.

UnsupervisedLearning-4096: Each word in TraTweet is represented as a 4096 dimensional vector that is extracted by using Unsupervised Sentiment Neuron [91], which learns an excellent representation of sentiment even though the model is trained only to predict the next character in the text.

EmojiFeature-1: Knowing that DeepMoji [38] model works only with 64 emoticons, we annotate them and assign values to each one of the 64 emoticon. We use DeepMoji model to produce different emoticons related to each tweet. Then by using our annotation, we build a one-dimensional vector to represent the tweet.

5.3.3 Network Architecture

Neural networks (NN) have recently become attractive to researchers for language modeling. A standard NN consists of simple connected processors called neurons, each producing a sequence of real-valued activations [101]. Deep learning model is composed of multiple processing layers to learn representations of data with multiple levels of abstraction. Although feed-forward networks can predict the next word of a sequence, the standard Recurrent Neural Networks (RNN) can take into account all of the predecessor words. RNNs are distinguished from feedforward networks by saying that RNNs have memory. A special kind of RNNs are Long Short Term Memory networks (LSTM) [48], while Convolutional Neural Networks (CNN) [66] are feed-forward neural networks. In recent years, both networks have become the state-of-the-art models for a variety of machine learning problems. A common architecture for LSTM is composed of a memory cell, an input gate, an output gate and a forget gate. The cell stores a value (or state), for either long or short time periods. This is achieved by using activation function for the memory cell. CNN makes an efficient use of layers with convolving filters that are applied to local features [66]. The CNN LSTM architecture involves using CNN layers for feature extraction on input data combined with LSTM to support sequence prediction. This architecture was originally referred to as a Long-term Recurrent Convolutional Network or LRCN model [27]. The current research uses feed-forward, LSTM, and CNN to predict the sentiment and emotion in a tweet. The following is a full description of the network architectures for both sub-models in SEDAT system:

Sub-Model 1: The input 4908 dimensional vector feeds into a fully connected neural network with three dense hidden layers of 500, 200, and 80 neurons for each layer, respectively. The activation function for each layer is ReLU [70]. The output layer consists of one sigmoid neuron, which predicts the intensity of the emotion or the sentiment between 0 and 1. Two dropouts are used in this network (0.3, 0.2) after the first and second layers, respectively. For optimization, we use Stochastic Gradient Descent (SGD) optimizer (lr=0.01, decay=1×10^{-6}, and momentum=0.9)[9], augmenting for MSE loss function and ACCURACY metrics. Early stopping is also applied to obtain best results. Best weights for the output predictions are saved to predict the testing datasets. The fit function uses number of epoch=40, batch size=8, validation split=33%.

Sub-Model 2: This sub-model uses a CNN LSTM architectures by adding CNN layer on the front end followed by LSTM layer with Dense layers on the output. The input vector (300, maxLengthOfTweet) feeds a CNN layer with 64 filters, the kernel size is three, and the activation function is ReLU. A maxpooling with pool size=2 is added, then the vectors will be directed to an LSTM of 256 neurons. To avoid over fitting, we use dropout 0.3 after LSTM layer. We add two dense hidden layers with 200 and 80 neurons and ReLU activation function. The output layer consists

[9] https://keras.io/optimizers/

Table 14: The Spearman correlation scores for SEDAT system and each sub-model in the system

Regression Task	**sub-model 1**	**sub-model 2**	**Final Prediction**
Anger Emotion	0.51	0.55	**0.595**
Joy Emotion	0.62	0.64	**0.747**
Fear Emotion	0.54	0.57	**0.622**
Sadness Emotion	0.459	0.649	**0.680**
Sentiment	0.78	0.74	**0.818**

of one sigmoid neuron, which predicts the intensity of the emotion or the sentiment between 0 and 1. For optimization, we use the same method as we use in Sub-Model 1. Also, best weights for the output predictions are saved to predict the testing datasets. The fit function uses the same epoch, patch size, and validation split parameters of Sub-Model 1.

5.3.4 Output and Results

Each sub-model in SEDAT system produces a real-valued number between 0 and 1. It has been shown that the prediction of second sub-model gives higher Pearson correlations than first sub-model for emotion task, and vice versa for sentiment Task. Also, using averaging method for both predictions provides better results than each one separately. After trying different weights for both predictions, we find that taking 40% of Sub-model 1 and 60% from Sub-model 2 gives better results for all emotions, and vice versa for Sentiment. Table 14 produces comprehensive details on how each prediction of each sub-model is produced. Also, it shows the final prediction results with 40% of sub-model1 and 60% of sub-model2.

We classify the final results of the real-valued number to one of the ordinal classes.

Table 15: Classify the output to ordinal classes for Arabic El-oc

Output class	**Angry**	**Joy**	**Fear**	**Sadness**
0: no emotion	0-0.40	0-0.31	0-0.45	0-0.47
1: low amount of emotion	0.40-0.55	0.31-0.51	0.45-0.56	0.47-0.54
2: moderate amount of emotion	0.55-0.64	0.51-0.75	0.56-0.76	0.54-0.67
3: high amount of emotion	0.64-1	0.75-1	0.76-1	0.67-1

Table 16: Classify the output to ordinal classes for English and Arabic V-oc

Output class	**Sentiment**
-3: very negative emotional state	0-0.20
-2: moderately negative emotional state	0.20-0.37
-1: slightly negative emotional state	0.37-0.43
0: neutral or mixed emotional state	0.43-0.56
1: slightly positive emotional state	0.56-0.69
2: moderately positive emotional state	0.69-0.81
3: very positive emotional state	0.81-1

Table 17: The Spearman correlation scores

Task	**Anger**	**Joy**	**Fear**	**Sadness**	**Final result**
Emotion Regression	0.595	0.747	0.622	0.68	**0.661**
Emotion Classification	0.504	0.537	0.526	0.611	**0.569**
Sentiment Regression	-	-	-	-	**0.817**
Sentiment Classification	-	-	-	-	**0.786**

We determine the ranges of values for each ordinal class by studying the annotated datasets. Tables 15 and 16 show the ranges of values to obtain the ordinal classes for emotions and sentiments, respectively.

The final results for the regression and classification tasks for both emotion and sentiments are shown in Table 17.

5.4 Analysis and Evaluations

Table 17 shows the performance of our system in regression and classification tasks for both sentiment and emotion. The performance of our system surpasses the SVM Unigrams Baseline model's performance, which is provided by the SemEval-Task 1's organizers. Figure 17 shows the difference between the two performances. SEDAT system also shows substantial improvements over TeamUNCC's system [6], which is the previous version of SEDAT system. SEDAT is only 0.01 to 0.02 points behind the first-ranked model in the challenge. It is worth mentioning that our results have been obtained using the task datasets without using any external data.

To gain insight on how SEDAT system performs in each language category, we manually split the testing dataset into MSA and two main dialects, Egyptian and Gulf dialects. We notice that the system performs best with MSA for Anger and Fear emotions, whereas Egyptian dialect preforms the best with Joy emotion and Gulf dialect with Sadness emotion (see Figure 18).

5.5 Conclusion

In this chapter, we have presented our system SEDAT that uses deep learning architectures for detecting the intensity of emotions and sentiments in Arabic tweets. The performance of the system surpasses the performance of the baseline's model, indicating that our approach is promising. In this system, we use word and document embedding models with feature vectors extracted from Arabic and translated tweets by using the AffectiveTweets package, Deepmoji, and Unsupervised Sentiment Neurons. These vectors feed different deep neural network architectures, feed-forward,

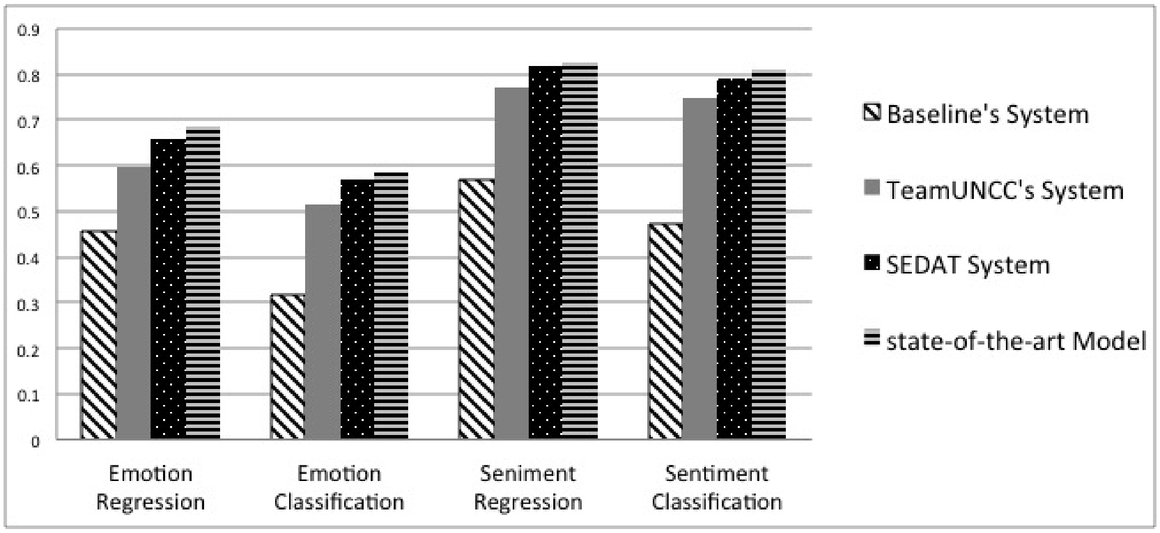


Figure 17: Comparing the Spearman correlation scores of SEDAT system with TeamUNCC and the baseline systems

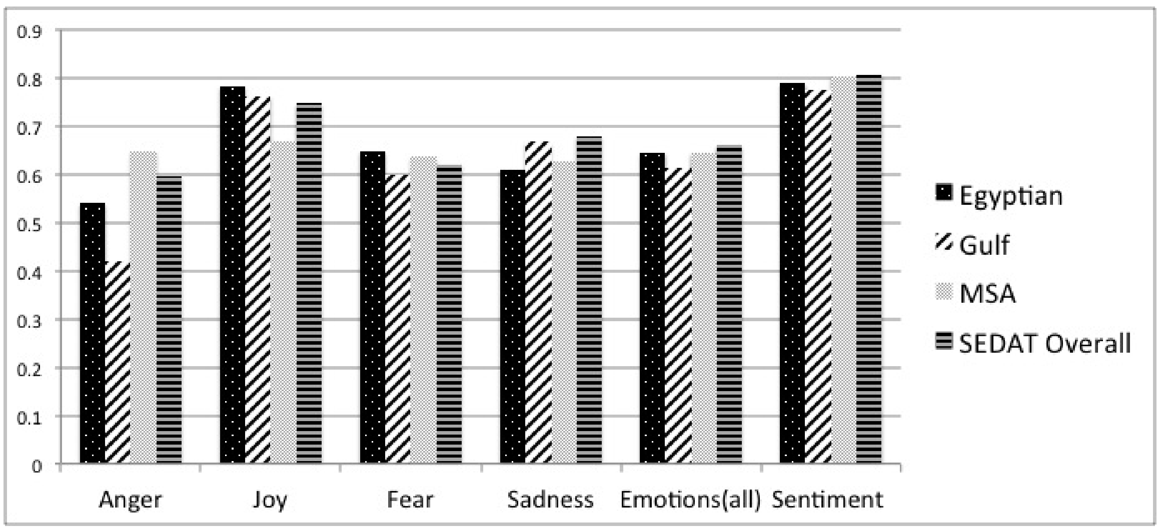


Figure 18: Analyzing the Spearman correlation scores of SEDAT system for each dialect

CNN, and LSTM to obtain the predictions. We use the SemEval-2018 Task 1's datasets as input for our system and shows that the performance of SEDAT is close to the performance of the first-ranked model in the task's challenge with a difference of 0.01-0.02 points .

CHAPTER 6: CONCLUSION

"A conclusion is simply the place
where you got tired of thinking."
–Dan Chaon

This dissertation has addressed the challenges of sentiment analysis and emotion detection on Twitter data. Numerous studies that analyzed people's opinions in English and other Indo-European languages have been reviewed. However, there are few studies that analyzed people's opinions in the Arabic language. A sophisticated categorization of a large number of recent articles has been reviewed in this dissertation to cover a wide variety of sentiment analysis in the Arabic language. The outcome of this study demonstrates a need for building and publishing additional lexicon Arabic resources with different genres and various dialects for both the public and research community. Assembling all lexicons for Arabic dialects from different geographical areas in the Middle East in one lexicon repository is a worthy goal.

This dissertation also has shown the process of performing sentiment analysis for social media platform tools by proceeding through multiple phases starting from crawling data, cleaning data, extracting the features, classifying the data, and visualizing results. The low efficiency of the used method shows a need to produce new systems for sentiment and emotion classification and regression tasks.

In consequence of the above-mentioned studies, we have proposed a system to detect the intensity of emotions and sentiments in English and Arabic tweets. The system outperformed the baseline models in all five subtasks of SemEval-2018 Task 1, indicating that our approach is promising. The system ranked third in El-oc for Arabic language and fourth in the other subtasks for the the Arabic language, too. In this system, we use word and document embedding models with feature vectors extracted from the tweets by using the AffectiveTweets package. These vectors feed the deep neural network layers to obtain the predictions. Our work uses the state-of-the-art approaches of deep learning and word/doc embeddings which have recently shown significant improvements over traditional machine learningbased approaches. To the best of our knowledge, there is no emotion detection system for Arabic tweets that use these approaches.

Although further investigations are needed to construct and to select new features for English tweets, we have chosen to investigate the Arabic language structure exclusively because we believe that proving a positive correlation for the syntactic features of language and emotion would have serious implications in this field. Translating from Arabic to English has been shown as an effective way to detect emotions; however, the system features depend on the morphology of a language, and Arabic has a more complicated morphology and a different structure than English. Therefor, we have extended our system to include features from the Arabic language itself. The new added features have improved the performance of the proposed system that is designed to determine the intensity of sentiment and emotion in the Arabic text.

Deep Neural Networks have recently shown significant improvements over traditional machine learningbased approaches on classification tasks. As a result, the new proposed system SEDAT has used CNN-LSTM architecture to train the model and determine the intensity of sentiment and emotion in Arabic tweets. The outcome of this new system shows a significant improvement over the previous system and baseline models. The system also proved a high proficiency in detecting sentiments and emotions in different dialectics and modern standard Arabic language.

There are no emotion lexicons that cover a wide variety of dialectical Arabic. Thus, further research should be conducted that involves creating and annotating emotional lexicons for different dialects of the Arabic language. Also, there is a need to annotate emotional tweets that can help train the model and improve predictions.

REFERENCES

[1] M. Abdul-Mageed and M. T. Diab. Awatif: A multi-genre corpus for modern standard arabic subjectivity and sentiment analysis. In *LREC*, pages 3907–3914. Citeseer, 2012.

[2] M. Abdul-Mageed and M. T. Diab. Sana: A large scale multi-genre, multi-dialect lexicon for arabic subjectivity and sentiment analysis. In *LREC*, pages 1162–1169, 2014.

[3] M. Abdul-Mageed, S. Kübler, and M. Diab. Samar: A system for subjectivity and sentiment analysis of arabic social media. In *Proceedings of the 3rd workshop in computational approaches to subjectivity and sentiment analysis*, pages 19–28. Association for Computational Linguistics, 2012.

[4] N. A. Abdulla, N. A. Ahmed, M. A. Shehab, and M. Al-Ayyoub. Arabic sentiment analysis: Lexicon-based and corpus-based. In *Applied Electrical Engineering and Computing Technologies (AEECT), 2013 IEEE Jordan Conference on*, pages 1–6. IEEE, 2013.

[5] M. Abdullah and M. Hadzikadic. Sentiment analysis on arabic tweets: Challenges to dissecting the language. In *International Conference on Social Computing and Social Media*, pages 191–202. Springer, 2017.

[6] M. Abdullah and S. Shaikh. Teamuncc at semeval-2018 task 1: Emotion detection in english and arabic tweets using deep learning. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 350–357, 2018.

[7] A. Agarwal, F. Biadsy, and K. R. Mckeown. Contextual phrase-level polarity analysis using lexical affect scoring and syntactic n-grams. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 24–32. Association for Computational Linguistics, 2009.

[8] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau. Sentiment analysis of twitter data. In *Proceedings of the workshop on languages in social media*, pages 30–38. Association for Computational Linguistics, 2011.

[9] S. Ahmed, M. Pasquier, and G. Qadah. Key issues in conducting sentiment analysis on arabic social media text. In *Innovations in Information Technology (IIT), 2013 9th International Conference on*, pages 72–77. IEEE, 2013.

[10] M. Al-Ayyoub, S. B. Essa, and I. Alsmadi. Lexicon-based sentiment analysis of arabic tweets. *International Journal of Social Network Mining*, 2(2):101–114, 2015.

[11] B. Alexandra and T. Marco. Multilingual sentiment analysis using machine translation. In *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*, pages 52–60, 2012.

[12] S. O. Alhumoud, M. I. Altuwaijri, T. M. Albuhairi, and W. M. Alohaideb. Survey on arabic sentiment analysis in twitter. *International Science Index*, 9(1):364–368, 2015.

[13] A. Alqudsi, N. Omar, and K. Shaker. Arabic machine translation: a survey. *Artificial Intelligence Review*, 42(4):549–572, 2014.

[14] S. Aman and S. Szpakowicz. Identifying expressions of emotion in text. In *International Conference on Text, Speech and Dialogue*, pages 196–205. Springer, 2007.

[15] D. Anuta, J. Churchin, and J. Luo. Election bias: Comparing polls and twitter in the 2016 us election. *arXiv preprint arXiv:1701.06232*, 2017.

[16] M. B. Arnold. Emotion and personality. 1960.

[17] A. Assiri, A. Emam, and H. Al-Dossari. Saudi twitter corpus for sentiment analysis. *World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering*, 10(2):272–275, 2016.

[18] A. Assiri, A. Emam, and H. Aldossari. Arabic sentiment analysis: a survey. *International Journal of Advanced Computer Science and Applications*, 6(12):75–85, 2015.

[19] I. Augenstein, T. Rocktäschel, A. Vlachos, and K. Bontcheva. Stance detection with bidirectional conditional encoding. *arXiv preprint arXiv:1606.05464*, 2016.

[20] A. Balahur, J. M. Hermida, and A. Montoyo. Detecting implicit expressions of emotion in text: A comparative analysis. *Decision Support Systems*, 53(4):742–753, 2012.

[21] M. Bautin, L. Vijayarenu, and S. Skiena. International sentiment analysis for news and blogs. In *ICWSM*, 2008.

[22] E. Boiy, P. Hens, K. Deschacht, and M.-F. Moens. Automatic sentiment analysis in on-line text. In *ELPUB*, pages 349–360, 2007.

[23] F. Bravo-Marquez, M. Mendoza, and B. Poblete. Meta-level sentiment models for big social data analysis. *Knowledge-Based Systems*, 69:86–99, 2014.

[24] D. Chin, A. Zappone, and J. Zhao. Analyzing twitter sentiment of the 2016 presidential candidates. *News & Publications: Stanford University*, 2016.

[25] W. O. Dailey, E. A. Hinck, and S. S. Hinck. *Politeness in presidential debates: Shaping political face in campaign debates from 1960 to 2004*. Rowman & Littlefield, 2008.

[26] H. Demuth and M. Beale. Neural network toolbox for use with matlab–user's guide verion 3.0. 1993.

[27] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015.

[28] V. Duppada, R. Jain, and S. Hiray. Seernet at semeval-2018 task 1: Domain adaptation for affect in tweets. *arXiv preprint arXiv:1804.06137*, 2018.

[29] R. Duwairi, R. Marji, N. Sha'ban, and S. Rushaidat. Sentiment analysis in arabic tweets. In *Information and communication systems (icics), 2014 5th international conference on*, pages 1–6. IEEE, 2014.

[30] P. Ekman. Are there basic emotions? 1992.

[31] P. Ekman. Facial expression and emotion. *American psychologist*, 48(4):384, 1993.

[32] P. Ekman, W. Friesen, and P. Ellsworth. Emotion in the face, 1982.

[33] S. R. El-Beltagy and A. Ali. Open issues in the sentiment analysis of arabic social media: A case study. In *Innovations in information technology (iit), 2013 9th international conference on*, pages 215–220. IEEE, 2013.

[34] S. R. El-Beltagy, T. Khalil, A. Halaby, and M. Hammad. Combining lexical features and a supervised learning approach for arabic sentiment analysis. *arXiv preprint arXiv:1710.08451*, 2017.

[35] H. ElSahar and S. R. El-Beltagy. Building large arabic multi-domain resources for sentiment analysis. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 23–34. Springer, 2015.

[36] Z. Eviatar and R. Ibrahim. Why is it hard to read arabic? In *Handbook of Arabic literacy*, pages 77–96. Springer, 2014.

[37] N. Farra, E. Challita, R. A. Assi, and H. Hajj. Sentence-level and document-level sentiment mining for arabic texts. In *2010 IEEE International Conference on Data Mining Workshops*, pages 1114–1119. IEEE, 2010.

[38] B. Felbo, A. Mislove, A. Søgaard, I. Rahwan, and S. Lehmann. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. *arXiv preprint arXiv:1708.00524*, 2017.

[39] D. Gayo Avello, P. T. Metaxas, and E. Mustafaraj. Limits of electoral predictions using twitter. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*. Association for the Advancement of Artificial Intelligence, 2011.

[40] P. Goel, D. Kulshreshtha, P. Jain, and K. K. Shukla. Prayas at emoint 2017: An ensemble of deep neural architectures for emotion intensity prediction in tweets. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 58–65, 2017.

[41] C. Guggilla, T. Miller, and I. Gurevych. Cnn-and lstm-based claim classification in online user comments. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2740–2751, 2016.

[42] I. Guyon and A. Elisseeff. An introduction to feature extraction. In *Feature extraction*, pages 1–25. Springer, 2006.

[43] N. Habash and R. M. Roth. Catib: The columbia arabic treebank. In *Proceedings of the ACL-IJCNLP 2009 conference short papers*, pages 221–224. Association for Computational Linguistics, 2009.

[44] N. Y. Habash. Introduction to arabic natural language processing. *Synthesis Lectures on Human Language Technologies*, 3(1):1–187, 2010.

[45] J. Hajic, O. Smrz, P. Zemánek, J. Šnaidauf, and E. Beška. Prague arabic dependency treebank: Development in data and tools. In *Proc. of the NEMLAR Intern. Conf. on Arabic Language Resources and Tools*, pages 110–117, 2004.

[46] F. Harrag. Estimating the sentiment of arabic social media contents: A survey. In *5th International Conference on Arabic Language Processing*, 2014.

[47] M. Hasan, E. Rundensteiner, and E. Agu. Emotex: Detecting emotions in twitter messages. 2014.

[48] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[49] M. Hu and B. Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM, 2004.

[50] H. S. Ibrahim, S. M. Abdou, and M. Gheith. Sentiment analysis for modern standard arabic and colloquial. *arXiv preprint arXiv:1505.03105*, 2015.

[51] S. M. Jimenez-Zafra, M. T. M. Valdivia, E. M. Camara, and L. A. Urena-Lopez. Studying the scope of negation for spanish sentiment analysis on twitter. *IEEE Transactions on Affective Computing*, 2017.

[52] T. Joachims. Text categorization with support vector machines: Learning with many relevant features. In *European conference on machine learning*, pages 137–142. Springer, 1998.

[53] J. Kamps, M. Marx, R. J. Mokken, and M. De Rijke. Using wordnet to measure semantic orientations of adjectives. In *LREC*, volume 4, pages 1115–1118. Citeseer, 2004.

[54] R. S. Katz and R. Kolodny. Party organization as an empty vessel: Parties in american politics. *How Parties Organize: Change and Adaptation in Party Organizations in Western Democracies*, 23:24–29, 1994.

[55] D. Kellner. *American Nightmare: Donald Trump, Media Spectacle, and Authoritarian Populism*, volume 117. Springer, 2016.

[56] J. Khairnar and M. Kinikar. Machine learning algorithms for opinion mining and sentiment classification. *International Journal of Scientific and Research Publications*, 3(6):1–6, 2013.

[57] A. A. Khrisat and Z. A. Alharthy. Arabic dialects and classical arabic language. *Advances in Social Sciences Research Journal*, 2(3), 2015.

[58] Y. Kim. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*, 2014.

[59] S. Kiritchenko, S. Mohammad, and M. Salameh. Semeval-2016 task 7: Determining sentiment intensity of english and arabic phrases. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 42–51, 2016.

[60] S. Kiritchenko, S. M. Mohammad, and M. Salameh. Semeval-2016 task 7: Determining sentiment intensity of english and arabic phrases. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '16, San Diego, California, June 2016.

[61] M. Korayem, D. Crandall, and M. Abdul-Mageed. Subjectivity and sentiment analysis of arabic: A survey. In *Advanced Machine Learning Technologies and Applications*, pages 128–139. Springer, 2012.

[62] E. Kouloumpis, T. Wilson, and J. D. Moore. Twitter sentiment analysis: The good the bad and the omg! *Icwsm*, 11(538-541):164, 2011.

[63] P. Kralj Novak, J. Smailović, B. Sluban, and I. Mozetič. Sentiment of emojis. *PLoS ONE*, 10(12):e0144296, 2015.

[64] J. H. Lau and T. Baldwin. An empirical evaluation of doc2vec with practical insights into document embedding generation. *arXiv preprint arXiv:1607.05368*, 2016.

[65] Q. Le and T. Mikolov. Distributed representations of sentences and documents. In *International Conference on Machine Learning*, pages 1188–1196, 2014.

[66] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[67] H. Liu and H. Motoda. Feature transformation and subset selection. *IEEE Intell Syst Their Appl*, 13(2):26–28, 1998.

[68] C.-Y. Lu, S.-H. Lin, J.-C. Liu, S. Cruz-Lara, and J.-S. Hong. Automatic event-level textual emotion sensing using mutual action histogram between entities. *Expert systems with applications*, 37(2):1643–1653, 2010.

[69] M. Maamouri, A. Bies, T. Buckwalter, and W. Mekki. The penn arabic treebank: Building a large-scale annotated arabic corpus. In *NEMLAR conference on Arabic language resources and tools*, volume 27, pages 466–467, 2004.

[70] A. L. Maas, A. Y. Hannun, and A. Y. Ng. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, volume 30, page 3, 2013.

[71] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[72] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[73] S. M. Mohammad. # emotional tweets. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 246–255. Association for Computational Linguistics, 2012.

[74] S. M. Mohammad and F. Bravo-Marquez. Wassa-2017 shared task on emotion intensity. *arXiv preprint arXiv:1708.03700*, 2017.

[75] S. M. Mohammad, F. Bravo-Marquez, M. Salameh, and S. Kiritchenko. Semeval-2018 Task 1: Affect in tweets. In *Proceedings of International Workshop on Semantic Evaluation (SemEval-2018)*, New Orleans, LA, USA, 2018.

[76] S. M. Mohammad and S. Kiritchenko. Using hashtags to capture fine emotion categories from tweets. *Computational Intelligence*, 31(2):301–326, 2015.

[77] S. M. Mohammad and S. Kiritchenko. Understanding emotions: A dataset of tweets to study interactions between affect categories. In *Proceedings of the 11th Edition of the Language Resources and Evaluation Conference*, Miyazaki, Japan, 2018.

[78] S. M. Mohammad, S. Kiritchenko, and X. Zhu. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. In *Second Joint Conference on Lexical and Computational Semantics (* SEM)*, volume 2, pages 321–327, 2013.

[79] S. M. Mohammad and P. D. Turney. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*, pages 26–34. Association for Computational Linguistics, 2010.

[80] A. Mourad and K. Darwish. Subjectivity and sentiment analysis of modern standard arabic and arabic microblogs. In *Proceedings of the 4th workshop on computational approaches to subjectivity, sentiment and social media analysis*, pages 55–64, 2013.

[81] M. Nabil, M. A. Aly, and A. F. Atiya. Labr: A large scale arabic book reviews dataset. *CoRR, abs/1411.6718*, 2014.

[82] T. Nasukawa and J. Yi. Sentiment analysis: Capturing favorability using natural language processing. In *Proceedings of the 2nd international conference on Knowledge capture*, pages 70–77. ACM, 2003.

[83] A. Ortony and T. J. Turner. What's basic about basic emotions? *Psychological review*, 97(3):315, 1990.

[84] N. Oscar, P. A. Fox, R. Croucher, R. Wernick, J. Keune, and K. Hooker. Machine learning, sentiment analysis, and tweets: an examination of alzheimers disease stigma on twitter. *Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, 72(5):742–751, 2017.

[85] A. Pak and P. Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *LREc*, volume 10, 2010.

[86] B. Pang and L. Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, page 271. Association for Computational Linguistics, 2004.

[87] R. Plutchik. A general psychoevolutionary theory of emotion. *Theories of emotion*, 1, 1980.

[88] R. Plutchik. Emotions and psychotherapy: A psychoevolutionary perspective. In *Emotion, psychopathology, and psychotherapy*, pages 3–41. Elsevier, 1990.

[89] N. W. Polsby, A. B. Wildavsky, and D. A. Hopkins. *Presidential elections: Strategies and structures of American politics*. Rowman & Littlefield, 2008.

[90] O. Rabie and C. Sturm. Feel the heat: Emotion detection in arabic social media content. In *The International Conference on Data Mining, Internet Computing, and Big Data (BigData2014)*, pages 37–49. The Society of Digital Information and Wireless Communication, 2014.

[91] A. Radford, R. Jozefowicz, and I. Sutskever. Learning to generate reviews and discovering sentiment. *arXiv preprint arXiv:1704.01444*, 2017.

[92] E. Refaee and V. Rieser. An arabic twitter corpus for subjectivity and sentiment analysis. In *LREC*, pages 2268–2273, 2014.

[93] K. Roberts, M. A. Roach, J. Johnson, J. Guthrie, and S. M. Harabagiu. Empatweet: Annotating and detecting emotions on twitter. In *LREC*, volume 12, pages 3806–3813. Citeseer, 2012.

[94] S. Rosenthal, N. Farra, and P. Nakov. Semeval-2017 task 4: Sentiment analysis in twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518, 2017.

[95] M. Rushdi-Saleh, M. T. Martín-Valdivia, L. A. Ureña-López, and J. M. Perea-Ortega. Oca: Opinion corpus for arabic. *Journal of the American Society for Information Science and Technology*, 62(10):2045–2054, 2011.

[96] H. Saif. Sentiment analysis of microblogs. *Mining the New World, Technical Report KMI-12-2 March*, 2012.

[97] M. S. Saif M. Mohammad and S. Kiritchenko. Sentiment lexicons for arabic social media. In *Proceedings of 10th edition of the the Language Resources and Evaluation Conference (LREC)*, Portorož, Slovenia, 2016.

[98] J. B. Salamah and A. Elkhlifi. Microblogging opinion mining approach for kuwaiti dialect. In *The International Conference on Computing Technology and Information Management (ICCTIM2014)*, pages 388–396. The Society of Digital Information and Wireless Communication, 2014.

[99] M. Salameh, S. M. Mohammad, and S. Kiritchenko. Sentiment after translation: A case-study on arabic social media posts. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 767–777, 2015.

[100] W. Salloum and N. Habash. Dialectal arabic to english machine translation: Pivoting through modern standard arabic. 2013.

[101] J. Schmidhuber. Deep learning in neural networks: An overview. *Neural networks*, 61:85–117, 2015.

[102] O. Selmer and M. Brevik. Classification and visualisation of twitter sentiment data. 2013.

[103] K. Shaalan. Nizar y. habash, introduction to arabic natural language processing (synthesis lectures on human language technologies). *Machine Translation*, 24(3):285–289, 2010.

[104] K. Shaalan. Rule-based approach in arabic natural language processing. *The International Journal on Information and Communication Technologies (IJICT)*, 3(3):11–19, 2010.

[105] L. Shi, N. Agarwal, A. Agrawal, R. Garg, and J. Spoelstra. Predicting us primary elections with twitter. *URL: http://snap. stanford. edu/social2012/papers/shi. pdf*, 2012.

[106] A. Shoukry and A. Rafea. Sentence-level arabic sentiment analysis. In *Collaboration Technologies and Systems (CTS), 2012 International Conference on*, pages 546–550. IEEE, 2012.

[107] S. K. Singh, D. S. Paul, and D. Kumar. Sentiment analysis approaches on different data set domain: Survey. *International Journal of Database Theory and Application*, 7(5):39–50, 2014.

[108] A. Smola and S. Vishwanathan. Introduction to machine learning. *Cambridge University, UK*, 32:34, 2008.

[109] A. B. Soliman, K. Eissa, and S. R. El-Beltagy. Aravec: A set of arabic word embedding models for use in arabic nlp. *Procedia Computer Science*, 117:256–265, 2017.

[110] S. V. Stehman. Selecting and interpreting measures of thematic classification accuracy. *Remote sensing of Environment*, 62(1):77–89, 1997.

[111] M. Thambirajah. *Psychological basis of psychiatry*. Churchill Livingstone, 2005.

[112] E. Tromp and M. Pechenizkiy. Rule-based emotion detection on social media: putting tweets on plutchik's wheel. *arXiv preprint arXiv:1412.4682*, 2014.

[113] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. *ICWSM*, 10(1):178–185, 2010.

[114] P. D. Turney. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 417–424. Association for Computational Linguistics, 2002.

[115] J. D. Ullman. *Mining of massive datasets*. Cambridge University Press, 2011.

[116] S. van den Bosch. Automatic feature generation and selection in predictive analytics solutions.

[117] T. Wilson, P. Hoffmann, S. Somasundaran, J. Kessler, J. Wiebe, Y. Choi, C. Cardie, E. Riloff, and S. Patwardhan. Opinionfinder: A system for subjectivity analysis. In *Proceedings of hlt/emnlp on interactive demonstrations*, pages 34–35. Association for Computational Linguistics, 2005.

[118] O. F. Zaidan and C. Callison-Burch. Arabic dialect identification. *Computational Linguistics*, 40(1):171–202, 2014.