DETECTING ABUSIVE ARABIC LANGUAGE TWITTER ACCOUNTS USING A MULTIDIMENSIONAL ANALYSIS MODEL

by

Ehab Abozinadah
A Dissertation
Submitted to the
Graduate Faculty
of
George Mason University
in Partial Fulfillment of
The Requirements for the Degree
of
Doctor of Philosophy
Information Technology

Committee:

_____	Dr. James H Jones, Dissertation Director
_____	Dr. Hemant Purohit, Committee Member
_____	Dr. Paulo Costa, Committee Member
_____	Dr. Ioulia Rytikova, Committee Member
_____	Dr. Stephen Nash, Senior Associate Dean
_____	Dr. Kenneth S. Ball, Dean, Volgenau School of Engineering

Date: _____ Fall Semester 2017
George Mason University
Fairfax, VA

Detecting Abusive Arabic Language Twitter Accounts Using a Multidimensional Analysis Model

A Dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at George Mason University

by

Ehab Abozinadah
Master of Science
Western Oregon University, 2008
Bachelor of Science
King Abdul-Aziz University, 2004

Director: James H Jones, Associate Professor
Department of Computer Forensics

Fall Semester 2017
George Mason University
Fairfax, VA

DEDICATION

This is dedicated to my parents, wife, children, siblings, close family and close friends.

ACKNOWLEDGEMENTS

I would like to express my appreciation and gratitude to my advisor and dissertation Chair, Dr. James H Jones for his patience, mentorship, expert guidance, and encouragement during this research. I also want to thank my committee members: Dr. Paulo Costa, Dr. Hemant Purohit and Dr. Ioulia Rytikova for their comments, input and support that made this research possible.

Also, I sincerely thankful to Kingdom of Saudi Arabia that presented by King Abdul-Aziz University, Ministry of Higher Education, and Saudi Arabia Culture Mission in U.S.A for the generous scholarship that covered all the expenses of all my graduate educations.

I also thankful for every reviewer who gave me a helpful feedback and comments that improved my work.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF EQUATIONS

LIST OF ABBREVIATIONS

BOW……... Bag of Words
PR …….……... PageRank
SO …….…..… Semantic Orientation
Min …….…….. Minimum
Max …….. Maximum
Avg ……..…….….. Average
Std ……...…..….. Standard Division

ABSTRACT

DETECTING ABUSIVE ARABIC LANGUAGE TWITTER ACCOUNTS USING A MULTIDIMENSIONAL ANALYSIS MODEL

Ehab Abozinadah, Ph.D.

George Mason University, 2017

Dissertation Director: Dr. James H Jones

Twitter is one of the most popular social media sources for disseminating news and propaganda in the Middle East. The increased use of social media has motivated spammers to post malicious content on social media sites. Some of these Arabic language spammers use adult content to further the distribution of their malicious activities. However, the extensive number of users posting adult content in social media degrades the experience for other users for whom the adult content is not desired or appropriate. These accounts would be suspended or terminated from Twitter whenever reported by Twitter's users as Twitter prohibits adult content in an image, a video, or a text. Moreover, some countries have attempted to detect these accounts, but have failed as these accounts use informal Arabic language and misspelled words that cannot be detected using blacklisted keywords.

In this research, I built a model to detect abusive Arabic language Twitter accounts that use obscenity, profanity, or inappropriate words in tweet content. The model is based on a multi-dimensional analysis approach by using independent lexical analysis, social graph analysis, and statistical analysis. Independent lexical analysis approaches are used to overcome the limitation of Arabic language analysis tools for correcting the misspelled words in the tweet, finding the abusive and non-abusive related words, and finding the concept related to the word. Social graph analysis is used to identify the user connectivity relationships on Twitter. Statistical analysis is used to identify the user's tweeting characteristics.

My analysis was based on real data collected from Twitter. The data was manually labeled to support a supervised machine learning technique (Support Vector Machine (SVM)). The constructed model contains 31 distinct features that are formed from profile information, social graph centrality measures, tweet elements' counts, and tweet lexical analysis measures. The model was evaluated against a previously unseen subset of the collected data and achieved 90% average accuracy.

CHAPTER ONE

In recent years, the widespread adoption and use of the social media has changed the way people communicate, obtain news, promote products, and evaluate services. The ubiquity of Internet access and mobile technology, especially smart phones and tablets, continues to drive global adoption and social media use. Examples of popular social media websites are: Twitter, Facebook, YouTube, etc.

Each social media site has user policies and guidelines about content that users are permitted to publish. Whenever users violate these policies and guidelines, either their content is deleted or the accounts are suspended. However, these restrictions have forced some users to maneuver around the boundaries set by these sites to generate content that otherwise would be violating these policies and guidelines. Such content is informal and may have misspellings, slang, vulgarity, profanity, emoticons, or meaningless words.

This research focuses on Twitter, which is a micro blogger social media provider that enables users to compose messages of 140 characters. These messages are called "tweets" and may contain text, pictures, videos, or hyperlinks. The usernames in Twitter start with a prefix symbol (@). Twitter users create their social networks through followers and following relationships. Tweets are posted on the user page and followers' timelines and can be found by Twitter's search engine. The tweets can be forwarded to the user's followers by clicking "retweet." At the same time, the tweet can be replayed

by mentioning the username prefixed by using @ in the tweet. The tweets' topics can be indexed using hashtags (#) for each topic and can be searched through Twitter's search engine.

Cybercriminals are notorious for using Twitter for spamming, sending scams, phishing, and recruiting innocent victims to criminal organizations. They use fake identity on Twitter, which enables them to commit crimes like sending bulk and unsolicited messages, spreading malicious links, generating fraudulent product and service reviews, sharing undesired content, and click-baiting and like-jacking (Tolentino, 2015).

Undesired content prohibited by Twitter includes but is not limited to adult content in the form of an image, a video, and a text. Furthermore, there are other restriction polices for the content and services, to meet with the regulation in certain countries, like promotion of dating services ("Adult or sexual products and services," 2017). However, cybercriminals continue to generate and disseminate this content using informal language to exploit their victims ("Digital extortion on the rise," 2015, "Saudi men prime target of social media blackmail," 2015).

These accounts use abusive content, which includes obscenity, profanity, insulting words, or inappropriate words. These accounts are called abusive accounts as they violate social media policies and abuse laws and regulations of some countries. Further still, there are neither tools nor sufficient research about these accounts despite their increasing activity on different social media platforms.

Arabic is the main language spoken in the 22 countries and the 6th top tweeting language (Fox, 2013). Arabic is a complex morphological, syntactical, and semantic language which varies in different regions of the Middle East (Muaidi & Al-tarawneh, 2012). Arabic language does not have capitalization but the diacritics are used above or below a letter indicate a different on pronunciation. Arabic has two forms: formal Arabic, also called Modern Standard Arabic (MSA), and informal Arabic. Formal Arabic is used in books, newspapers, academia, and other forms of formal literature, while informal Arabic comprises local words and slang words within different regions of the Arabic-speaking world.

Moreover, Arabic language is challenging due the limitation of sentiment analysis tools that capable to interact with Arabic dialects and slang (Mubarak, Darwish, & Magdy, 2017). This increases the complexity of understanding the concepts in Arabic tweets. Most researches have been conducted on lexical analysis and machine learning on English corpora (Benevenuto, Magno, Rodrigues, & Almeida, 2010; McCord & Chuah, 2011; A. H. Wang, 2010). There is limited research on Arabic text mainly due to its morphological complexity and limited lexical resources (Alsaleem, 2011; Rsheed & Khan, 2014; Heider A. Wahsheh, Al-kabi, & Alsmadi, 2012). In addition, informal language including slang and dialects are different from region to region, and from generation to generation, and is not covered by word dictionaries (Diab & Habash, 2007). The state of the art for the current censorship systems in Arab nations uses blacklists to identify the abusive accounts, an approach which has significant limitations (Chaabane et al., 2014).

In this research, Arabic tweets were analyzed to detect abusive Arabic language Twitter accounts. This is because some Arab nations have attempted but failed to censor Internet traffic to block content from abusive accounts ("Four govt agencies struggling to control porn on social media," 2014) . These attempts have failed because most of these tweets are created using slang, misspelled words, or words with multiple meaning to evade Internet censorship mechanisms. Also, the existing detection tools are targeting spammers who are using English language; however, such mechanisms are ineffective for detecting abusive accounts in Arabic language content (H.A. Wahsheh, Al-Kabi, & Alsmadi, 2013; Heider A. Wahsheh et al., 2012).

In this research, I analyzed three aspects of the subject on Twitter accounts: user's profile, social graph, and tweets. These aspects are divided into five feature sets: profile-based features, social graph features, tweet statistical features, tweet PageRank (PR) feature set, and tweet semantic orientation (SO) feature set. Below is a brief description of the feature sets used in this work:

- Profile feature set: reflect the user's interaction activities in Twitter, which contains features extracted from the profile page of a Twitter account.
- Social graph feature set: reflect the user's social graph connectives, which comprises features extracted from the social graph centrality measures.
- Tweet statistical feature set: reflect the user's tweeting characteristics, which extracted from counting each component in the tweet.

- Tweet PageRank feature set: reflect the lexical knowledge to understand the tweeting behavior of abusive accounts.
- Tweet Semantic orientation feature set: reflect the word meaning based on lexical semantics.

These feature sets were used to construct a multidimensional analysis model to detect variety characteristics of abusive accounts. These feature sets were created based on statistical analysis, social graph analysis, and lexical analysis. A simple statistical analysis was used to count the number of items in each tweet. The social graph analysis identifies the user connectivity relationships within the social media. The lexical analysis was used to correct the misspelled words in the tweets, find the co-occurrence relationship between words based on lexical knowledge, and find the concept of the word based on lexical semantics, which identifies the word meaning based on its closeness to either a positive or negative word.

Lexical knowledge is understanding the word meaning based its surrounding context (Lesk, 1986; McCarthy, Koeling, Weeds, & Carroll, 2004). This method has been combined with a graph-based approach and shows remarkable improvement on identifying word meaning (Agirre & Soroa, 2009; Mihalcea, 2005; Mihalcea & Tarau, 2004; Navigli & Lapata, 2007). In this research, I used the PageRank with weighted edge algorithm to rank each word in the tweet based on it used by abusive and non-abusive accounts.

Semantic analysis helps in understanding the meaning of words and the correlation between the words such as synonym, antonym, hyponym, or other associated

words (Cruse, 1986). In this research, lexical semantic analysis was used to identify the meaning of words based on their closeness to either abusive or non-abusive words.

Motivation

This research is motivated by the following:

- News events explaining frustrations of the people of Arab's countries to find a technique for blocking or detecting the accounts with profane and pornographic content ("Saudi govt. agencies struggling to fight porn on social media - Al Arabiya News," 2014).
- Twitter has policies against adult content in an image, a video, and a text; however, malicious Twitter users continue to violate these policies and post tweets with profane and pornographic content ("Adult or sexual products and services," 2017; Wagner, 2016).
- Preliminary research was conducted using sample data of 500 labeled tweet accounts analyzed using Bag-of-words (BOW) approach (E. A. Abozinadah, Mbaziira, & Jones, 2015). The results of the analysis revealed an improvement on the classifier performance when the features reduced from 3,000 to 100 features, which most of the words were useless by using BOW approach. Another drawback of this approach is that the model fail to generalize to datasets that were not part of the training set. I also observed it was difficult for the classifiers in the BOW approach to discriminate between properly spelled words and misspellings.

- Children may view obscenity and profanity in social media. Children 13 years and under easily fake their ages and gain access to adult content (Reporter, 2014). Research reported that three-quarters of children between the ages of 11-12 years had faked their ages and more than two-thirds of young people did not report offensive language in social media (Coughlan, 2016).
- There is abundant research on detecting nudity in images and videos (Lopes, Avila, Peixoto, Oliveira, & Araújo, 2009; Santos, Santos, & Souto, 2012). However, there is limited research on detecting abusive accounts by analyzing tweets. Some research uses hashtag, mentions, URL, or social graph only and ignores the tweet content for its complexity (Cheng, Xing, Liu, & Lv, 2015; Singh, Bansal, & Sofat, 2016).
- Arabic is a complex morphological language and has limited lexical analysis resources that require a method that overcome this limitation (Duwairi, Rehab M, 2007). In addition, tweets are full of slang and local dialect that do not exist in formal Arabic dictionaries, which can increase the challenge of analyzing Arabic tweet contents.

Purpose Statement

The purpose of this research is to build a multidimensional analysis model for detecting abusive accounts based on Twitter profile information, tweets and social activities to overcome the limitation of blacklisted keywords. Moreover, analyzing the tweets using independent lexical approaches to defeat the limitation of Arabic language analysis tools.

The model was generated based on live data that collected from Twitter to understand the behavior of abusive accounts with Arabic tweets. To the best of our knowledge there are no existing datasets for detecting abusive accounts with Arabic tweets. The concluded model would be suitable for Twitter or other operators to have these accounts either blocked or filtered.

Research Problem

Abusive accounts with Arabic tweets are not detectable by the current filtering systems that based on blacklisted keywords or reported accounts as these accounts are hiding on the crowd to look legitimate. Moreover, it is important to note that some decent words in abusive tweets can have an inappropriate meaning based on the tweet content. Also, informal Arabic language such as dialect, and slang are not recognized by lexicon resources. Furthermore, the tweet can contain misspellings that can lead to misunderstanding the meaning of the tweets, but correcting the misspelling words can give different meanings if not corrected based on the tweet content and the informal words. However, the misspellings can be one technique that abusive accounts use to bypass content filters and censorship in some countries.

This research aims to detect different characteristic of abusive accounts to build a multidimensional analysis model based on machine learning.

Research Questions

1. Can correcting the misspelled slang and informal words enhance the performance of the classifier?

2. How can I design an effective multi-dimensional classifier that incorporates both lexical and statistical techniques to detect the abusive accounts with Arabic tweets?
3. What set of features gives higher predictive accuracies for detecting the abusive accounts?
4. How can I characterize the tweeting behavior of abusive accounts?

Research Contribution

This research makes five major contributions and part of it was the pioneering work in detecting abusive accounts (Singh et al., 2016). First, the dataset was collected from Twitter using tools that were designed specifically for this research. The dataset was comprised of more than one million tweets. Additionally, the collected data was manually analyzed to establish a ground truth for this research.

Second, a misspelled correction approach was developed to fit with the tweet content. The method uses domain specific lexicons where the spell checker is based on a co-occurrence relationship between the words in each tweet. This approach showed an improvement in detection of abusive accounts compared to the existing spell checker by choosing the most frequent words or the closest words for the misspelling.

Third, the tweets were statistically analyzed, where each element of the tweet was counted to understand the tweeting behavior of abusive accounts. Moreover, statistical summary measures were used such as average, minimum, maximum, and standard deviation to reflect the tweeting behavior for each Twitter account.

Fourth, the tweets were analyzed using graph-based lexical knowledge to understand the meaning of tweet contents by weighting the co-occurrence relationship between words in the tweets. The PageRank algorithm with edge weights was used to study co-occurrence relationships between the words and understand the meaning of tweets. Each account had two PageRank results from two graphs. Each graph has words as nodes, the co-occurrence relationship between the words as edge, and the co-occurrence frequency as edge weight. One graph was built based on the abusive accounts and one based on non-abusive accounts. Using two PageRank results present the tweet meaning based on abusive content and non-abusive content.

Fifth, I analyzed the tweets using an independent lexical semantic analysis to understand the meaning of each word. The meaning of the words was based on finding the closeness of the word to a positive or negative word. When the word was close to a positive word in the graph, it reflected a positive meaning and vice versa. Therefore, we used a semantic orientation method and applied it to the collected tweets to understand the meaning of words based on the tweeting behavior of Twitter accounts.

CHAPTER TWO

LITERATURE REVIEW

Twitter is a popular social media platform in the Arab region. With over five million active users, the top three tweeting Arab countries since March 2014 are Saudi Arabia, Egypt, and Kuwait with 40%, 17%, 10% usage respectively ("Twitter in the Arab Region," 2011). Countries in the Middle East restrict and regulate the use of social media by the public and government employees to reduce the incidents of exploitation from spammers (Elbadawi, 2012). Spammers create and stockpile social media accounts, especially on Twitter because of its simplicity to create new accounts due to weak account opening and verification mechanisms ("The Twitter Rules," n.d.). Spammers use these accounts to launch spamming campaigns that contain profanity, curse words, adult content, promotion of child pornography and exploitation, and harassment (Singh et al., 2016). Spammers then disseminate targeted Twitter spam by exploiting weaknesses of the internet censorship and content filtering systems that use the blacklisted keywords, blacklisted URLs, and blacklisted spamming words (Chaabane et al., 2014).

In this chapter, overview of related works was covered. First, literature review of concepts about Arabic language and research on classifying Arabic words. Second, literature review on existing misspelling correction methods that deal with informal language and slang. Third, literature review of lexical analysis approaches. Last,

overview of several machine learning techniques that focus on detecting spam and adult content on Twitter.

Arabic Word Classifying

The Arabic language consists of 28 letters where each letter has a variety of shapes based on its location on a word. The direction of Arabic writing is from right to left, compared to other languages that use non-Arabic language systems reading from left to right. The words are segmented by a whitespace unlike the Farsi language that use the final form of letters instead of whitespace (Miangah, 2013). Arabic grammar also uses accent symbols to stress pronunciation and the meaning of words. Other forms of the Arabic language is Arabic slang, which has various forms across age-group and location within the Middle East (Versteegh, 2014).

In (Duwairi, Marji, Sha'ban, & Rushaidat, 2014), the authors use more than 25,000 labeled tweets were classified. The tweets were normalized using three domain specific lexicon dictionaries that translate the informal words to MSA. This research shows the benefit of understanding the informal words in the tweets. Interestingly, this work finds stemming the tweets by reducing each word to their base or root would weaken the classification accuracy.

In (Sallam, Mousa, & Hussein, 2016), the authors compared the results of using three datasets of MSA, namely: non-normalized and non-stemmed, only normalized, and only stemmed. The "only normalized" data set has the best result as it outperforms the other two data sets. Hence, normalization has a higher impact on the result than the stemmer.

Other studies (Rsheed & Khan, 2014) investigated the popularity of trending Arabic news instead of focusing on the popularity of words by comparing three classification algorithms: Decision Tree, Naïve Bayes, and rule-based classifiers to find features that increase the popularity of the trending Arabic news in Twitter. The features were divided into two types: external and internal. External features included the article source, website and the number of tweets that contain the article URL. The internal features include the title and the description of the article. In this research, internal features were weak and did not yield a good result because of the complexity of the Arabic language and lack of lists to indicate the popularity of words.

Another research effort evaluated two classification algorithms, Naïve Bayes and Support Vector Machines, to classify Saudi Arabian newspaper content (Alsaleem, 2011). The evaluation used three metrics: recall, precision, and F1 measure. Both classifiers registered good performance outcomes.

In a different effort, Arabic web data was classified into five categories, namely health, business, culture, science, and sport (El Kourdi, Bensaid, & Rachidi, 2004). The classification technique used was Naïve Bayes. The average accuracy rate was 68.78% for the experiments used in this research. This outcome reflected the challenges for a Naïve Bayes classifier to learn from Arabic text and successfully predict outcomes.

Another study used a supervised learning approach based on Support Vector Machines (SVM) to classify Arabic documents by comparing the result of stemmed and non-stemmed documents (Alsmadi, Al-Kabi, Wahbeh, Al-Radaideh, & Al-Shawakfa, 2011). The stemmed documents had approximately a 3% lower result compared to non-

stemmed documents. The researchers concluded that dealing with non-stem documents helps the classifier perform better in Arabic text classification problem. It is important to note that a 3% difference is not large, but it can be useful in deciding between the use of non-stemmed and stemmed documents given dataset properties such as limited word count within corpora like Twitter content.

Other studies used the Naïve Bayes algorithm with the Chi square features selection method to evaluation different Arabic text categorizations (Thabtah, Eljinini, Zamzeer, & Hadi, 2009). The dataset used in this research contained 1,000 features; however, the classifier registered the best performance when the dataset was reduced to 800 features. The result of reduced features demonstrates the benefit of using feature selection methods.

Another study used three classifiers, Naïve Bayes, k-Nearest Neighbors (kNN), and distance-based classifier to categorize 1,000 Arabic text corpus documents into 10 categories (Duwairi, Rehab M, 2007). In this study, the Naïve Bayes classifier outperforms the other two classifiers based on measures of recall, precision, error rate, and fallout measures. This study reflected the practicality of Naïve Bayes classifier in classifying Arabic text.

Other studies evaluated machine learning classifiers and built frameworks for addressing spam detection (H.A. Wahsheh et al., 2013). For instance, the authors built a framework to detect spam in Arabic opinions of the user feedback and comments on the web content or news. The framework had two categories and subcategories. The first category is the spammer and contains two subcategories: high level spammers and low-

level spammers. The second category is a non-spammer and contains three subcategories: positive, neutral, and negative. The user is considered a spammer if he or she uses a URL or five consecutive numbers. Therefore, if the user uses a legitimate URL to explain his or her opinions, it will count as a spammer, which can be a drawback to this study.

Arabic Word Normalization

One study developed an automatic spell checker for standard Arabic and Egyptian dialects (Shaalan, Allam, & Gomah, 2003). The study created different lists of common Arabic spelling errors to choose corrected words from several Arabic dictionaries. The first one was the Reading Errors list which contained a group of letters that are similar of each other. The second dictionary was the Hearing Errors list that contained a group of letters with similar pronunciation. The third dictionary was the Touch-Typing Errors list that contained a group of letters close to each other on the keyboard. The fourth dictionary was the Morphological Errors list that contained a list of common words based on Arabic morphology. The final dictionary was the Editing Errors list that deal with typing mistakes such as insertion, deletion, and substitutions. This approach corrects the word based on a detected error type, which may result of unfitting word correction for different dialects.

Another study divided each word into bigrams of two letters to develop an Arabic spell-checker (Muaidi & Al-tarawneh, 2012). Each bigram was given a score and the scores are for the letter location, such as at end of the word, anywhere in the word, or not in the word, which were assigned values of 2, 1, or 0 respectively. Each word was compared with a list of words with similar bigrams. A word was considered correct if it

has score of one for all the bigrams in the beginning and middle of the word, and a score of two for the last bigram; otherwise, the word will be considered wrong and has a score of zero. Unfortunately, this approach would correct the word without having any consideration to the word location on the sentence as the word can be corrected but unsuitable with the sentence.

In other research, a dictionary with more than 9 million words was used for Arabic spell correction (Shaalan, Attia, Pecina, Samih, & Genabith, 2003). A word is considered misspelled if it was not part of the dictionary list, then the Edit Distance algorithm was applied to retrieve a list of candidate words. Each candidate word was scored based on a noisy channel model that used a one-million-word corpus, and then the word with the highest score was chosen. This approach was applied to the MSA correction corpus, but it did not cover the informal Arabic word corpus.

Another study created eleven candidate patterns of polarities in tweets with the Egyptian Arabic language, which overcame the weakness of existing parts of speech in dealing with colloquial Arabic (Elsahar & El-Beltagy, 2014). That research studied the Arabic pattern from Egyptian slang; however, Arabic slang differs within different cities in Egypt and between different Arab speaking countries. Existing lexicon tools dependent on Arabic language do not support slang and misspelled words; therefore, using an independent approach could overcome this issue.

Lexical Analysis

One prior study used a domain specific lexicon to classify the reviews as thumbs up if the phrases' average was closely associated with word “excellent,” (i.e., five stars),

and thumbs down if the phrases' average is closely associated with word “poor” (i.e., one star) (Turney, 2002). Part of Speech (PoS) tags were applied to identify a set of patterns to extract two-word phrases from the reviews. The researcher then use Pointwise Mutual Information (PMI) to define the probability that the two words occurred together as follows (Turney, 2001):

Equation 1. Pointwise Mutual Information (PMI)

$$PMI(w_i, w_j) = \log_2 \left[\frac{P(w_i, w_j)}{P(w_i) * P(w_j)} \right]$$

Where $P(w_i, w_j)$ defines the probability of the words co-occurring in the document, and $P(w)$ defines the probability of the word in the document. PMI measures the degree of statistical dependence between the words. The Semantic Orientation (SO) of the phrase is calculated as shown in Equation 2 below. The review polarity is based on the average SO of all phrases in each review. This study used a lexical semantic technique that estimated the word meaning based on its closeness to positive or negative words, which is a practical approach to learn the meaning of slang or unknown words.

Equation 2.The Semantic Orientation (SO)

$$SO(w_i) = PMI(w_i, \text{excellent}) - PMI(w_i, \text{poor})$$

Another study created a domain specific lexicon and compared it with a general purpose lexicon on tweets (Tai & Kao, 2013). The domain specific lexicon is based on a co-occurrence graph, where the nodes were the words and the edges were the similarities between the words. The similarities between words constructed were based on three different dependent lexicon resources from WordNet (Miller, 1995), Conjunction rules (Hatzivassiloglou & McKeown, 1997), and SOC-PMI (Turney, 2001), where the polarity of the word graph was assigned based on negative and positive seeds. This study demonstrated the advantage of using domain specific lexicon polarities with a focus on a specific domain.

Other researchers used linguistic analysis to identify conversational tweets from non-conversational tweets for emergency responses of disaster events (Purohit et al., 2013). The conversational tweets focused on the form of replies, retweets, and mentions, which used as a conversational indicator. The conversational tweet was classified correctly by applying simple heuristics features from pronouns, dialogue management, word count, and pre-defined categories of words. This approach suggests that the use of statistical analysis would improve the classification performance when dealing with informal text.

Other researchers designed a social media offensive language detection model on lexicon and users' language profile features using Lexical Syntactic Feature (LSF). Lexicon features are based on three aspects: offensive word dictionary, syntactic intensifier, and offensive value. The dictionary was based on the Urban Dictionary ("Urban Dictionary," 1999), which was based on slang and informal words from different

domains and had a limited number of offensive words(Chen, Zhou, Zhu, & Xu, 2012). Some offensive words could not be recognized by the dictionary as offensive when it appeared in some domain versus others. Therefore, using an independent lexical approach to identify the co-occurrence relationship between the words would identify the category of the unknown word based on the closeness of the word to either an offensive word or a decent word.

Moreover, other researchers used labeled tweets from three languages and used them to bootstrap Twitter specific lexicons (Volkova, Wilson, & Yarowsky, 2013). The approach was based on a semi-supervised learning approach, using several subjective seeds for each language to identify the subjective tweets, and label the tweets as positive or negative based on the appearance of the similar words from the previously labeled tweets. The new words *t* from unlabeled tweets and that are not on the list of words from labeled tweets are added to the word list based on the probability of the identified polarity of the tweet. However, the measure of similarity would perform better on a blog that has no misspelling corrections (E. Abozinadah & Jones, 2016).

Another study proposed TextRank approach that use the PageRank algorithm with link weighting to extract keywords and sentences from documents (Mihalcea & Tarau, 2004). Co-occurrence relations was used to locate the words on the graph, where the words are a node in a graph, the edge is corresponding to words co-occurring with a distance set of words, and the edge weight is the frequent co-occurrence of the words. The result reflected better performance using the PageRank algorithm with link weighting than using word frequency count. That approach showed the benefit of using

PageRank algorithm other than ranking webpages. In addition, the benefit of using the graph as a lexical knowledge approach is to identify the most important keyword in the document.

Another study used direct graph-based key phrase extractor without any language dependent methods (Litvak, Last, Aizenman, Gobits, & Kandel, 2011). The graph was constructed by using nodes representing words, edges representing the order relationships between words, and a unique identifier number for each sentence to be assigned to the edge. The keyword was the most connected node to other nodes, and the key phrase was the sequence of up to three nodes with the higher score. This approach outperforms the TextRank (Mihalcea & Tarau, 2004) and GenEx (Turney, 2000) when it came to multi-language and overpassing the lexical analysis limitation tools in some languages.

Other research manually identified the polarity of subjective expression based on emoticon, hashtags, negation words, where the negation words are used to convert the polarity of the word (Palanisamy, Prabu, Yadav, & Elchuri, 2013). The study contains 9,451 subjective expressions from Twitter, and the polarity for each expression is the sum of the sentiments from all the entities, where the expression counts as positive when the sum is larger than zero, negative is smaller than zero, and zero is neutral.

Machine Learning and Social Media

Analyzing tweet content would enable better detection of the spammers. In (Singh et al., 2016), the authors used six profane keywords in Twitter searches to collect a pornographic spammer dataset. The dataset contained more than 73,000 tweets and more than 18,000 users. The result of analyzing the tweets' content distinguished spammer

from non-spammer accounts, and had a better performance than using the profile information alone. Therefore, the number of the followers and following of the spammers did not show any difference from a celebrity account from an unknown account, but the tweet content reflected the spammer's behavior uniquely.

Shekar et al. analyzed pharmaceutical spammers on Twitter, where the study showed improved results by using two lists of words instead of one list (Shekar, Liszka, & Chan, 2011). The first list was the name of the product, and the second list was the words associated with the products, for example organic, tablet, refill, etc. The classifying result had less false positives when the second list was used.

Irani et al. studied the top trending topic on Twitter by connecting the tweet content to the URL content on the web (Irani, Webb, & Pu, 2010). The findings showed that spammers were using the top trending topic in their tweet as either a hashtag, or text. Additionally, the URL content was not related to the tweet topic, and the study used information gain measures to reduce the noise features to improve performance of the classifiers.

Wang et al. detected the unidentified spammers on social media by studying suspended spammers' accounts (D. Wang, Irani, & Pu, 2011). In their approach, they first matched the URLs, IP addresses, and hashtags with suspended account data to predict the spammers' accounts. Then classified profile data, text data, and the webpage content to determine similarities between spammer behavior in different social networks. One of the finding shows that the spammers use profane words in the online community by

replacing letters with symbols to bypass filtering systems, which present the usefulness of correcting the misspell words to detect the spammer accounts.

Yoon et al. determined the correct spelling of profane words that have symbols in them (Yoon, Park, & Cho, 2010). Each word was checked against a list of regular words. If not identified, the word was checked against a profane word list. If not recognized in this list either, then a similarity letter process was applied. The similarity process checked for symbols in the word and replaced them with corresponding letters from a list of letters with matched symbols. After replacing the symbols, the word was again checked against the regular word list and the profanity word list. This study presents the benefit of understanding the chatting behavior and correcting the word accordingly to its domain to have meaningful correction and not lose the meaning of the word.

In Benevenuto, et al., spammers exploited trending topics to have their tweets visible and have a higher chance of creating more traffic to their malicious URL (Benevenuto et al., 2010). They studied characteristics of tweets and user behavior to predict the spammers and non-spammers who are using the top trending topic on their tweets. This study focused on English language trending topics and ignored other languages. The evaluation was conducted by using a Support Vector Machine (SVM) that detects 70% of the spammers and 96% of non-spammers. This study showed how the spammers post unrelated topics in the tweet to get more traffic, and to hide their malicious posting activities under legitimate topics.

In other research, the authors evaluated user-based and content-based features to distinguish between spammer and non-spammer accounts (2011). This study used four

different classifiers that include Random Forest, SMO, Naïve Bayesian, and k Nearest Neighbors. The Random Forest outperformed the other classifiers. One of the user-based features used in this study was the distribution of tweets over a 24-hour period, where the authors suggested that the spammers are tweeting during the morning hours, while the non-spammers are less active during the night. This feature could be misleading the classifier, because the spammers do not present true information about their location. In addition, this study used 100 recent tweets to classify the spammers' accounts, which would take more than four days of tracking the tweeting behavior as the average tweeting per user per day is 22 tweets (Zarrella, n.d.).

Moreover, researchers in (Mbaziira, Abozinadah, & Jones Jr, 2015) used machine learning to evaluate a criminal network of 419 bilingual scams in Facebook by using two data sets, one containing English language comments only, and the second one containing English and Nigerian Pidgin. The evaluation performance of using bilingual comments had the better result, even on sub-dataset with unigram and bigram words. Therefore, this research presents insight into using classifiers to analyze bilingual cybercriminal behavioral.

In another study, detected spammers took advantage of top trending topics in Twitter to spread malicious links (Irani et al., 2010). The detection mechanism was based on building a set of 12,000 features of the tweets and 500,000 features of the associated web pages. Information gain method was used to reduce the features and have a more accurate classification measure. Each features of the dataset were reduced to 1,000 features and 5,000 features, which are 91% and 99% respectively. Datasets of 100 and

1,000 features were constructed and tested on three classifiers: Naïve Bayes, C4.5 Decision Trees, and Decision Stump. The classifiers performed better on the 100 features dataset compared to that with 1,000 features.

Other researchers used a machine learning algorithm with nine features to detect spammers distributing pornographic content in Twitter (Singh et al., 2016). Their approach achieved an accuracy rate of 91%. The spammer's dataset was collected based on six keywords related to adult content and identified any account as spam if it used a sexual word even in a legitimate context. However, this approach ignored the tweet content for its complexity and the model was not evaluated on testing set with unknown type.

In another study, the researchers proposed a graph-based and collective correlation model to detect adult content accounts in Twitter (Cheng et al., 2015). The graph contained two types of nodes: Twitter accounts and the tweets' entities that included hashtags and mentions on each tweet. The collective correlation model was used to distinguish between the normal and adult content accounts based on the account's neighbors and the account's hashtags. The account is considered to have adult content if it is followed by accounts that are following many other known adult content accounts and has some adult content hashtags. This approach ignores the tweet content except the hashtags for the simplicity. In addition, this approach would misclassify a normal account as an adult content account if it is followed by Twitter account, whose users are interested in adult content accounts and the account used some adult hashtags for teaching purposes.

CHAPTER THREE

METHODOLOGY

This research used a multi-dimensional analysis approach to detect abusive accounts with Arabic language tweets. The analysis is based on the account profile information, the account tweets, and the account social graph. The benefit of using a multidimensional analysis approach is to detect as much of the abusive accounts' behavior as possible. These accounts use many techniques to appear to be legitimate regarding the Twitter adult content policy and to avoid being reported by other users.

Moreover, misspelling correction method based in Twitter content was proposed to improve the classifier performance. The misspelling correction approach was based on the tweet content to correct the misspelled slang that does not exist in a standard Arabic dictionary. Additionally, the tweet was analyzed by using independent lexical analysis that did not use lexicon tools such as natural language processing (NLP), name entity (NE), or blacklisted words to overcome the limitation of Arabic language analysis tools.

The diagram in Figure 1 shows each component of the proposed methodology. The research started by collecting data from Twitter and manually analyzing part of the data as abusive accounts and non-abusive accounts. The data was organized based on five types of feature sets: Profile-based, Tweet content statistical based, Tweet PageRank

based, Semantic Orientation based, and Social graph-based. The dataset has been preprocessed to eliminate the noise and correct misspelled words.

This approach extracted the features from each set and trimmed them using feature selection to eliminate the noisy features. Three common machine learning models have been used: Naïve Bayes (NB), Support Vector Machine (SVM), and Decision Tree (J48) to find the model that can detect the abusive accounts with minimum error. The dataset was divided into a training and testing set to evaluate the proposed model with a previously unseen dataset. The classifier performance evaluation was based on five measures: accuracy, precision, recall, f-measure, and ROC. Each process is explained in the following sections.

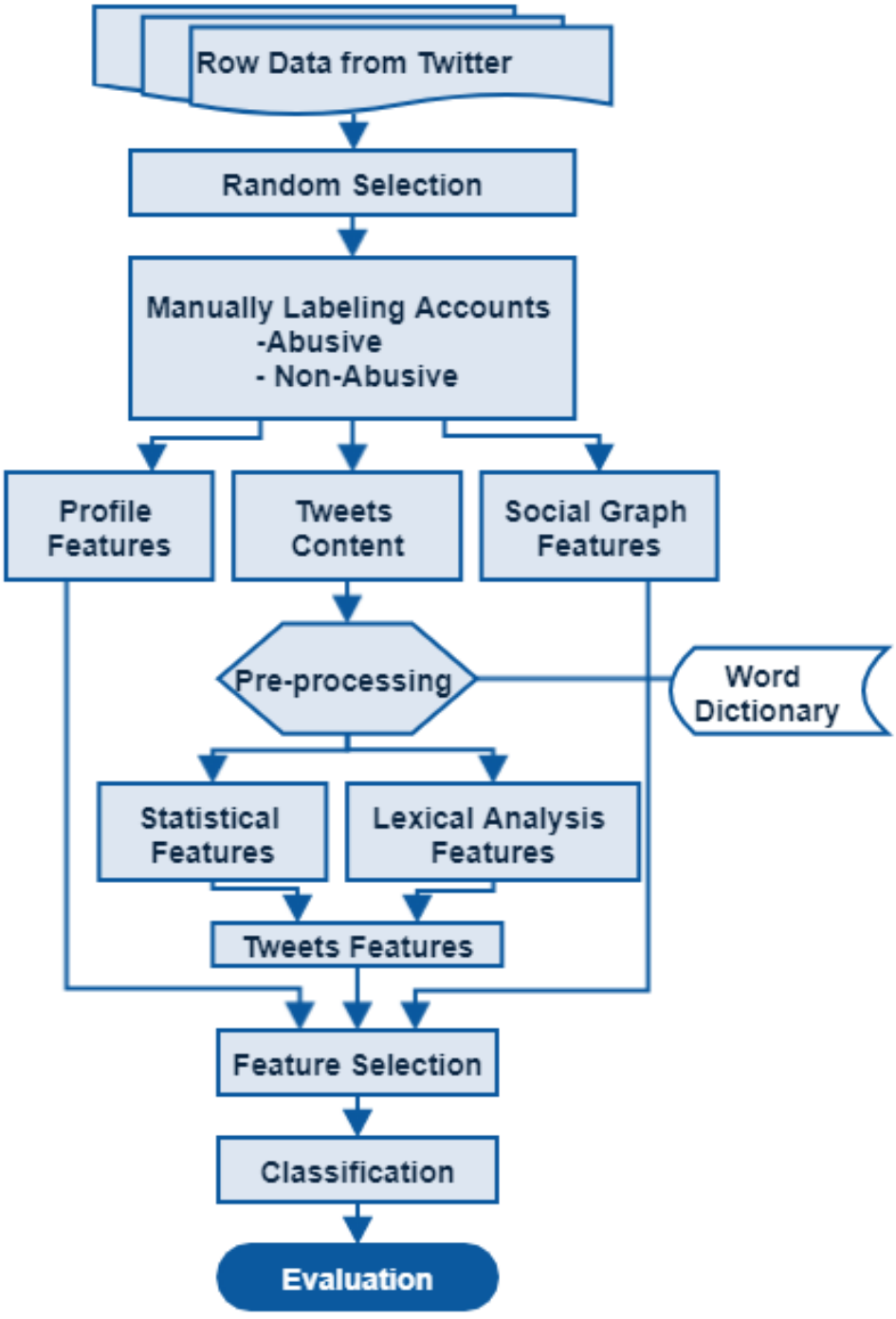

Figure 1. Diagram of Multi-Dimensional Analysis Approach

Data Collection

This research used supervised learning approach, where each record in the dataset was mapped to a class label. The data for this research was collected from Twitter by using a customized Python tools to scrape data from Twitter without using the restricted API. The data was collected for a period of three months from April 1, 2014 to June 30, 2014. The data collection started by using the top five Arabic insulting words obtained from a website with Arabic insulting words ("how do I swear in Arabic from insults.net," 1999) and is presented in APPENDIX A. These words were used as searching seeds in the Twitter search engine, which the most resent 800 tweets were collected for each search query. However, all collected tweets were tweeted by 255 unique Twitter accounts. Furthermore, the follower, followings, profile information, and the most 50 recent tweets were scraped that include Arabic words, English words, numbers, characters, hashes, mentions, and links to have the full social network of the abusive accounts. Moreover, the same process had been applied to collect information about the followers and the followings of the followers' seed accounts. A summary of the data collected is shown in Table 1.

Table 1. Summary of Collected Data

Type of Content	Total
Seed Words	5
Main Accounts	255
Accounts	350,000
Tweets	1,300,000
Hashes	530,000
Links (URLs)	1,150,000
Followers	925,000
Followings	19,000

Manually Analyzed Data

From the collected data, 2500 accounts were randomly selected, where each account tweeted more than 100 tweets. These accounts were manually analyzed by three graduate students who labeled the accounts based on the content of the tweet as an abusive account, non-abusive account or unknown. The abusive accounts are those with tweets contain Arabic profane words, Arabic obscenity, Arabic insult words, or had sexual meaning out of the Arabic tweets. However, images, URLS, and the tweets in other languages (not Arabic) were ignored. The analyzers had to detect five tweets that had abusive attention which have been posted on two different days or more to identify the account as an abusive account. The non-abusive accounts contained tweets without abusive words. The unknown accounts were the accounts with tweets in other languages than Arabic, have URLS, pictures with no text, or it was unclear for the analyzers.

Analyst aggregation is based on voting, where a minimum of two analyzers should agree on the type of the account (abusive or non-abusive). From the manually analyzed accounts, a balance dataset of 400 non-abusive accounts, and 400 abusive accounts, with 50 recent tweets for each account were randomly selected. These accounts were divided into training and testing as 80%, and 20% respectively to evaluate the performance of the classifier. The training set was used across all stages of implementing and improving the feature set, and classifiers. The testing set was only used once at the last classifier evaluating process to ensure our approach was applicable for use with unseen data.

Moreover, the cross validation method was used on the training set to assist the performance of the classifiers. In addition, the training set was used to construct three graphs that include the co-occurrence words graph of abusive accounts, the co-occurrence word graph of non-abusive accounts, and the graph of the words that associated with identified non-abusive word or abusive word, which two graphs used for tweet PageRank method, and one for semantic orientation method, respectively. It also was used on the developed corpus based misspell correction for this domain.

Data Preprocessing

The tweets were normalized using two steps. The first step was cleaning up the tweet, and the second step was correcting the misspelled words on the tweet.

Cleaning up Tweets

This step was to reduce the noise on the tweet and keep the words on Arabic language by applying the following steps:

- Removed all non-Arabic words.
- Removed all symbols.
- Removed all digits.
- Removed all the stop words by using the stop word list in ("nltk.stem.isri — NLTK 3.0 documentation," n.d., "PyArabic 0.5 : Python Package Index," n.d., "stop-words - Stop words - Google Project Hosting," n.d.).
- Removed all diacritics.
- Removed all extra whitespaces.

- Removed all sequences of letters in the words except the name of God (Allah-الله) (E. A. Abozinadah et al., 2015). The sequences of letter in Arabic words were commonly used in casual manner because Arabic language does not have a capitalizing letter. Therefore, the user used sequences of letters on the word to emphasize the point or anger. Table 2 shows the number of sequence letters for each letter on the collected tweets.

Table 2. Set of Sequence for Each Letter

Letter	Sequence Set	Letter	Sequence Set
ل	700	ش	10
ض	132	س	10
خ	22	ي	92
هـ	454	ب	47
ع	22	ا	494
غ	2	ت	86
ف	6	ن	37
ق	13	م	607
ث	2	ك	58
ص	132	ؤ	15
ض	1	و	316
		ز	7

- Replaced all numbers in the words with corresponding letters as shown in Table 3. Using numbers in words is a way of typing in social medias and chat rooms (Saleem, 2014).

Table 3. Numbers with Associated Letters

Numbers	Letters
2	ء
3	ع
4	ذ
5	خ
6	ط
7	ح
8	هـ
9	ص
‘9	ض

Correct Misspelled Words

A domain-specific lexicon word correction approach was used for choosing the correct words from candidate words based on the content of the tweets. As shown in Figure 2, the process has three main phases: identifying misspelled words with corresponding candidate words, building a list of n-gram words with their frequency, and lastly choosing the correct word.

Identifying Misspelled Words

Each word was compared against two word dictionaries that contain data from the Arabic Hunspell dictionary containing 300,000 words (“Ayaspell project,” n.d.), and The Mo3jam dictionary for Arabic crowed source dictionary containing 9,595 words (“Mo3jam,” 2013). The string matching is based on Levenshtein Distance algorithm (Levenshtein, 1966) where the edit distance of 0 is an exact word match from the dictionary. The words that match with the word dictionary lists are considered correct and no further processing is required; otherwise, the word is considered misspelled and will have a candidate word list.

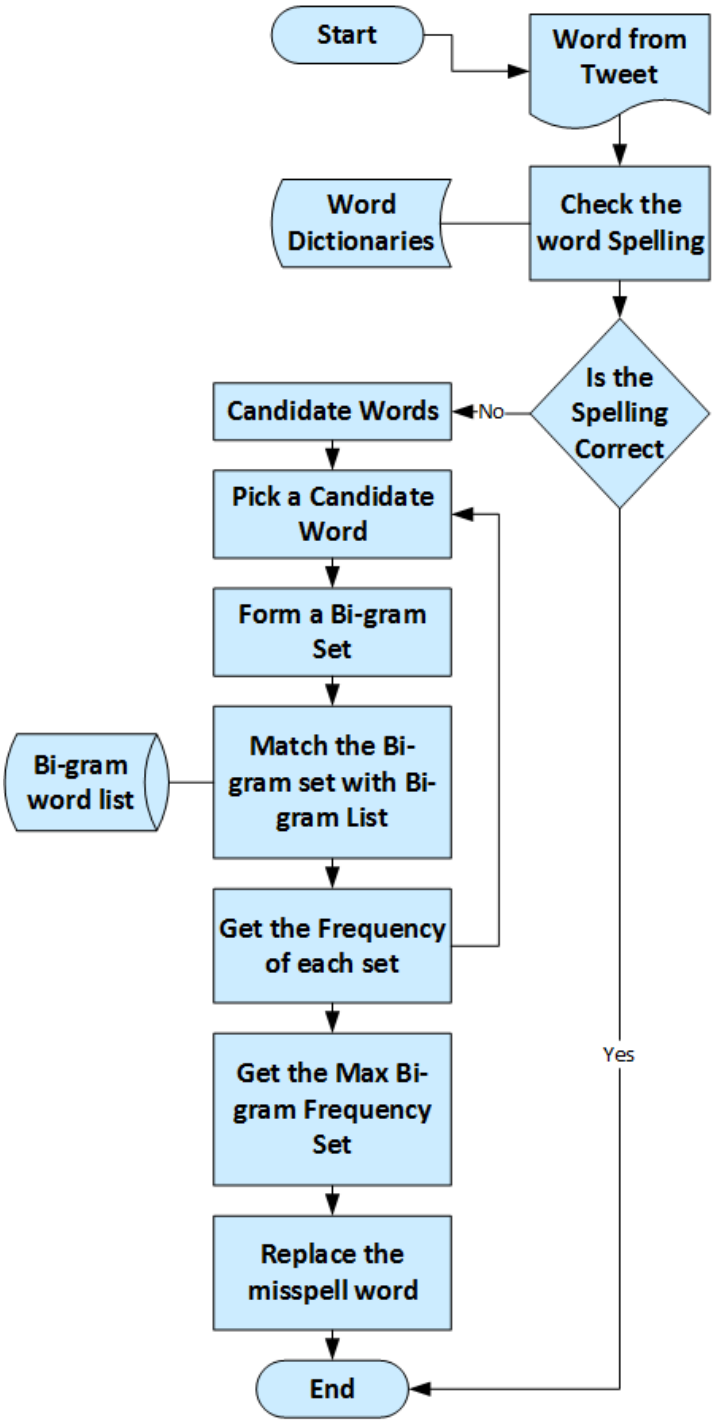

Figure 2. Word Correction Approach

The candidate word list is based on the following operation in each letter on the word: insertion, deletion, substitution, and transposition. Insertion is adding one letter in different places to the word. Deletion is removing a letter from the word. Substitution is replacing a letter with another letter in the word. Transposition is changing the letter's place with another letter in the word. These operations lead to find matching word list, that fit the tweet contents. I used Edit distance of 1 to have limited number of words on the candidate word list that will correct the word with one error and ignore words with multiple spelling mistakes as it can lead to different meaning. This approach corrects the misspelled words with a word that matched the tweet content, but does not replace a wrong word that is correctly spelled.

N-gram word with Frequency Count List

In this phase, I prepared a n-gram words list to choose the correct word that fit the tweet meaning. Each tweet was divided into n-gram words and counted the frequency of each n-gram words. This n-gram words list was used to pick the correct word out of the suggested candidate word list.

This list was built by using 1,300,000 tweets that came from the dataset that explained in our research. The correction result of three different sizes of n-gram were compared that include unigram (1-gram), bigram (2-gram), and trigram (3-gram) to choose the right size of n for the n-gram list. Randomly 300 tweets were picked that had misspelled words and ran them against the three n-gram word lists. The spelling correction of each set was evaluated manually by three graduate students. The

performance of each set was analyzed by counting the number of misspelled words that were replaced and the number of replaced words that fit the tweet's meaning.

Table 4. N-Gram List Comparison

	Replaced misspelled	Replaced with correct word
Uni-gram	89%	64%
Bi-gram	80%	91%
Tri-gram	10%	33%

Table 5. Tweet with Misspelling that Corrected by N-Gram Words List

Tweet	Correct	Uni-gram	Bi-gram	Tri-gram
الف مبروك فوز المنتخب العراق العراق بجميع طوائفه وقومياته يمثله هذا المنتخب	ألف	فلا	ألف	الف
الحمد لله رجاء و رخاء و شدة و طاع و الحمد لله يوما و شهرا و عمرا	طاعه	طالعه	طاعه	طاعه

Table 6. Summary of Bi-Gram Words List

Number of tweets	1,300,000
Number of bigrams in word list	2,000,000
Bi-gram word sets with frequency >=5	100,000

As shown in Table 4, the first column represents the percentage of misspelled words that were replaced with correct words, whereas the rest of the misspelled words were not replaced because there were no matching sequences of words in the n-gram list. The results showed the trigram words list is not effective on matching three words and finding the correct word, while the other two n-gram words lists replaced more than 80% of misspelled words. However, the second column presents the percentage of the

corrected words that fit within the tweet, as the misspelled word could be corrected by using a word that changes the tweet's meaning. For example, in **Error! Reference source not found.**, the unigram replaced the misspelled word “thousand” (الف) which is missing Hamza(أ), with word “Don’t” (فلا), but the bigram list detected the misspelled part and corrected it. The bigram word list replaced the misspelled word with the correct word more accurately than the unigram word list as shown in Table 4. Based on the comparison of the three sizes of n-gram word lists, the bigram word had the highest percentage of correct correction that matches the tweet's meaning.

The total tweets shown in Table 6 produced 2,000,000 bigram words. This list contains 5% of bigram words with frequency of five or more, and the rest of the list had a frequency less than five. Therefore, The top 5% of bi-gram words were used in correcting the tweets as the rest of words are rarely appeared in tweets.

Choosing the Correct Word

This approach assigned the most suitable word from the suggested candidate word list to replace the misspelled word. The process of choosing the correct word was based on replacing the misspell word by one word from the suggested candidate word list at time. The replaced word was used with the next word in the tweet to form a bigram words set. Each bigram words set was compared against the bigram word list to find the most suitable word. The set with higher frequency was used as the corrected word. However, if the word was not part of the bigram word list, the word was identified as unknown, and not replaced.

The word dictionary has a limited word set that does not cover all dialects and slang. The tweets contain some informal Arabic words that cannot be found in MSA dictionary and crowd source dictionary. Therefore, using a domain specific lexicon method can overcome the limitation of dictionaries.

Features

In this section, each feature was explained that I extracted from Twitter accounts. The features have been generated based on three analyses processes that include statistical analysis, social graph analysis, and independent lexical analysis. Based on these analyses five feature sets were extracted that include: one profile-based feature set contain four features, one social graph-based feature set containing nine features and three tweet-based feature sets. The tweet-based feature set includes tweet statistical feature set containing 78 features, tweet PageRank feature set contain 9 features, and tweet semantic orientation feature set contain 4 features.

Moreover, statistical summary measures were used that included average, minimum, maximum, and standard deviation in each feature set. These measures reflect the overall behavior for each Twitter account. The expected reflection is explained as the following:

- Average (Avg): it reflects how frequently the item appeared in the account's tweets.
- Minimum (Min): it reflects how infrequently the item appeared in the account's tweets.

- Maximum (Max): it reflects the highest frequency of the item in the account's tweets.
- Standard deviation (Std): it reflects the variation in a distribution of an item appearing in the account's tweets.

Profile-Based Features

Profile-based features are properties extracted from account information on the home page of each account as show in Table 7. The profile objects comprise the number of the tweets, followers, and following. Also, obtain the account reputation score that reflect the ratio of followers to followings as the Twitter accounts with number of following higher than number of following would consider spam. The reputation score is between [1,0] where score closer to 1 reflects a higher reputation as the account has large number of follower than the number of following. The score closer to 0 reflects low reputation as the account has smaller number of follower than the number of following. The reputation is calculated based on the formula below (A. H. Wang, 2010):

Equation 3. Reputation

$$\text{Reputation} = \frac{\text{Followers}}{(\text{Followers} + \text{Followings})}$$

Table 7. Profile Based Features

Features
Number of tweets
Number of followers
Number of following
Reputation

The features from profile information have shown its effectiveness on identifying the user type in much research such as detecting the spammer in Twitter (Thomas, Grier, Song, & Paxson, 2011), identifying the celebrities in Twitter (Marwick & boyd, 2011), or detecting fake accounts in Twitter (Thomas, McCoy, Grier, Kolcz, & Paxson, 2013). For instance, the celebrities would have more followers than following, whereas the opposite is true of the scammers. Also, the scammer would have almost an equal number of followers and following as the scammers try to follow many users as theses users would following them back (A. H. Wang, 2010).

Tweet-Based Features

Twitter is micro-blog social media, which the user activities are presented on the tweet or around the tweet. In the tweet is the way they write it and around the tweet is how they share the tweet with others. Therefore, Tweet-based features were conducted to analyze the content of each tweet that includes the text, hashes, links, URLs, pictures and mentions. These features were conducted based on statistical analysis and lexical analysis approaches, which are generated onto three feature sets: tweet statistical feature set, tweet PR feature set, and tweet OS feature set.

Tweet Statistical Analysis

The statistical analysis is based on counting each element in the tweets to understand the behavior of the abusive accounts. The statistical summary measures explained above will be used for each account to have overview of tweeting behavior of each account. The elements extracted from the tweet are:

- Stop-words.
- Arabic words in the tweet.

- English words in the tweet.
- Numbers in the tweet.
- Characters in the tweet.
- Letters in the tweet.
- Hashtags in the tweet.
- Mentions in the tweet.
- Diacritics in the tweet.
- Slang Arabic words in the tweet.
- Formal Arabic words in the tweet.
- Unknown Arabic words in the tweet.
- Correct Slang words in the tweet.
- Sequence of letters on the words in the tweet.

These elements were measured to build the tweet content statistical feature set that are presented in Table 8.

Table 8. Tweet Content Statistical Feature Set.

Features	
Count_pic_max	Count_number_std
Count_pic_min	Count_slang_max
Count_pic_avg	Count_slang_min
Count_pic_std	Count_slang_avg
Count_tweet_letters_max	Count_slang_std
Count_tweet_letters_min	Count_stand_max
Count_tweet_letters_avg	Count_stand_min
Count_tweet_letters_std	Count_stand_avg
Count_clean_letters_max	Count_stand_std

Count_clean_letters_min	Count_unknown_max
Count_clean_letters_avg	Count_unknown_min
Count_clean_letters_std	Count_unknown_avg
Count_ar_max	Count_unknown_std
Count_ar_min	Count_correct_slang_max
Count_ar_avg	Count_correct_slang_min
Count_ar_std	Count_correct_slang_avg
Count_en_max	Count_correct_slang_std
Count_en_min	Count_correct_stand_max
Count_en_avg	Count_correct_stand_min
Count_en_std	Count_correct_stnad_avg
Count_hash_max	Count_correct_stand_std
Count_hash_min	Count_sequence_max
Count_hash_avg	Count_sequence_min
Count_hash_std	Count_sequence_avg
Count_http_max	Count_sequence_std
Count_http_min	Count_stop_max
Count_http_avg	Count_stop_min
Count_http_std	Count_stop_avg
Count_mention_max	Count_stop_std
Count_mention_min	Count_correct_letters_max
Count_mention_avg	Count_correct_letters_min
Count_mention_std	Count_correct_letter_avg
Count_non_max	Count_correct_letters_std
Count_non_min	Count_diacritize_words_max
Count_non_avg	Count_diacritize_words_min
Count_non_std	Count_diacritize_words_avg
Count_number_max	Count_diacritize_words_std
Count_number_min	Mention_rate
Count_number_avg	Hash_rate

Tweet Lexical Analysis

The Arabic lexicon analysis had limitations (Al-Sughaiyer & Al-Kharashi, 2004; Black et al., 2006; Farghaly & Shaalan, 2009; Mamoun & Ahmed, 2016) and encouraged us to propose a method that is lexical independent. Moreover, preliminary experiment

was conducted by using a BOW approach to detect the abusive accounts, where the result shoed the abusive accounts having their own way of tweeting. Their tweets included a more informal Arabic language with vague words that were meaningless when alone, but had abusive meanings when the full tweet was analyzed. Therefore, Tweet PageRank approach was proposed based on PageRank algorithm to identify the tweet meaning based on independent lexical knowledge from a graph, and tweet Semantic Orientation approach to identify the word meaning based on semantic lexical by measuring the word closeness to either positive or negative word. Each lexical method has a set of features, which is explained in detail in the following sections:

Tweet PageRank (PR) Feature Set

The PageRank (PR) algorithm was developed by Larry Page, one of the founders of Google (Page, Brin, Motwani, & Winograd, 1998). This algorithm ranks the importance of websites in a search engine index. In addition, this algorithm has been used in different environments, such as keyword extraction (Litvak et al., 2011), sentence extraction (Mihalcea & Tarau, 2004), and influences on Twitter (Wu, Hofman, Mason, & Watts, 2011).

This algorithm measures the importance of website pages by assigning a weight to each website (Brin & Page, 1998). The score of each website is defined by the following formula:

Equation 4. PageRank Algorithm

$$\mathrm{PR}(V_i) = (1 - \mathrm{d}) + \mathrm{d} * \sum_{j \in In(V_i)} \frac{\mathrm{PR}(V_j)}{\mathrm{Out}(V_j)}$$

Where $\mathrm{PR}(V_i)$ is the PageRank of page V_i, d is the damping factor that is set between [0-1], V_j is the page's links to page A, and Out (V_j) is the outbound links on V_j. The dumping factor is the probability of jumping from one page to another, which is usually set to 0.85 (Brin & Page, 1998).

The main benefit of using PR is to have a value for each word to reflect the influence of the word in the tweet. In addition, some words have different meanings based on the sentence structure. Moreover, using a PR with edge weight algorithm would measure the co-occurrence relationship between the words. In this study, the PageRank with edge weight algorithm was used, which presented by (Mihalcea & Tarau, 2004) and the calculation formula is shown below:

Equation 5. Weighted PageRank with (WPR)

$$\mathrm{WPR}(V_i) = (1 - \mathrm{d}) + \mathrm{d} * \sum_{j \in In(V_i)} \frac{w_{ji}}{\sum_{k \in \mathrm{Out}(V_j)} w_{jk}} WPR(V_j)$$

The algorithm reflects the importance of the co-occurrence relation between the words in the tweet. Two graphs of the tweets' normalized words were used to conduct two PR score for each word. First graph was based on words from the tweets of labeled

abusive accounts and the second graph was based on words from the tweets of labeled non-abusive accounts. The content of each graph is presented in Table 9.

The graph G = (V, E) where V is a set of nodes for each unique word in the tweets. E is the edge between nodes that reflect the co-occurrence relationship between each word with the next word in each tweet. The edges on the graph are weighted based on the frequency co-occurrence relationship between each word. The weight reflects the weakness or strength between the two nodes, which can differentiate the most common co-occurrence words from the rare ones.

Table 9. The PR Graphs content

Graph	Nodes	Edge
Graph with words from abusive accounts	15395	41422
Graph with words from non-abusive accounts	54853	170230

For instance, abusive words had higher PR from the graph with words from abusive accounts than the graph from non-abusive accounts. The PR of the tweet was calculated by taking the average PR of ranked words in the tweet. Furthermore, the PR of the Twitter account was conducted by applying the statistical summary measures for all tweets.

The overview of each graph is presented on Figure 3. The word is represented by circles, the co-occurrence relationship is represented by arrows, the frequency of co-occurrence relationship is represented by the thickness of the arrows and the word influence represented by the size of the circles.

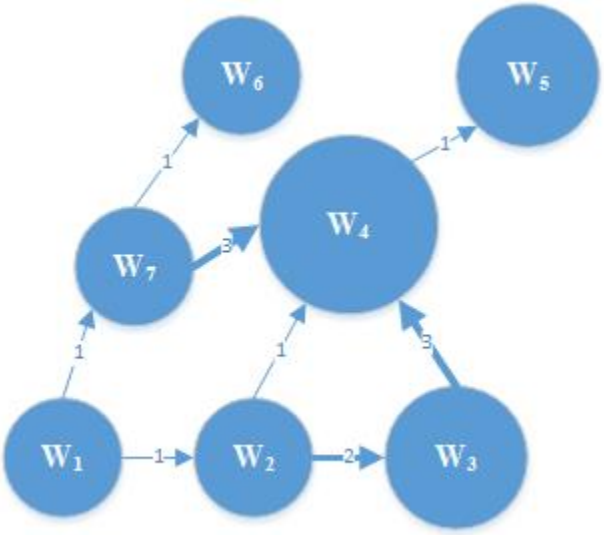

Figure 3. Co-occurrence Word Graph

Tweet Semantic Orientation (SO) Feature Set

The Semantic orientation of each extracted phase finds the closeness of the word to a positive word or to a negative word (Hatzivassiloglou & McKeown, 1997). If the association word was positive the word would be considered positive and vice versa for negative.

The SO was calculated based on the PMI-IR method that was presented in (Turney, 2002). The PMI-IR algorithm contained in Point-wise Mutual information (PMI) and Information retrieval (Turney, 2001) that measured the similarity between two words. A given word would be associated with a positive word to measure the similarity of the word to the positive word, and would be associated with a negative word to measure the similarity of the word to the negative word. The direction of the phrase's

semantic orientation would be based on the equation below that was proposed in (Turney, 2002).

Equation 6. Semantic Orientation

$$SO(Phrase) = \log_2 \left[\frac{hits\ (Phrase\ Near\ "Positive")\ hits("Negative")}{hits\ (Phrase\ Near\ "Negative")\ hits("\text{Positive}")} \right]$$

Where $hits\ (Phrase\ Near\ "Positive")$ is the frequency of the given word that associated with the positive word, $hits("Negative")$ is the frequency of the negative word alone, $hits\ (Phrase\ Near\ "Negative")$ is the frequency of the given word that associated with the negative word, and $hit("Positive")$ is the frequency of the positive word alone. To prevent division by zero 0.01 was added to each hit.

In this research, the positive word and the negative words were identified based on our dataset as the positive word should reflect non-abusive word and negative word has to reflect abusive word. Therefore, we calculated the frequency of each word in the dataset. The word "الله" means God, and the word "سكس" means sex were the most frequent words in our dataset. However, the tweet PR approach, which was explained above, has the word "الله" mean God and word "سكس" as highest PR scores from the word graph of non-abusive accounts and word graph of abusive accounts, respectively.

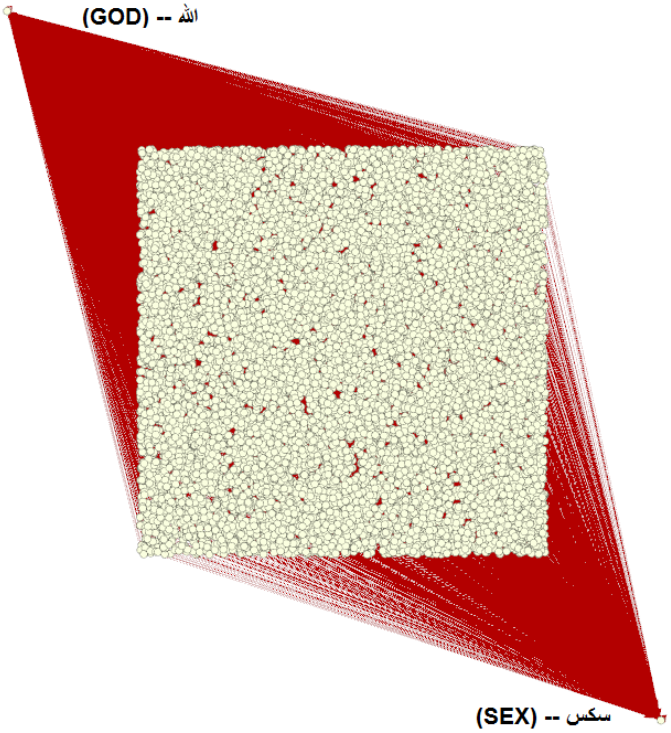

Figure 4. The Word Graph with the Positive and Negative edges

Furthermore, two Twitter accounts were picked to find which word belonged to which class. First, pornographic distribution account that had been followed by many adult accounts, was full of insult and obscene words, and has been suspended by Twitter, which represent the abusive account. The second account was for the professor engaged in social activities that ranked one of the top following accounts in Saudi Arabia, which would represent the non-abusive account. the most recent thousand tweets of each account were collected and counted the frequency of words. The word "الله" God was the most frequent word in the non-abusive account, and word "سكس" sex was the most frequent word in the abusive account. Therefore, word "الله" God used to reflect the positive word, and word "سكس" sex used to reflect the negative word.

To conduct the similarity measure of each word in the tweet, the training dataset with nearly 1,000,000 tweets was used to count the frequency of each word alone, and counted the frequency of each word that associated with word "الله" God or "سكس" sex at any location in the tweet.

For each tweet, the SO for each word was used to calculate the average of SO of the tweet. Lastly, the statistical summary measures were applied on all tweets from each user to reflect the behavior of each account as shown in Table 10.

Table 10. Tweet SO Feature Set

Features
SO_max
SO_min
SO_avg
SO_std

Social Graph Features

Social graph features are extracted from concepts of social graph theory (Zafarani, Abbasi, & Liu, 2014). The abusive accounts are engaging on Twitter by forming social activities of having a number followings and followers, whom they may not know. It can be useful to locate each account on the network and find if the social network of abusive accounts is reflecting a different behavior than the non-abusive accounts. Therefore, the most common centrality measures presented in Table 11 have been used to identify the Twitter accounts' social activities.

Eigenvector measures the user influence on the network (Easley, 2010). In-degree measures the number of connections directed to the user, while out-degree measures the number of connections directed from the user to other users. Degree measures the number of connection going in and out of the node. Betweenness computes the number of times the node was part of shortest path between other nodes (Bergamini & Meyerhenke, 2015). Closeness measures how much closer the node is to every other node. The higher closeness score indicates the node has many connections to other nodes (Zafarani et al., 2014).

Moreover, the shortness path for each Twitter account to positive and negative accounts were measured by sing the two accounts that have been used on the tweet SO method. The positive account reflects the non-abusive account and the negative account reflects the abusive account. Also, the shortest path of the account of being close to a positive or to a negative account has been calculated by the following formula:

Equation 7. Shortest Path

$$Short\ Path\ Compression = Shortness\ path\ to\ Positive - Shortness\ path\ to\ Negative$$

The compression will have a total greater than 0 if the shortness path is closer to positive account than negative account and would have total less than 0 if the shortness path is closer to a negative account than a positive account.

Table 11. Social Graph Features

Features
In_degree
Out_degree
Eigenvector
Betweenneess
Closeness
Short_path_pos
Short_path_neg
Short_path_total
Degree

Features Selection

The main purpose of the features selection is to decrease the size of the features that add noise to the classifier and lead to wrong classifications. Moreover, it minimizes the number of the features to the most relevant and reduces the calculation time to improve its accuracy.

Feature selection methods are categorized into three categories: filter, wrapper, and embedded methods. The Filtering method is a pre-processing approach that selects the features independent of the learning algorithm. It ranks each feature based on the information gain to the class without giving any attention to the dependent relationship between the features. The Wrapper method is based on the performance of a given learning machine with subset of features, where it explores the dependent relationships between features (Cateni, Vannucci, Vannocci, & Colla, 2012). The Embedded method is selecting features based on a specific learning machine.

In this research, the Wrapper method was used as it compares the performance of different subset of features, and choose features based on its performance with chosen

classifier. Wrapper has two greedy search strategies to choose the features. The first one is the Sequential Forward Selection (SFS) that starts with minimum number of a subset and adds to it (Guyon & Elisseeff, 2003). The second one is the Sequential Backward Selection (SBS), which starts with a full feature set and deletes the least important features one by one (Colla, 2012). I used the second strategy as it covers a larger set of features, where SFS could stop prematurely when it has a greater accuracy result.

Furthermore, I compared the feature set that reduced by the Wrapper method with the feature set reduced by the Filtering method to see if the selected features are cross checked on all proposed features sets. The compression was based on the size of the Wrapper method result, as the Filtering method can have any size of features.

Classifiers

Classifiers are data mining algorithms that classify the data into categories. The most comment three classifiers for text mining were used namely: Naïve Bayes, Support Vector Machine (SVM), and Decision Tree (J48).

Naïve Bayes (NB) is a simple probabilistic classifier based on Bayes theorem with the assumption that all attributes are strongly independent. Posterior probabilities are computed from prior probabilities, which are derived from previous experience (Irani et al., 2010).

Support Vector Machine (SVM) is a set of associated supervised learning methods that classify the data based on dimensional patterns (Tong & Koller, 2002).

Decision Tree (J48) is based on a predictive model, which maps the dataset into a tree structure that divides the data into subsets. The tree will contain decision nodes,

leaves, nodes, and branches. The decision nodes are the questioner nodes that feed the leaf nodes with the data subset (Friedl & Brodley, 1997).

Table 12. Confusion Matrix

Type	Prediction	
	Non-Abusive	**Abusive**
Non-Abusive	True Positive (TP_R)	False Negative (FN_R)
Abusive	False Positive (FP_R)	True Negative (TN_R)

The performance evaluate of each classifier is based on average precision (P), average recall (R), average F-measure (F), accuracy (A), and Receiver Operation Characteristic Curve (ROC). All four measures are computed from the confusion matrix as show in Table 12.

Where the confusion matrix in Table 12 represents the following:

- True Positive (TP_R): represents the number of non-abusive accounts correctly classified as non-abusive accounts.
- False Negative (FN_R): represents number of non-abusive account incorrectly classified as abusive account.
- True Negative (TN_R): represents number of abusive accounts correctly classified as abusive accounts.
- False Positive (FP_R): represents number of abusive accounts incorrectly classified as non-abusive accounts.

The precision (P), and recall (R) are measures of completeness and exactness respectively indicated in the formulas below.

Equation 8. Precision (P)

$$P = \frac{TP\_R}{(TP\_R + TN\_R)}$$

And

Equation 9. Recall (R)

$$R = \frac{TP\_R}{(TP\_R + FN\_R)}$$

F-measure (F) is the harmonic mean of the precision and recall values that are computed as:

Equation 10. F-measure (F)

$$F = \frac{2PR}{(P + R)}$$

Accuracy (A) is the correct result compared to all results that are computed as:

Equation 11. Accuracy (A)

$$A = \frac{TP\_R + TN\_R}{(TP\_R + FP\_R + TN\_R + FN\_R)}$$

Receiver Operator Characteristic Curve (ROC) is a graphical approach for displaying the tradeoff between a true positive rate (TPR) and false positive rate (FPR). Where TPR and FPR are computed as:

Equation 12. True Positive Rate (TPR)

$$\mathrm{TPR} = \frac{\mathrm{TP\_R}}{(\mathrm{TP\_R} + \mathrm{FN\_R})}$$

Equation 13. False Positive Rate (FPR)

$$\mathrm{FPR} = \frac{\mathrm{FP\_R}}{(\mathrm{FP\_R} + \mathrm{TN\_R})}$$

CHAPTER FOUR

FINDINGS AND ANALYSIS

In this chapter, I present my findings and analysis of the multidimensional analysis model for detecting the abusive accounts with Arabic tweets. Several experiments were conducted to construct the model, improve the model performance, and evaluate the model with unknown dataset. Each experiment was explained on the following sections and summarized as follow:

- Corrected misspellings using the proposed method presented on the methodology section.
- Evaluated classifier performances using 5-fold cross validation on the full features.
- Applied filtering method for feature selection to reduce the number of features and evaluate classifier performance of using minimal features.
- Used hypotheses testing to choose the classifier that showed better performance in detecting abusive accounts.
- Applied Wrapper method of feature selection with the chosen classifier to obtain best feature set that demonstrated better classifier performance.
- Compared the classifier performance with the same dimension features that were selected by two feature selection methods.

- Used the test set to evaluate the performance of the predicted model.

Dataset Pre-process and Misspell Correction

All the tweets used in this research were pre-processed to remove noise to improve classifier performance. The techniques I used to pre-process the tweets are:

- Removing all non-Arabic words.
- Removing all symbols.
- Removing all digits.
- removing all the stop words by using the stop word list in ("nltk.stem.isri — NLTK 3.0 documentation," n.d., "PyArabic 0.5 : Python Package Index," n.d., "stop-words - Stop words - Google Project Hosting," n.d.).
- removing all sequences of letters in the words except the name of God (Allah-الله) (E. A. Abozinadah et al., 2015).
- Removing all diacritics.
- Removing all extra whitespaces.

Furthermore, Correcting misspelled words generally improves text performance and the overall result (Bassil, 2012; Miangah, 2013; Nguyen, Nguyen, & Snasel, 2015; Sallam et al., 2016). Therefore, The misspelling words were corrected in the dataset using the proposed correction approach and compared it with other existing correction approaches to reflect the advantage of use this approach.

The normalized dataset was used to create eight data sets, which were used to compare the classifier performance of each set. The eight datasets are: clear tweets, basic normalization, edit distance, proposed approach, and four other datasets that are the same

previous four sets with the light stemmer applied. The light stemmer was used to remove prefixes and suffixes without approaching an infix or getting the root of the words. Arabic text mining has better performance with a light stemmer than a root stemmer (Saad & Ashour, 2010; Sallam et al., 2016). The two most common Arabic light stemmers are ISRI ("nltk.stem.isri — NLTK 3.0 documentation," n.d.; Taghva, Elkhoury, & Coombs, 2005) and Tashaphyne ("Tashaphyne 0.2 : Python Package Index," n.d.). In our work, the ISRI light stemmer was applied. The description of the first four sets is:

- Clear tweets dataset is the normalized dataset without applying any further process of correcting misspelled words.
- Basic normalization dataset is the normalized dataset with the basic Arabic normalization process to correct the most common Arabic misspellings that were presented on (Darwish, Magdy, & Mourad, 2012; Sallam et al., 2016):
  - Converting آ – إ – أ to ا
  - Converting ى to ي
  - Converting ة to ه
  - Converting ؤ – ئ to ء
- Edit Distance dataset is normalized by choosing the correct word by using an edit distance of 1. The edit distance of 1 can detect an error within a word which is caused by the following operations: deletion, substitution, indentation, and transposition (Levenshtein, 1966).
- Proposed Approach dataset uses normalized dataset in which we applied the proposed approach of correcting misspelled words as described above.

Each tweet in the dataset was tokenized into bag of words (BOW), which gave a set of features as shown in **Error! Reference source not found.**.

Table 13. Number of Features in the Eight Datasets

	Without stem	With stem
Clean dataset	6414	5682
Basic Dataset	6079	5618
Edit Distance dataset	6181	5744
Propose Approach dataset	6049	5600

The Performance Evaluate of Misspell Correction Method

The eight datasets were classified using The Support Vector Machine (SVM) algorithm with five-fold cross-validation to evaluate the performance of the eight models. The classifier predictive accuracy of each model was over 90%. However, the models without light stemming performed better than models with light stemming. Detailed results on performance of the models in discriminating between abusive and non-abusive Arabic tweets are shown in Table 14. The classifier with the dataset of proposed approach without light stemming got 1% better than the rest of the datasets. Also, the confusion matrices shown in Table 15 present eight false negatives, and nineteen false positives, which imply the use of word correction improved the abusive account detection performance.

Table 14. Classifier Performance of Eight Data-Sets

Data Sets	Accuracy	Precision	Recall	F-measure
Clean Data-set	95.7%	95.8%	95.7%	95.7%
Basic Normalized Data-set	96.0%	96.1%	96.0%	96.0%
Edit Distance Data-set	95.7%	95.7%	95.7%	95.7%
Bi-gram Approach Data-Set	96.5%	96.6%	96.5%	96.5%
Stem-Clean Data-set	96.3%	96.3%	96.3%	96.3%
Stem-Basic Normalized Data-set	96.0%	96.1%	96.0%	96.0%
Stem-Edit Distance Data-set	96.3%	96.3%	96.3%	96.3%
Stem-Propose Bi-gram Data-Set	96.3%	96.3%	96.3%	96.3%

Table 15. Misspell Correction Confusion Matrices

	Non-Abusive	Abusive
Non-Abusive	392	8
Abusive	19	381

Additionally, the basic normalization method corrects four common typing mistakes, but the proposed approach has the ability of correcting the most common words that appear in Twitter.

The stem datasets performed worse than the non-stemmed datasets, which reflected the needs for the full word length in Arabic text classification. The full word length reflected the gender, time, and population, which these parts would be lost when using the stemmed word. For example, an abusive account tweeting behavior is to talk about their interactions in the present tense, not about what they did, which would be missed by stemming. The approach used in my research outperformed other common approaches on Arabic word correction in Twitter (Al-Jefri & Mohammed, 2015; Muaidi & Al-tarawneh, 2012). The performance result illustrated the drawback of using stemming in Arabic language tweets. Therefore, I used this approach on correcting the

tweet and count of the misspelled words. Correcting the tweet also decreased the number of the nodes on the graph for the tweet PageRank method, and the semantic orientation (SO) method.

Feature Selection and Classifier Evaluation

All the feature sets were combined into one large set as show in Table 17 to evaluate the performance of using all features. A total of 104 features based on the following feature sets: profile feature set had 4 features, social graph feature set had 9 features, tweet content feature set had 78 features, tweet PR feature set had 9 features, and tweet SO feature set had 4 features.

The training models were evaluated using the 5-cross validation approach. The training dataset comprised of manually labeled tweets from 320 non-abusive accounts and 320 abusive accounts. The performance of the three classifiers in discriminating between abusive and non-abusive accounts was above 90%, which reflect the usefulness of the features in detecting the abusive accounts. However, some of the features that were included in the models, can reduce the performance of the classifiers.

Table 16. Evaluation Performance of Full Features Set

Classifier	Accuracy	Precision	Recall	F-Measure	ROC
NB	95.5%	95.6%	95.5%	95.5%	97.6%
SVM	96.3%	96.3%	96.3%	96.2%	96.3%
J48	93.9%	93.9%	93.9%	93.9%	94.9%

Table 17. Full Feature Sets Contain 104 Features

Feature Sets	Features	
Profile set	Num_tweet	Num_following

	Num_followers	Reputation
Social Graph set	In_degree	Short_path_pos
	Out_degree	Short_path_neg
	Eigenvector	Short_path_total
	Betweenneess	Degree
	Closeness	
Tweet Statistical set	Count_pic_max	Count_slang_max
	Count_pic_min	Count_slang_min
	Count_pic_avg	Count_slang_avg
	Count_pic_std	Count_slang_std
	Count_tweet_letters_max	Count_stand_max
	Count_tweet_letters_min	Count_stand_min
	Count_tweet_letters_avg	Count_stand_avg
	Count_tweet_letters_std	Count_stand_std
	Count_clean_letters_max	Count_unknown_max
	Count_clean_letters_min	Count_unknown_min
	Count_clean_letters_avg	Count_unknown_avg
	Count_clean_letters_std	Count_unknown_std
	Count_ar_max	Count_correct_slang_max
	Count_ar_min	Count_correct_slang_min
	Count_ar_avg	Count_correct_slang_avg
	Count_ar_std	Count_correct_slang_std
	Count_en_max	Count_correct_stand_max
	Count_en_min	Count_correct_stand_min
	Count_en_avg	Count_correct_stnad_avg
	Count_en_std	Count_correct_stand_std
	Count_hash_max	Count_sequence_max
	Count_hash_min	Count_sequence_min
	Count_hash_avg	Count_sequence_avg
	Count_hash_std	Count_sequence_std
	Count_http_max	Count_stop_max
	Count_http_min	Count_stop_min
	Count_http_avg	Count_stop_avg
	Count_http_std	Count_stop_std
	Count_mention_max	Count_correct_letters_max
	Count_mention_min	Count_correct_letters_min
	Count_mention_avg	Count_correct_letter_avg
	Count_mention_std	Count_correct_letters_std
	Count_non_max	Count_diacritics_words_max

	Count_non_min	Count_diacritics_words_min
	Count_non_avg	Count_diacritics_words_avg
	Count_non_std	Count_diacritics_words_std
	Count_number_max	Mention_rate
	Count_number_min	Hash_rate
	Count_number_avg	
	Count_number_std	
Tweet PR set	PR_n_avg_stop	PR_s_avg_stop
	PR_n_max_stop	PR_s_std_stop
	PR_n_min_stop	PR_s_max_stop
	PR_n_std_stop	PR_s_min_stop
		PR_avg_match_word
Tweet SO set	SO_max	SO_avg
	SO_min	SO_std
Total Features	**104**	

Therefore, the filtering method was used for feature selection that ranked the features based on the information gain. The features based on the information gain were ranked, which many features had information gain of less than 10% information gain. Therefore, a threshold of 10% information gain was used to select the features with equal or greater than 10% information gain. The result was 46 features as shown in Table 18. The features were across all the sets and divided as follow: 2 features associated with profile feature set, 6 features associated with social graph feature set, 27 features associated with tweet content feature set, 7 features associated with tweet PR feature set, and 4 features associated with tweet SO feature set.

Table 18. Feature Selection with Filtering Method Contain 46 Features

Feature Sets	**Features**	
Profile set	Num_following	Reputation

Social Graph set	In_degree	Short_path_neg
	Out_degree	Short_path_total
	Eigenvector	Closeness
Tweet Statistical set	Count_stop_avg	Count_pic_avg
	Count_ar_avg	Count_unknown_avg
	Count_stop_std	Count_correct_slang_avg
	Count_stand_avg	Count_unknown_max
	Count_ar_max	Count_correct_slang_std
	Count_stand_max	Count_mention_avg
	Count_diacritize_words_avg	Count_slang_std
	Count_stop_max	Count_hash_std
	Count_ar_std	Count_sequence_avg
	Count_diacritics_words_std	Count_hash_max
	Count_diacritics_words_max	Count_mention_max
	Count_stand_std	Count_ar_min
	Count_tweet_letters_avg	Count_correct_slang_max
	Count_unknown_std	
Tweet PR set	PR_avg_normal	PR_std_abusive
	PR_max_normal	PR_max_abusive
	PR_min_normal	avg_match_word
	PR_avg_abusive	
Tweet SO set	SO_avg	SO_max
	SO_std	SO_min
Total Features	**46**	

5-cross validation was used to evaluate classifiers' performance with the 46 selected features. As shown in Table 19. All the performance evaluations for the three classifiers were over 90% and had a similar evaluation result of using the full 104 features. The evaluation results showed the usefulness of using feature selection on reducing the number of the features as the result of the evaluation did not decreased. Additionally, the selected features are reflective of all the feature sets, which highlight the benefit of using the multidimensional analysis approach for Twitter accounts.

The performance of classifiers were evaluated using accuracy rate, precision, recall, f-measure and receiver operating curves. Table 15 shows evaluations of the classifiers for models trained on 46 features. The Support Vector Machine (SVM) performed better than Naïve Bayes (NB) and decision trees (J48) on four out of five measures. Moreover, the SVM has an accuracy rate of 1% higher than Naïve Base (NB) and 3% higher than J48; therefore, classifiers' performance can be sorted based on the better performance as follows: SVM, NB, and J48. Furthermore, to ensure that the SVM has the best performance I have used the hypotheses testing that is explained in the following section.

Table 19. Performance Evaluation of Selected Features of 46 Features

Classifier	Accuracy	Precision	Recall	F-Measure	ROC
NB	95.8%	95.9%	95.8%	95.8%	98.1%
SVM	96.7%	96.7%	96.7%	96.7%	96.7%
J48	93.8%	93.8%	93.8%	93.7%	95.0%

Hypotheses Testing

Several hypotheses testing were conducted to choose the best classifier out of the three classifiers. Two hypotheses testing were used. First, 5-cross validation that present above on Table 19. Second, t-testing to ensure of choosing the best classifier (Bouckaert, 2003). I used t-testing in four evaluation measures. These measures are the average accuracy, average precision, average ROC, and average F-measure. The result of the four-evaluation measure is presented as follows:

First, I used three null hypotheses testing as shown in Table 20 to compare the average accuracy of SVM, NB and J48 as follow:

Hypotheses testing in comparison to the SVM with NB:

- H_0: the average accuracy of SVM is equal or less than the average accuracy of NB.
- H_1: the average accuracy of SVM is greater than the average accuracy of NB.

Hypotheses testing in comparison to the SVM with J48:

- H_0: the average accuracy of SVM is equal or less than the average accuracy of J48.
- H_1: the average accuracy of SVM is greater than the average accuracy of J48.

Hypotheses testing in comparison to the J48 with NB:

- H_0: the average accuracy of J48 is equal or less than the average accuracy of NB.
- H_1: the average accuracy of J48 is greater than the average accuracy of NB.

Table 20. Average Accuracy Hypotheses Testing

Null Hypotheses(H_0)	Alternative Hypotheses(H_1)	Average Accuracy	Hypotheses (α=0.5)
SVM =< NB	SVM > NB	0.96 ±0.04	Reject
SVM =< J48	SVM > J48	0.96 ±0.03	Reject
J48 =< NB	J48 > NB	0.96 ±0.04	Reject

The result of hypotheses testing of average accuracy that is presented in Table 20 show that all the null hypotheses H_0 have been rejected, which reflect an acceptance for all the

alternative hypotheses H_1. Though, SVM had better performance than NB and J48, and J48 has better performance than NB.

Second, I used three null hypotheses testing as shown in Table 21 to compare the average precision of SVM, NB and J48 as follow:

Hypotheses testing of comparing the SVM with NB:

- H_0: the average precision of SVM is equal or less than the average precision of NB.
- H_1: the average precision of SVM is greater than the average precision of NB.

Hypotheses testing of comparing the SVM with J48:

- H_0: the average precision of SVM is equal or less than the average precision of J48.
- H_1: the average precision of SVM is greater than the average precision of J48.

Hypotheses testing of comparing the J48 with NB:

- H_0: the average precision of J48 is equal or less than the average precision of NB.
- H_1: the average precision of J48 is greater than the average precision of NB.

Table 21. Average Precision Hypotheses Testing

Null Hypotheses(H_0)	Alternative Hypotheses(H_1)	Average Precision	Hypotheses (α=0.5)
SVM =< NB	SVM > NB	0.96 ±0.03	Reject
SVM =< J48	SVM > J48	0.97 ±0.03	Reject
J48 =< NB	J48 > NB	0.96 ±0.03	Reject

The result of hypotheses testing of average precision is presented in Table 21 show that all the null hypotheses H_0 have been rejected, which reflect an acceptance for all the alternative hypotheses H_1. Though, SVM had better performance than NB and J48, and J48 had better performance than NB.

Third, I used three null hypotheses testing as shown in Table 22 to compare the average ROC of SVM, NB and J48 as follow:

Hypotheses testing of comparing the SVM with NB:

- H_0: the average ROC of SVM is equal or less than the average ROC of NB.
- H_1: the average ROC of SVM is greater than the average ROC of NB.

Hypotheses testing of comparing the SVM with J48:

- H_0: the average ROC of SVM is equal or less than the average ROC of J48.
- H_1: the average ROC of SVM is greater than the average ROC of J48.

Hypotheses testing of comparing the J48 with NB:

- H_0: the average ROC of J48 is equal or less than the average ROC of NB.
- H_1: the average ROC of J48 is greater than the average ROC of NB.

Table 22. Average ROC Hypotheses Testing

Null Hypotheses(H_0)	Alternative Hypotheses(H_1)	Average ROC	Hypotheses (α=0.5)
SVM =< NB	SVM > NB	0.98 ±0.02	Fail to Reject
SVM =< J48	SVM > J48	0.96 ±0.05	Reject
J48 =< NB	J48 > NB	0.98 ±0.02	Fail to Reject

The results of the hypotheses testing of average ROC that are presented in Table 22 show that one null hypothesis H_0 have been rejected, and two have been failed to reject. Though NB had better performance than SVM and J48.

Fourth, I used three null hypotheses testing as shown in Table 23 to compare the average F-measure of SVM, NB and J48 as follow:

Hypotheses testing of comparing the SVM with NB:

- H_0: the average F-measure of SVM is equal or less than the average F-measure of NB.
- H_1: the average F-measure of SVM is greater than the average F-measure of NB.

Hypotheses testing of comparing the SVM with J48:

- H_0: the average F-measure of SVM is equal or less than the average F-measure of J48.
- H_1: the average F-measure of SVM is greater than the average F-measure of J48.

Hypotheses testing of comparing the J48 with NB:

- H_0: the average F-measure of J48 is equal or less than the average F-measure of NB.
- H_1: the average F-measure of J48 is greater than the average F-measure of NB.

Table 23. Average F-measure Hypotheses Testing

Null Hypotheses(H_0)	Alternative Hypotheses(H_1)	Average F-measure	Hypotheses (α=0.5)

SVM =< NB	SVM > NB	0.96 ±0.04	Reject
SVM =< J48	SVM > J48	0.96 ±0.03	Reject
J48 =< NB	J48 > NB	0.96 ±0.04	Reject

The result of hypotheses testing of average F-measure that is presented in Table 23 show that all the null hypotheses H_0 have been rejected, which reflect an acceptance for all the alternative hypotheses H_1. Though, SVM had better performance than NB and J48, and J48 had better performance than NB.

From the four hypotheses test sets the SVM outperform J48 and NB into three evaluation measures: average accuracy, average precision, and average F-measure. The NB outperformed the J48 and SVM into the average ROC. In the end, SVM had a better evaluation on three out of four measures.

Therefore, I concluded the SVM was the appropriate classifier for the abusive accounts detection model as it had the highest evaluation measures of using t-test and the best evaluation performance by using 5-cross validation.

Feature Selection

Moreover, the research method was based on an analysis process of covering the user profile, social network, and tweets to build a multidimensional analysis model that detected multiple aspects of the abusive accounts' behavior. I compared the evaluation performance of three different grouping of feature sets to ensure the usefulness of each feature set. The evaluation result is presented in Table 24. The comparison includes: All five feature sets together, four feature sets together and ignored one set, and one feature set alone. The selected model of five feature sets performed better than the models with

four feature sets, or models with one feature set, which the accuracy rate achieved 96%, 95% and 90% respectively. In comparison, the models of one feature set obtained accuracy rates from 72% to 93%, which reflected the variable discriminative power of each feature set. However, the feature sets based on a tweet that included content set, PR set, and SO set reflected the richness of data in the tweet that lead to detecting abusive accounts. Much prior research had been conducted on Twitter without analyzing the tweet content, as it is complex to normalize and has limited lexicon resources on some languages.

Moreover, I used a wrapper feature selection method that compared the performance of subset of features to find the best feature set. Also, it selects the features based on their performance with the classifier. The wrapper method was run using SVM and backward feature selection. The backward feature selection works by using all features and then randomly drops one feature at time and stops dropping when the classifier performance decreases.

Table 24. Comparing Verity of Feature Sets Combination

Feature Set	# Features	Accuracy	Precision	Recall	F-Measure	ROC
All Feature sets	46	96.7%	96.7%	96.7%	96.7%	96.7%
All Feature sets - Profile set	44	95.3%	95.3%	95.3%	95.3%	95.3%
All Feature sets - Social Graph set	40	94.8%	94.8%	94.8%	94.8%	94.8%
All Feature sets - Statistical set	19	95.5%	95.6%	95.5%	95.5%	95.5%
All Feature sets - PR set	39	95.2%	95.2%	95.2%	95.2%	95.2%

All Feature sets - SO set	42	95.3%	95.3%	95.3%	95.3%	95.3%
Profile set	2	71.6%	72.1%	71.6%	71.4%	71.6%
Social Graph set	6	76.1%	76.6%	76.1%	76.0%	76.1%
Statistical set	27	93.3%	93.3%	93.3%	93.3%	93.3%
PR set	7	88.6%	89.1%	88.6%	88.6%	88.6%
SO set	4	92.2%	92.6%	92.2%	92.2%	92.2%

The selected features of the wrapper method were 31 features that are presented in Table 25. The selected features reflect all the feature sets that are divided as follows: 2 features associated with profile feature set, 3 features associated with social graph set, 17 features associated with tweet content feature set, 5 features associated with tweet PR feature set, and 4 features associated with tweet SO feature set.

Table 25. Wrapper Method with 31 Features

Feature Sets	**Features**	
Profile Feature set	Num_following	Reputation
Social Graph set	In_degree	Eigenvector
	Short_path_neg	
Tweet Statistical set	Count_stand_avg	Count_mention_max
	Count_slang_std	Count_mention_avg
	Count_ar_max	Count_stop_max
	Count_ar_std	Count_sequence_avg
	Count_ar_min	Count_hash_max
	Count_diacritics_words_std	Count_correct_slang_std
	Count_unknown_std	Count_correct_slang_avg
	Count_unknown_avg	Count_correct_slang_max
	Count_pic_avg	

Tweet PR set	PR_avg_normal	PR_std_abusive
	PR_min_normal	PR_max_abusive
	PR_avg_abusive	
Tweet SO set	SO _avg	SO _max
	SO _std	SO _min
Total Features	**31**	

Feature Selection Methods' Comparisons

In this section, I used the filtering method of feature selection to select 31 features to perform several comparisons. First, I checked if the selected features were associated with all feature sets as it was in the wrapper method. Second, I compared the classifier performance of the two feature selection methods. Third, I determined the matching features between the two feature selection methods, and how the matching features associated with all feature sets. Fourth, I compared the classifier performance with and without feature selection. The aim of this comparisons was to ensure the usefulness of every feature set and the different result of using different feature selection methods.

The first comparison is shown in Table 26. The selected features are associated with all feature sets as follows: 1 feature associated with profile feature set, 3 features associated with social graph feature set, 18 features associated with tweet statistical feature set, 5 features associated with tweet PR feature set, and 4 features associated with tweet SO feature set.

Table 26. Feature Selection - Filtering Method - contain 31 Features

Features Set	**Features**	
Profile set	Reputation	

Social Graph set	In_degree	Out_degree
	Eigenvector	
Tweet Statistical set	Count_stop_avg	Count_ar_max
	Count_ar_avg	Count_stand_max
	Count_stop_std	Count_diacritize_words_avg
	Count_stand_avg	Count_stop_max
	Count_tweet_letters_avg	Count_ar_std
	Count_unknown_std	Count_diacritics_words_std
	Count_unknown_avg	Count_diacritics_words_max
	Count_correct_slang_avg	Count_stand_std
	Count_unknown_max	Count_pic_avg
Tweet PR set	PR_avg_abusive	PR_max_abusive
	PR_std_abusive	PR_max_normal
	avg_match_word	
Tweet SO set	SO _avg	SO_max
	SO _min	SO_std
Total Features	**31**	

For the second comparison, I evaluated the SVM performance of features selected by wrapper method and filtering method. One model contained 31 features selected by wrapper method, and the second model contained the top 31 features of filtering method as shown in Table 25 and Table 26 respectively. The evaluation results shown in Figure 5 represent 96% accuracy rate and higher for both methods, which reflect the usefulness of both models. However, the accuracy performance of wrapper method is approximately 2% better than the filtering method, which reflects the selected features by wrapper has a higher impact on SVM than filtering method. Moreover, both methods have features that across all the five feature sets, which illustrate the benefit of using the multidimensional analysis approach on detecting the abusive accounts.

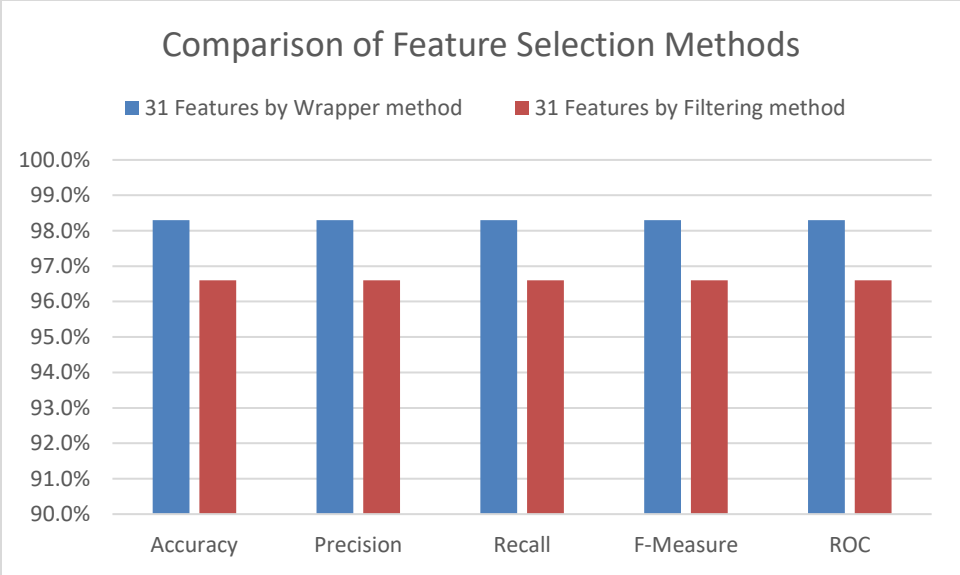

Figure 5. Comparison of Feature Selection Methods

In the third comparison, the features on both models were compared as the features in each model are slightly different. As shown in Table 27, each method has some similar features and some features are unique. Therefore, each feature was returned to the original set to identify which set had the better impact on the feature selection method. Also, the reasons for having higher performance of the model built with wrapper method was identified. Each feature set is presented below:

- Profile feature set size: Wrapper method > Filtering method.
- Social graph feature set size: Wrapper method = Filtering method.
- Tweet statistical feature set size: Wrapper method < Filtering method.
- Tweet PR feature set size: Wrapper method = Filtering method.
- Tweet SO feature set size: Wrapper method = Filtering method.

The comparison of both methods had almost similar distribution features over the feature sets. However, the slight improvement of the wrapper method reflects the usefulness of using more features from the profile feature set, and less features from tweet statistical feature set.

Table 27. Matching Features of Filtering Method and Wrapper Filtering Method

Feature Sets	Wrapper Method	Filtering Method	Match
Profile set	Reputation	Reputation	X
	Num_following		
Social Graph set	In_degree	In_degree	X
		Out_degree	
	Eigenvector	Eigenvector	X
	Short_path_neg		
Tweet Statistical set		Count_stop_avg	
	Count_stop_max	Count_stop_max	X
		Count_stop_std	
		Count_tweet_letters_avg	
		Count_unknown_max	
	Count_unknown_std	Count_unknown_std	X
	Count_unknown_avg	Count_unknown_avg	X
	Count_correct _slang_avg	Count_correct _slang_avg	X
	Count_ar_min		
		Count_ar_avg	
	Count_ar_max	Count_ar_max	X
	Count_ar_std	Count_ar_std	X
		Count_stand_max	
		Count_stand_std	
	Count_stand_avg	Count_stand_avg	X
	Count_slang_std		
	Count_correct _slang_std		
	Count_correct _slang_max		
		Count_diacritics	

		_words_avg	
	Count_diacritics _words_std	Count_diacritics _words_std	X
		Count_diacritics _words_max	
	Count_pic_avg	Count_pic_avg	X
	Count_sequence_avg		
	Count_hash_max		
	Count_mention_max		
	Count_mention_avg		
Tweet PR set	PR_avg_abusive	PR_avg_abusive	X
	PR_std_abusive	PR_std_abusive	X
	PR_max_abusive	PR_max_abusive	X
		PR_max_normal	
	PR_avg_normal		
	PR_min_normal		
		PR_avg_match_word	
Tweet SO set	SO_avg	SO_avg	X
	SO_min	SO_min	X
	SO_max	SO_max	X
	SO_std	SO_std	X
Total Features	**31**	**31**	**19**

In the fourth comparison, the SVM evaluation performances were compared with and without feature selection. As shown in Figure 6, the full feature are 104 features without using feature selection, and the second and third sets are selected features by using filtering method to reduce the features to 46 features and 31 features respectively. The classifier performance reached 96% rate in all evaluation measures in all the three-feature sets, which reflected no reduction in the classifier performance by using less features. The feature reduction reflects the benefit of the feature selection method in reducing the number of useless features that has no positive impact on the performance of the classifier.

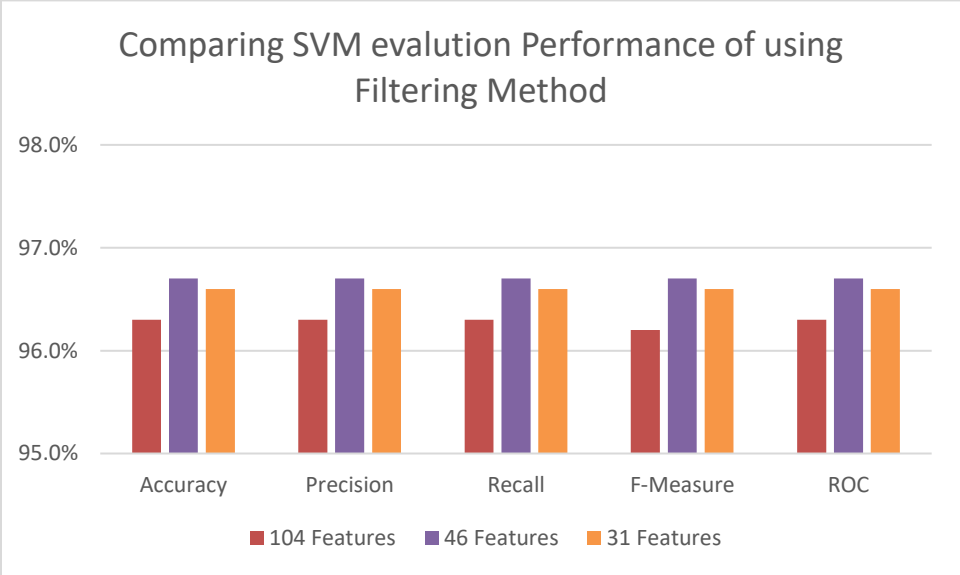

Figure 6. SVM Performance Evaluation and Comparison of using Filtering Method.

Lastly, the performances of using and not using the wrapper method were compared. As shown in Figure 7 the 31 features that selected with wrapper method had the highest evaluation result of 98% in all measures and it out perform the full features of 104 features and 46 selected features.

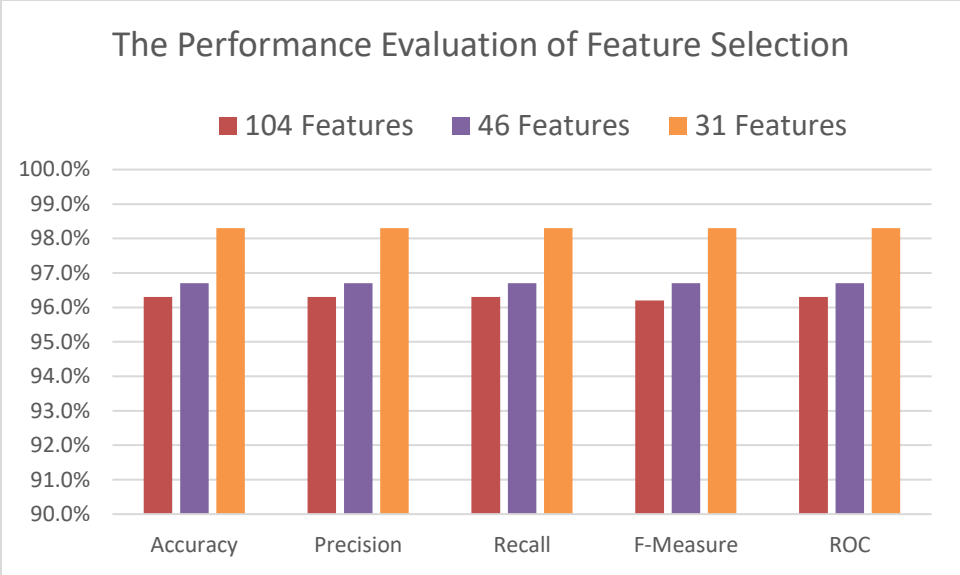

Figure 7. SVM Performance Evaluation and Comparison of using Wrapper Method.

In conclusion, the four comparisons of using feature selection methods illustrated a better performance evaluation by using the model with 31 features selected by wrapper method. Also, the selected features were distributed over all the feature sets, which reflected the usefulness of using the multidimensional features on detecting the abusive accounts. In addition, the classifier evaluation measures based on 5-cross validation of training dataset exceeded a 96% rate with all the models that were presented above. The model generated by wrapper method reached a 98% accuracy rate, which is better than the model generated by filtering method for detecting the abusive accounts. Therefore, the constructed model based on the wrapper method is more suitable for detecting abusive accounts.

Using Test Set to Evaluate Conducted Models

The proposed methodology produced a model with 31 features of multidimensional analysis sets. This model had been evaluated by using the training

dataset with 5-cross validation to prepare it for use with unseen datasets. The testing set that has been explained above is unseen datasets and has not been used through all above experiments. It contained 80 non-abusive accounts and 80 abusive accounts that tested the learned models. The classifier performance evaluation of constructed model is shown in Table 28. The Evaluation reveals an average accuracy rate of 90%, an average precision rate of 91%, an average recall rate of 90%, an average F-measure rate of 90%, and an average ROC rate of 90% of detecting abusive accounts. The model's evaluation shows the benefit of using a multidimensional analysis approach on analyzing the Twitter accounts.

Table 28. Performance Evaluation with 20% Testing Set

SVM	Accuracy	Precision	Recall	F-Measure	ROC
Wrapper method with 31 Features	90.6%	91.7%	90.6%	90.6%	90.6%

Table 29. The Confusion Matrix and Performance Evaluation per Class

Class	TP_R	FP_R	Precision	Recall	F-Measure	ROC
Non-Abusive	98.8%	17.5%	84.9%	98.8%	91.3%	90.6%
Abusive	82.5%	1.3%	98.5%	82.5%	89.8%	90.6%
Average	90.6%	9.4%	91.7%	90.6%	90.6%	90.6%

Moreover, the confusion matrix of the test set as shown in Table 29 present 98.8% of non-abusive accounts, and 82.5% of abusive accounts were identified correctly. However, 17.5% of abusive accounts were misclassified as they were using less hashtags, less pictures, less mentions, less spelling mistakes and more Arabic words. Therefore, the

behavior of each class was analyzed as shown in **Error! Reference source not found.**. This analysis represents the average class behavior of each feature. Under Figure 8, a comparison of features' usage between the abusive and non-abusive class.

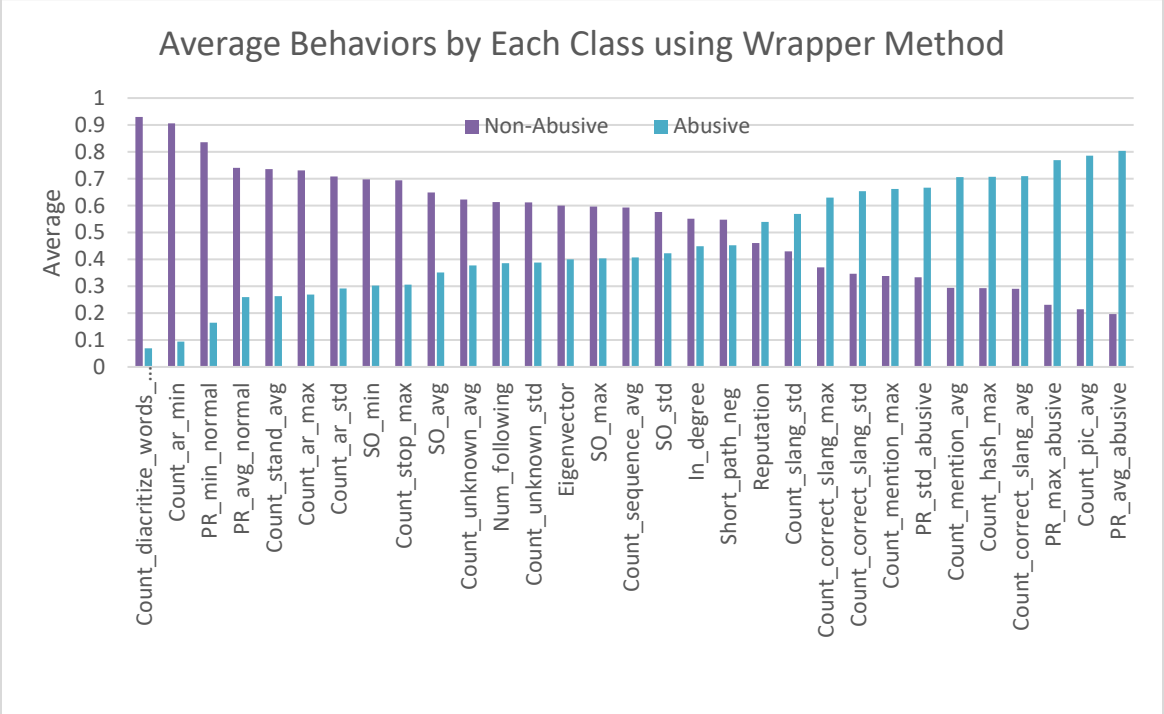

Figure 8. Comparison of the Average Behaviors per Class with Wrapper Method

- Profile_Num_following: None-Abusive accounts > Abusive accounts
- Profile_Reputation: None-Abusive accounts < Abusive accounts
- Social_Eigenvector: None-Abusive accounts < Abusive accounts
- Social_In_degree: None-Abusive accounts > Abusive accounts
- Social_Short_path_neg: None-Abusive accounts > Abusive accounts
- Tweet_Count_correct_slang_max: None-Abusive accounts < Abusive accounts

- Tweet_Count_slang_std: None-Abusive accounts > Abusive accounts
- Tweet_Count_ar_min: None-Abusive accounts > Abusive accounts
- Tweet_Count_stand_avg: None-Abusive accounts > Abusive accounts
- Tweet_Count_mention_max: None-Abusive accounts < Abusive accounts
- Tweet_Count_hash_max: None-Abusive accounts < Abusive accounts
- Tweet_Count_correct_slang_std: None-Abusive accounts < Abusive accounts
- Tweet_Count_pic_avg: None-Abusive accounts < Abusive accounts
- Tweet_Count_ar_max: None-Abusive accounts > Abusive accounts
- Tweet_Count_mention_avg: None-Abusive accounts < Abusive accounts
- Tweet_Count_stop_max: None-Abusive accounts > Abusive accounts
- Tweet_Count_correct_slang_avg: None-Abusive accounts < Abusive accounts
- Tweet_Count_ar_std: None-Abusive accounts > Abusive accounts
- Tweet_Count_diacritics_std: None-Abusive accounts > Abusive accounts
- Tweet_Count_sequence_avg: None-Abusive accounts > Abusive accounts
- Tweet_Count_unknown_avg: None-Abusive accounts > Abusive accounts
- Tweet_Count_unknown_std: None-Abusive accounts > Abusive accounts
- PR_min_normal: None-Abusive accounts > Abusive accounts
- PR_avg_normal: None-Abusive accounts > Abusive accounts
- PR_std_abusive: None-Abusive accounts < Abusive accounts
- PR_avg_abusive: None-Abusive accounts < Abusive accounts
- PR_max_abusive: None-Abusive accounts < Abusive accounts

- SO_avg: None-Abusive accounts < Abusive accounts
- SO_std: None-Abusive accounts > Abusive accounts
- SO_max: None-Abusive accounts > Abusive accounts
- SO_min: None-Abusive accounts > Abusive accounts

This comparison revealed unique characteristics of the abusive and non-abusive accounts. Also, identified the usage behavior of each feature.

Table 30. The Features' Usage per Class

Feature Sets	Features	Average of Non-Abusive	Average of Abusive
Profile Feature set	Num_following	824.088	518.61
	Reputation	0.672	0.786
Social Graph set	In_degree	0.0000898	0.000073
	Short_path_neg	2.847	2.353
	Eigenvector	0.003	0.002
Tweet Statistical set	Count_stand_avg	7.616	2.729
	Count_slang_std	0.992	1.313
	Count_ar_max	55.794	20.547
	Count_ar_std	11.355	4.68
	Count_ar_min	1.5	0.156
	Count_diacritics _words_std	4.149	0.311
	Count_unknown_std	1.601	1.017
	Count_unknown_avg	1.413	0.856
	Count_pic_avg	0.12	0.439
	Count_mention_max	1.391	2.728
	Count_mention_avg	0.22	0.527
	Count_stop_max	7.559	3.331
	Count_sequence_avg	0.704	0.484
	Count_hash_max	2.037	4.928
	Count_correct _slang_std	0.187	0.352
	Count_correct	0.064	0.156

	_slang_avg		
	Count_correct _slang_max	0.928	1.575
Tweet PR set	PR_avg_normal	0.001	0.00035
	PR_min_normal	0.000029	0.0000057
	PR_avg_abusive	0.00049	0.002
	PR_std_abusive	0.001	0.002
	PR_max_abusive	0.003	0.01
Tweet SO set	SO_avg	1.743	0.943
	SO_std	0.496	0.364
	SO_max	3.06	2.074
	SO_min	0.909	0.394
Total Features	**31**		

The behavior of non-abusive accounts and abusive accounts can be identified based on the average features' usage per class as represented in Table 30.

Average Usage of Feature Sets

The reflected behavior of each feature set is presented on following sections:

Profile Feature Set

Contains two features that reflecting the following:

- The feature of counting the number of the following (Num_following) shows the non-abusive accounts had a higher number of following than abusive accounts, which presents a normal behavior of a legitimate user of following large number of users to get news, knowledge, or make friends.
- One Twitter policy is that users should have more followers than followings, or the account would be flagged as spam. the reputation feature shows that both non-abusive accounts and abusive accounts were maintaining of having over a

65% reputation rate, which reflects that the active accounts do have more than 15% number of followers than number of following.

- The abusive accounts are having a lower number of followings but higher number of reputation, which indicates abusive accounts have more followers compared to the number of followings.

Social Graph Feature Set

Contains three features reflecting the following:

- The non-abusive accounts had more In-degree centrality measure (In_degree) than the abusive accounts, which indicate that the non-abusive accounts have more influence on the social network more than the abusive accounts.
- The short path to negative account feature (Short_path_neg) show the abusive accounts are closer to one of well-known pornographic distributor accounts in Arabic tweet than the non-abusive accounts.
- The non-abusive accounts have higher Eigenvector than abusive accounts, which reflect the non-abusive do form a social graph with mutual connection.

Tweet Statistical Feature Set

Contains 17 features that reflecting the following:

- The feature of counting the average number of standard Arabic word in the tweet (Count_stand_avg) shows that the non-abusive accounts did use the standard Arabic words two times more the abusive accounts.

- The feature of counting the standard division of slang (Count_slang_std) shows the non-abusive accounts were rarely using the slang words, while the abusive accounts were commonly using the slang words.
- The feature of counting the maximum number of Arabic words in the tweet (Count_ar_max) shows that the non-abusive accounts used 55 words in one of their tweets, while the abusive accounts used either less Arabic words or using other languages.
- The feature of the standard division of counting the number of Arabic words in the tweet shows that the non-abusive accounts were using 11 Arabic words' deviation from the average Arabic words' count in the tweet, while the abusive accounts were using 4 Arabic words' deviation from the average Arabic words in the tweet. Clearly, the abusive accounts do use a limited size of Arabic words in their tweet.
- Counting the minimum Arabic words in the tweet (Count_ar_min) shows that the abusive accounts were using no Arabic words or digits in some of their tweets, while the non-abusive accounts were using at least one Arabic word in their tweet.
- Counting the Arabic diacritics (Count_diacritics_words_std) shows that the non-abusive accounts do use the Arabic diacritics much more than the abusive accounts. Diacritics are considered part of the formal Arabic words, which commonly used in official letters, newspapers, or government websites.

Therefore, the abusive accounts are not wasting time to have tweets that are grammatically correct, which is considered part of their behavior.

- Counting the unknown Arabic words in the tweet (Count_unknown_std, Count_unknown_avg) shows that the non-abusive accounts use more unknown words that are not in standard Arabic or slang dictionaries. These words are the dialect words on some Arab regions that are not well known in other regions.
- Counting the average number of pictures for each Twitter account (Count_pic_avg) shows that the abusive accounts were using pictures in almost half of their tweets, which was much more than the number of pictures that post by non-abusive account, which counts as one tenth of their tweets.
- Counting the mentions in each tweet (Count_mention_max, Count_mention_avg) shows that the abusive accounts had almost three mentions in one of their tweets and at least one mention in half of their tweets. However, the non-abusive accounts were using at least one mention in one of their tweets, and almost 20% of their tweets had one mention.
- Counting the Arabic stop word in each tweet (Count_stop_max) shows the non-abusive accounts were using Arabic stop words two times more than the abusive accounts in one of their tweets. The larger number of stop words present a a correctly syntactic tweet that contains verbs and nouns that connect by stop words. The use of Arabic stop words by non-abusive accounts do match with the previous point of finding the non-abusive accounts do use more Arabic words than abusive accounts in their tweet.

- Counting the sequence of the same Arabic letters in words (Count_sequence_avg) shows that the non-abusive accounts were using this technique in their tweet more than the abusive accounts. This technique does present happiness, anger, importance, or inspiration word in the tweet, as no capitalization in Arabic language. Therefore, obvious use of this technique in non-abusive accounts as they are explaining or displaying their emotion. Commonly, abusive account users try to bypass the internet censorship or Twitter filtering by using this technique to post inappropriate words. Twitter has a restrictive policy prohibiting the profane text to be posted on the tweet, but the abusive accounts ignore this policy.
- Counting the hashtags in the tweet (Count_hash_max) shows that the non-abusive accounts had a maximum of two hashtags in one of their tweets, while the abusive accounts used approximately five hashtags in one of their tweets. The maximum number of hashtags shows that the abusive accounts do use more hashtags. The benefit of using many hashtags is to tag their tweet under many topic or trending topics to attract many users to their tweet and have their tweets viewed frequently.
- Counting the corrected misspelled slang words in the tweet (Count_correct_slang_std, Count_correct_slang_avg, Count_correct_slang_max) shows that the abusive accounts used large number of misspelled slangs in their tweet. The misspelled slangs would help the abusive accounts to post words excluded from blacklisted words in some Arab

countries. Therefore, correcting these words would identify the slang word sand can compare it against blacklists or inappropriate words. Additionally, correcting the slang word was based on the proposed correction approach, which reflects the abusive accounts behavior of using slang words. Therefore, correcting the misspelled slang incorrectly would ruin the tweet meaning, which is one of the abusive accounts behavior.

Tweet PageRank Features Set

Contains 5 features that reflect the following:

- The tweet PageRank of non-abusive accounts (PR_non-abusive_avg, PR_non-abusive_min) shows that the abusive accounts had very low tweet PageRank when it evaluated on the non-abusive word graph. Moreover, the non-abusive accounts had three times higher tweet PageRank score from non-abusive graph as the non-abusive accounts do use similar words and co-occurrence relationship between the words. In addition, the low tweet PR of abusive accounts represents fewer words that were commonly used between the abusive accounts and non-abusive accounts.
- The non-abusive graph has word (الله) God as the highest PR word as shown in Figure 9 and an abusive graph has the word (سكس) sex as the highest PR word as indicated in Figure 10. These high PR words in each graph are meaningful as the non-abusive accounts use some daily words possibly related to religion but the abusive accounts use adult words that more related to their sexual interest.

- The second highest PR word on abusive account graph is (ريتويت) “re-tweet” word as shown in Figure 10, which indicates the abusive accounts want their account to get popular by requesting the user to re-tweet their posted tweet. However, this word has low PR on the non-abusive graph as the non-abusive accounts are not eager on having their tweet re-tweeted unless the other accounts like it and feel it is worth being re-tweeted.

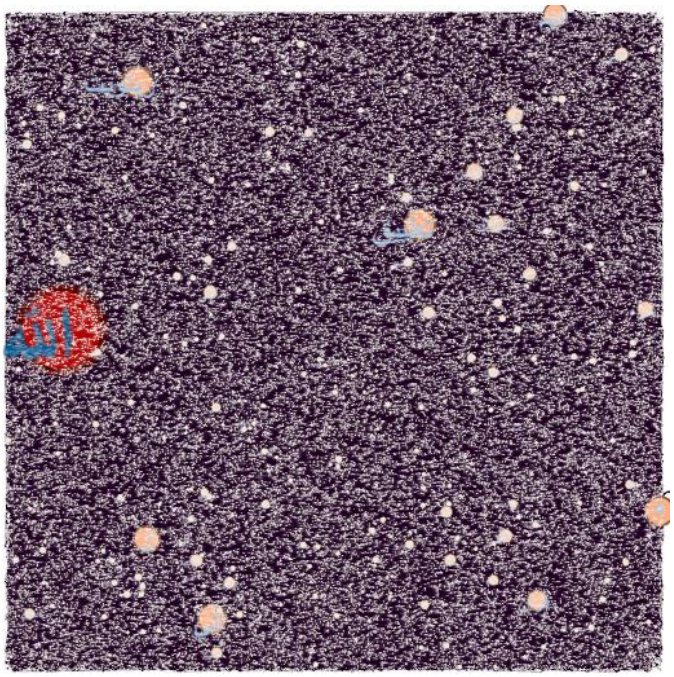

Figure 9. Word Graph of Non-Abusive Accounts

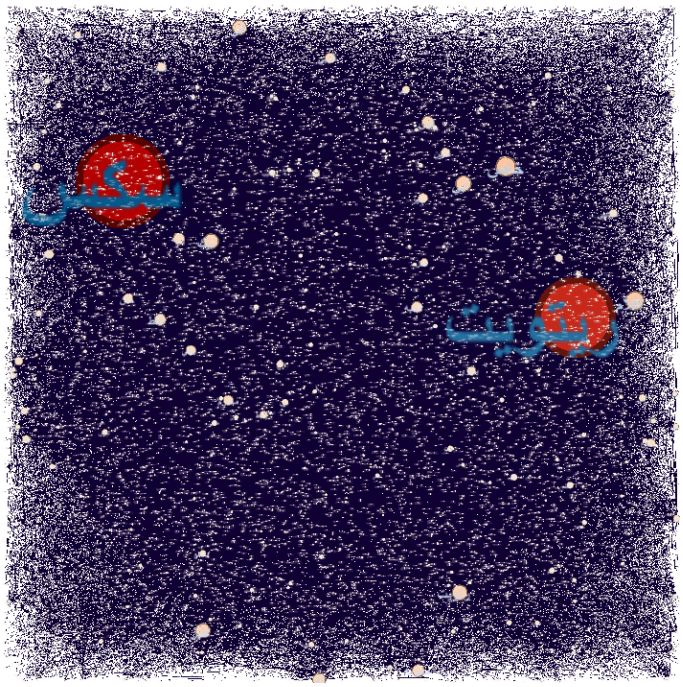

Figure 10. Word Graph of Abusive Accounts

- The tweet PageRank of abusive accounts (PR_abusive_avg, PR_abusive_max, PR_abusive_std) shows that the abusive accounts had a higher tweet PR from the abusive graph than non-abusive accounts as the non-abusive accounts have few matching words with abusive accounts.

Tweet Semantic Orientation Feature Set

Contains 4 features that reflects the following:

- The tweet semantic orientation (SO_avg, SO_std, SO_max, SO_min) shows the non-abusive accounts do use words that are relatively closer to the positive word than the non-abusive accounts. Moreover, the maximum SO of the abusive accounts were less than the non-abusive accounts as the abusive accounts using words that far away from the most positive word. On the

opposite side of the SO measure, the SO minimum score for the non-abusive accounts is larger than the abusive accounts. The SO minimum shows that the non-abusive accounts do use more words closer to positive words.

CHAPTER FIVE

DISCUSSION AND CONCLUSION

Cybercriminals continue to distribute abusive content by exploiting vulnerabilities and weaknesses in social networks and attempted controls. Some of this abusive content includes obscene and profane words that are prohibited and considered a crime in some countries ("Saudi govt. agencies struggling to fight porn on social media," n.d.). This research has covered a reasonable sample size of these accounts and generated a model that detected these accounts with an accuracy rate of 90%.

In this research, tweets were analyzed using independent lexical and statistical analysis. Independent lexical analysis overcome the limitation of dependent lexicon analysis tools, such as Natural Language Processing (NLP), Name Entity(NE), and Lexicon ontology and can be useful with other languages too. Some of statistical analysis of a tweet can be used with other languages as it's based on counting the elements on the tweet. Some unique features that can be use with other languages are the full features of tweet PR feature set, the full features of tweet SO feature set, and some of tweet statistical feature set that include count of mentions, count stop-words, and count hashtags.

The dimension of the initial features was 104 features as shown in Table 17. Some of these features were helpful in detecting abusive accounts and other were not. For

instance, the number of tweets and number of letters on each tweet was calculated, but both features were useless in identifying the abusive accounts. Feature selection methods were used to determine relevant features from the five-feature sets. The dimension of the features was reduced to 46 features. However, all the selected features covered the five feature sets. Having all the feature sets in the model reflects the distinct behavior of the abusive accounts across each set.

Furthermore, the wrapper method reduced the features to 31 features as shown in Table 25, which the reduced features were adding noise to the model. Moreover, the comparison of the two feature selection methods confirmed the usability of the five feature sets as all the features are across the five sets.

Contributions

This research covered six contributions that were part of the construction process of proposed model. First, we collected a dataset that was manually analyzed to build a ground truth for this research, as no dataset has been existed for abusive accounts with Arabic tweets. This dataset was collected by using customized Python code to overcome the limitations of API Twitter. Moreover, when this research started in 2014, all the pictures in tweets were viewable without any warning of nudity content, but on 2015 Twitter started using age restrictions and warning massages for any nude pictures. Therefore, Twitter solved part of the picture issues but not the abusive content issues, which is the focus of this research.

Second, the preliminary result was conducted by analyzing a set of Twitter accounts by using Bag of Word (BOW) approach. The approach had five hundred

features from the tweet alone (E. A. Abozinadah et al., 2015). As result, we observed the importance of using the tweet content to detect the abusive accounts. Additionally, the misspelled words were widespread in Twitter, and how it affected the classifier performance.

Third, the proposed misspelled correction method was capable of correcting misspelled words that do not exist in Arabic dictionaries (E. Abozinadah & Jones, 2016). However, using existing misspelled correction based on minimum string matching technique does replace the misspell word with the similar word from the dictionary, but the corrected word can be incorrect fit for tweet's content.

Fourth, independent lexical analysis approaches were implemented, to measure the tweet's content to overcome the limitation of Arabic language analysis tools, and BOW approach (Wagner, 2016). The tweet PageRank (PR) with weighted edge approach identified the words' influence in the tweet and the tweet semantic Orientation (OS) identified the words' semantics of being closer to positive or negative meaning.

Fifth, the multidimensional analysis model that based on five feature sets achieved 90% accuracy rate, which showed the benefit of analyzing the user's profile, tweet and social activities. In addition, when the number of the features have been reduced from 104 to 31 features the model still hold features from each feature set.

Limitations and Future Directions

There are limited researches on detecting abusive accounts in social media using parts of tweets such as hashtags, mention, or keywords, but not the full content of the tweet as it is considered a complex process. The complexity of analyzing the tweet arises

from the informality of writing tweets, given the 140 characters' maximum and the evolving nature of written language that blends both formal and informal speech. This presents challenges in analyzing tweets with incorrect grammar using natural language processing (NLP) approach, and incorrect spelling can affect both the name entity (NE) approach and finding synonyms.

Moreover, the tweet can contain words from many languages, but most of the research is using one language to analyze the tweet. Using one language and ignoring other languages in the tweet can be one of the limitations of this research, as complaining other languages could have a better result. In addition, writing the pronunciation of English word in the Arabic language does affect the classifier performance. Having the Arabic word in two formats one in English pronunciation and one in Arabic pronunciation can be identified separately as it has different spelling. This issue can be resolved by building a tool that can correct or convert the word based on pronunciations.

Furthermore, the process of building the multidimensional model for detecting the abusive accounts can be used to build a model for detecting terrorists network, phishers network, or spammers with other languages. Each of these criminal activities has its own behavior, which can be detected based on analyzing their behavior.

Conclusion

The popularity of social media has attracted cybercriminals to implement their activities on Twitter. The purposed model has been designed to detect the abusive accounts that are posting abusive content on Twitter. These accounts do post profanity,

obscenity or inappropriate words, which Twitter has prohibited any tweet that contains adult content in an image, a video, or a text.

The model has been built based on a multidimensional analysis approach of five sets that include the profile information analysis, social graph measures, tweet statistical content, and independent lexical analysis based on tweet PageRank, and tweet semantic orientation. These sets had a total of 104 features that been reduced to 31 features by using feature selection method to ensure the effectiveness of each feature and eliminate the noisy features.

In addition, correcting the misspell words on Twitter based on the tweet's content improved the classifier performance as the word dictionaries do not cover dialect and slang. Moreover, counting the number of misspell words, slang Arabic words, formal Arabic words, and corrected words did reflect different tweeting behavior between abusive and non-abusive accounts.

The model has been tested on unknown datasets and successfully reached an accuracy rate of 90% in detecting the abusive accounts. The model can be applied on Twitter or an internet filtering system to detect the unknown abusive accounts with low error rates.

APPENDIX

Table 31. Seed Words for Data Collection

Seed words w/o “@”	
ب@ز	P@e@n@i@s
ز@ي@ط	B@u@t@t
م@ر@ا@ح@م	I@n@c@e@s@t
ه@ن@و@ح@م@م	H@o@r@n@y
ك@ي@ن	S@e@x

REFERENCES

Abozinadah, E. A., Mbaziira, A. V., & Jones, J. H. J. (2015). Detection of Abusive Accounts with Arabic Tweets. *International Journal of Knowledge Engineering-IACSIT*, *1*(2), 113–119. https://doi.org/10.7763/IJKE.2015.V1.19

Abozinadah, E., & Jones, J. H. J. (2016). Improved Micor-Blog Classification for Detecting Abusive Arabic Twitter Accounts. *International Journal of Data Mining & Knowledge Management Process (IJDKP)*, *6*(6), 17–28.

Adult or sexual products and services. (2017). Retrieved April 3, 2017, from https://help.twitter.com/articles/20170427?lang=en

Agirre, E., & Soroa, A. (2009). Personalizing PageRank for Word Sense Disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 33–41). Stroudsburg, PA, USA: Association for Computational Linguistics. Retrieved from http://dl.acm.org/citation.cfm?id=1609067.1609070

Al-Jefri, M. M., & Mohammed, S. A. (2015). *Arabic spell checking technique*. Google Patents. Retrieved from https://www.google.com/patents/US9037967

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

Alsmadi, I., Al-Kabi, M., Wahbeh, A., Al-Radaideh, Q., & Al-Shawakfa, E. (2011). The Effect of Stemming on Arabic Text Classification: An Empirical Study. *Int. J. Inf. Retr. Res.*, *1*(3), 54–70. https://doi.org/10.4018/IJIRR.2011070104

Al-Sughaiyer, I. A., & Al-Kharashi, I. A. (2004). Arabic Morphological Analysis Techniques: A Comprehensive Survey. *J. Am. Soc. Inf. Sci. Technol.*, *55*(3), 189–213. https://doi.org/10.1002/asi.10368

Ayaspell project. (n.d.). Retrieved October 6, 2016, from http://ayaspell.sourceforge.net/index.html

Bassil, Y. (2012). Parallel Spell-Checking Algorithm Based on Yahoo! N-Grams Dataset. *arXiv:1204.0184 [cs]*. Retrieved from http://arxiv.org/abs/1204.0184

Benevenuto, F., Magno, G., Rodrigues, T., & Almeida, V. (2010). Detecting spammers on twitter. In *In Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS*.

Bergamini, E., & Meyerhenke, H. (2015). Approximating Betweenness Centrality in Fully-dynamic Networks. *arXiv:1510.07971 [cs]*. Retrieved from http://arxiv.org/abs/1510.07971

Black, W., Elkateb, S., Rodriguez, H., Alkhalifa, M., Vossen, P., Pease, A., & Fellbaum, C. (2006). Introducing the Arabic wordnet project. In *Proceedings of the Third International WordNet Conference* (pp. 295–300).

Bouckaert, R. R. (2003). Choosing between two learning algorithms based on calibrated tests. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)* (pp. 51–58).

Brin, S., & Page, L. (1998). The Anatomy of a Large-scale Hypertextual Web Search Engine. In *Proceedings of the Seventh International Conference on World Wide Web 7* (pp. 107–117). Amsterdam, The Netherlands, The Netherlands: Elsevier Science Publishers B. V. Retrieved from http://dl.acm.org/citation.cfm?id=297805.297827

Cateni, S., Vannucci, M., Vannocci, M., & Colla, V. (2012). Variable selection and feature extraction through artificial intelligence techniques. *Multivariate Analysis in Management, Engineering and the Science*, 103–118.

Chaabane, A., Chen, T., Cunche, M., De Cristofaro, E., Friedman, A., & Kaafar, M. A. (2014). Censorship in the Wild: Analyzing Internet Filtering in Syria. In *Proceedings of the 2014 Conference on Internet Measurement Conference* (pp. 285–298). New York, NY, USA: ACM. https://doi.org/10.1145/2663716.2663720

Cheng, H., Xing, X., Liu, X., & Lv, Q. (2015). ISC: An Iterative Social Based Classifier for Adult Account Detection on Twitter. *IEEE Transactions on Knowledge and Data Engineering*, *27*(4), 1045–1056. https://doi.org/10.1109/TKDE.2014.2357012

Chen, Y., Zhou, Y., Zhu, S., & Xu, H. (2012). Detecting Offensive Language in Social Media to Protect Adolescent Online Safety. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social*

Computing (SocialCom) (pp. 71–80). https://doi.org/10.1109/SocialCom-PASSAT.2012.55

Colla, V. (2012). A Genetic Algorithm-based approach for selecting input variables and setting relevant network parameters of a SOM-based classifier. Retrieved December 26, 2015, from http://www.percro.org/node/621

Coughlan, S. (2016, February 9). Safer Internet Day: Young ignore "social media age limit." *BBC News*. Retrieved from http://www.bbc.com/news/education-35524429

Cruse, D. A. (1986). *Lexical Semantics*. Cambridge University Press.

Darwish, K., Magdy, W., & Mourad, A. (2012). Language processing for arabic microblog retrieval. In *Proceedings of the 21st ACM international conference on Information and knowledge management* (pp. 2427–2430). ACM.

Diab, M., & Habash, N. (2007). Arabic dialect processing tutorial. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Tutorial Abstracts* (pp. 5–6). Association for Computational Linguistics.

Digital extortion on the rise. (2015, April 16). Retrieved December 27, 2015, from http://gulfnews.com/business/sectors/technology/digital-extortion-on-the-rise-1.1491734

Duwairi, Rehab M. (2007). Arabic Text Categorization. *Int. Arab J. Inf. Technol.*, *4*(2), 125–132.

Duwairi, R. M., Marji, R., Sha'ban, N., & Rushaidat, S. (2014). Sentiment Analysis in Arabic tweets. In *2014 5th International Conference on Information and Communication Systems (ICICS)* (pp. 1–6). https://doi.org/10.1109/IACS.2014.6841964

Easley, D. (2010). *Networks, crowds, and markets: reasoning about a highly connected world*. New York: Cambridge University Press.

Elbadawi, I. (2012). Social Media Usage Guidelines for the Government of the United Arab Emirates. In *Proceedings of the 6th International Conference on Theory and Practice of Electronic Governance* (pp. 508–510). New York, NY, USA: ACM. https://doi.org/10.1145/2463728.2463839

El Kourdi, M., Bensaid, A., & Rachidi, T. (2004). Automatic Arabic Document Categorization Based on the Naive Bayes Algorithm. In *Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages* (pp. 51–58). Stroudsburg, PA, USA: Association for Computational Linguistics. Retrieved from http://dl.acm.org/citation.cfm?id=1621804.1621819

Elsahar, H., & El-Beltagy, S. R. (2014). A Fully Automated Approach for Arabic Slang Lexicon Extraction from Microblogs. In *Proceedings of the 15th International Conference on Computational Linguistics and Intelligent Text Processing - Volume 8403* (pp. 79–91). New York, NY, USA: Springer-Verlag New York, Inc. https://doi.org/10.1007/978-3-642-54906-9_7

Farghaly, A., & Shaalan, K. (2009). Arabic Natural Language Processing: Challenges and Solutions, *8*(4), 14:1–14:22. https://doi.org/10.1145/1644879.1644881

Four govt agencies struggling to control porn on social media. (2014, March 18). Retrieved September 5, 2017, from http://saudigazette.com.sa/article/78244/Four-govt-agencies-struggling-to-control-porn-on-social-media

Fox, Z. (2013). Top 10 Most Popular Languages on Twitter [CHART]. Retrieved October 13, 2017, from http://mashable.com/2013/12/17/twitter-popular-languages/

Friedl, M. A., & Brodley, C. E. (1997). Decision tree classification of land cover from remotely sensed data. *Remote Sensing of Environment*, *61*(3), 399–409. https://doi.org/10.1016/S0034-4257(97)00049-7

Guyon, I., & Elisseeff, A. (2003). An Introduction to Variable and Feature Selection. *J. Mach. Learn. Res.*, *3*, 1157–1182.

Hatzivassiloglou, V., & McKeown, K. R. (1997). Predicting the Semantic Orientation of Adjectives. In *Proceedings of the Eighth Conference on European Chapter of the Association for Computational Linguistics* (pp. 174–181). Stroudsburg, PA, USA: Association for Computational Linguistics. https://doi.org/10.3115/979617.979640

how do I swear in aribic from insults.net. (1999). Retrieved November 22, 2014, from http://www.insults.net/html/swear/arabic.html

Irani, D., Webb, S., & Pu, C. (2010). Study of trend-stuffing on twitter through text classification. In *In Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS*.

Lesk, M. (1986). Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone. In *Proceedings of the 5th Annual International Conference on Systems Documentation* (pp. 24–26). New York, NY, USA: ACM. https://doi.org/10.1145/318723.318728

Levenshtein, V. (1966). Binary Codes Capable of Correcting Deletions, Insertions and Reversals, *10*(707).

Litvak, M., Last, M., Aizenman, H., Gobits, I., & Kandel, A. (2011). DegExt — A Language-Independent Graph-Based Keyphrase Extractor. In E. Mugellini, P. S. Szczepaniak, M. C. Pettenati, & M. Sokhn (Eds.), *Advances in Intelligent Web Mastering – 3* (pp. 121–130). Springer Berlin Heidelberg. Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-18029-3_13

Lopes, A. P. B., Avila, S. E. F. de, Peixoto, A. N. A., Oliveira, R. S., & Araújo, A. de A. (2009). A bag-of-features approach based on Hue-SIFT descriptor for nude detection. In *2009 17th European Signal Processing Conference* (pp. 1552–1556).

Mamoun, R., & Ahmed, M. (2016). Arabic text stemming: Comparative analysis. In *2016 Conference of Basic Sciences and Engineering Studies (SGCAC)* (pp. 88–93). https://doi.org/10.1109/SGCAC.2016.7458011

Marwick, A., & boyd, danah. (2011). To See and Be Seen: Celebrity Practice on Twitter. *Convergence*, *17*(2), 139–158. https://doi.org/10.1177/1354856510394539

Mbaziira, A. V., Abozinadah, E., & Jones Jr, J. H. (2015). Evaluating Classifiers in Detecting 419 Scams in Bilingual Cybercriminal Communities. *arXiv Preprint arXiv:1508.04123*. Retrieved from http://arxiv.org/abs/1508.04123

McCarthy, D., Koeling, R., Weeds, J., & Carroll, J. (2004). Finding Predominant Word Senses in Untagged Text. In *Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics. https://doi.org/10.3115/1218955.1218991

McCord, M., & Chuah, M. (2011). Spam Detection on Twitter Using Traditional Classifiers. In J. M. A. Calero, L. T. Yang, F. G. Mármol, L. J. G. Villalba, A. X. Li, & Y. Wang (Eds.), *Autonomic and Trusted Computing* (pp. 175–186). Springer Berlin Heidelberg. Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-23496-5_13

Miangah, T. M. (2013). FarsiSpell: A spell-checking system for Persian using a large monolingual corpus. *Literary and Linguistic Computing*, fqt008. https://doi.org/10.1093/llc/fqt008

Mihalcea, R. (2005). Unsupervised Large-vocabulary Word Sense Disambiguation with Graph-based Algorithms for Sequence Data Labeling. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural*

Language Processing (pp. 411–418). Stroudsburg, PA, USA: Association for Computational Linguistics. https://doi.org/10.3115/1220575.1220627

Mihalcea, R., & Tarau, P. (2004). TextRank: Bringing Order into Texts (pp. 404–411). Presented at the Proceedings of EMNLP 2004, Association for Computational Linguistics. Retrieved from http://aclasb.dfki.de/nlp/bib/W04-3252

Miller, G. A. (1995). WordNet: A Lexical Database for English. *Commun. ACM*, *38*(11), 39–41. https://doi.org/10.1145/219717.219748

Mo3jam: Dictionary of colloquial Arabic / Arabic Slang. (2013). Retrieved January 1, 2016, from http://en.mo3jam.com/

Muaidi, H., & Al-tarawneh, R. (2012). Towards Arabic Spell-Checker Based on N-Grams Scores. *International Journal of Computer Applications*, *53*(3). https://doi.org/http://dx.doi.org/10.5120/8400-2168

Mubarak, H., Darwish, K., & Magdy, W. (2017). Abusive Language Detection on Arabic Social Media. *ACL 2017*, 52.

Navigli, R., & Lapata, M. (2007). Graph Connectivity Measures for Unsupervised Word Sense Disambiguation. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence* (pp. 1683–1688). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. Retrieved from http://dl.acm.org/citation.cfm?id=1625275.1625548

Nguyen, V. H., Nguyen, H. T., & Snasel, V. (2015). Normalization of vietnamese tweets on twitter. In *Intelligent Data Analysis and Applications* (pp. 179–189). Springer.

nltk.stem.isri — NLTK 3.0 documentation. (n.d.). Retrieved October 7, 2016, from http://www.nltk.org/_modules/nltk/stem/isri.html

Page, L., Brin, S., Motwani, R., & Winograd, T. (1998). The PageRank Citation Ranking: Bringing Order to the Web. [Techreport]. Retrieved September 6, 2017, from http://ilpubs.stanford.edu:8090/422/

Palanisamy, Prabu, Yadav, V., & Elchuri, H. (2013). Serendio: Simple and Practical Lexicon based approach to Sentiment Analysis. *Atlanta, Georgia*.

Purohit, H., Hampton, A., Shalin, V. L., Sheth, A. P., Flach, J., & Bhatt, S. (2013). What kind of #conversation is Twitter? Mining #psycholinguistic cues for emergency coordination. *Computers in Human Behavior*, *29*(6), 2438–2447. https://doi.org/10.1016/j.chb.2013.05.007

PyArabic 0.5 : Python Package Index. (n.d.). Retrieved October 7, 2016, from https://pypi.python.org/pypi/PyArabic/0.5

Reporter, B. D. M. (2014, February 6). More than half of children use social media by the age of 10. Retrieved September 5, 2017, from http://www.dailymail.co.uk/news/article-2552658/More-half-children-use-social-media-age-10-Facebook-popular-site-youngsters-join.html

Rsheed, N. A., & Khan, M. B. (2014). Predicting the Popularity of Trending Arabic News on Twitter. In *Proceedings of the 6th International Conference on Management of Emergent Digital EcoSystems* (pp. 3:15–3:19). New York, NY, USA: ACM. https://doi.org/10.1145/2668260.2668285

Saad, M. K., & Ashour, W. (2010). Arabic morphological tools for text mining. *Corpora*, *18*, 19.

Saleem, Z. (2014). احرف الشات بالارقام - موضوع. Retrieved January 1, 2016, from http://mawdoo3.com/%D8%A7%D8%AD%D8%B1%D9%81_%D8%A7%D9%84%D8%B4%D8%A7%D8%AA_%D8%A8%D8%A7%D9%84%D8%A7%D8%B1%D9%82%D8%A7%D9%85

Sallam, R., Mousa, H., & Hussein, M. (2016). Improving Arabic Text Categorization using Normalization and Stemming Techniques. *International Journal of Computer Applications*, *135*(2), 38–43. https://doi.org/10.5120/ijca2016908328

Santos, C., Santos, E. M. dos, & Souto, E. (2012). Nudity detection based on image zoning. In *2012 11th International Conference on Information Science, Signal Processing and their Applications (ISSPA)* (pp. 1098–1103). https://doi.org/10.1109/ISSPA.2012.6310454

Saudi govt. agencies struggling to fight porn on social media. (n.d.). Retrieved November 21, 2014, from http://english.alarabiya.net/en/media/digital/2014/03/19/Saudi-govt-agencies-struggling-to-fight-porn-on-social-media.html

Saudi govt. agencies struggling to fight porn on social media - Al Arabiya News. (2014, March 19). Retrieved April 17, 2015, from http://english.alarabiya.net/en/media/digital/2014/03/19/Saudi-govt-agencies-struggling-to-fight-porn-on-social-media.html

Saudi men prime target of social media blackmail. (2015, October 4). Retrieved December 27, 2015, from http://www.arabnews.com/node/815351

Shaalan, K., Allam, A., & Gomah, A. (2003). Towards automatic spell checking for Arabic. In *Proceedings of the 4th Conference on Language Engineering, Egyptian Society of Language Engineering (ELSE), Cairo, Egypt* (pp. 21–22).

Shaalan, K., Attia, M., Pecina, P., Samih, Y., & Genabith, J. van. (2003). Arabic Word Generation and Modelling for Spell Checking. In N. C. (Conference Chair), K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, … S. Piperidis (Eds.), *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey: European Language Resources Association (ELRA).

Shekar, R., Liszka, K. J., & Chan, C. (2011). *Twitter on Drugs: Pharmaceutical Spam in Tweets*. Conference Proceedings of the 2011 International Conference on Security and Management. Las Vegas.

Singh, M., Bansal, D., & Sofat, S. (2016). Behavioral analysis and classification of spammers distributing pornographic content in social media. *Social Network Analysis and Mining*, *6*(1), 41. https://doi.org/10.1007/s13278-016-0350-0

stop-words - Stop words - Google Project Hosting. (n.d.). Retrieved November 22, 2014, from https://code.google.com/p/stop-words/

Taghva, K., Elkhoury, R., & Coombs, J. (2005). Arabic stemming without a root dictionary. In *International Conference on Information Technology: Coding and Computing*

(ITCC'05) - Volume II (Vol. 1, pp. 152–157 Vol. 1). https://doi.org/10.1109/ITCC.2005.90

Tai, Y.-J., & Kao, H.-Y. (2013). Automatic Domain-Specific Sentiment Lexicon Generation with Label Propagation. In *Proceedings of International Conference on Information Integration and Web-based Applications & Services* (pp. 53:53–53:62). New York, NY, USA: ACM. https://doi.org/10.1145/2539150.2539190

Tashaphyne 0.2 : Python Package Index. (n.d.). Retrieved October 7, 2016, from https://pypi.python.org/pypi/Tashaphyne/

Thabtah, F., Eljinini, M., Zamzeer, M., & Hadi, W. 'el. (2009). Naïve Bayesian Based on Chi Square to Categorize Arabic. *IBIMA*, *10*, 158–163.

The Twitter Rules. (n.d.). Retrieved June 17, 2017, from https://help.twitter.com/articles/18311?lang=en

Thomas, K., Grier, C., Song, D., & Paxson, V. (2011). Suspended Accounts in Retrospect: An Analysis of Twitter Spam. In *Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference* (pp. 243–258). New York, NY, USA: ACM. https://doi.org/10.1145/2068816.2068840

Thomas, K., McCoy, D., Grier, C., Kolcz, A., & Paxson, V. (2013). Trafficking Fraudulent Accounts: The Role of the Underground Market in Twitter Spam and Abuse. In *Proceedings of the 22Nd USENIX Conference on Security* (pp. 195–210). Berkeley, CA, USA: USENIX Association. Retrieved from http://dl.acm.org/citation.cfm?id=2534766.2534784

Tolentino, J. (2015, April 6). 5 Types of Social Spam (and How to Prevent Them). Retrieved December 26, 2015, from http://thenextweb.com/future-of-communications/2015/04/06/5-types-of-social-spam-and-how-to-prevent-them/

Tong, S., & Koller, D. (2002). Support Vector Machine Active Learning with Applications to Text Classification. *J. Mach. Learn. Res.*, *2*, 45–66. https://doi.org/10.1162/153244302760185243

Turney, P. D. (2000). Learning Algorithms for Keyphrase Extraction. *Information Retrieval*, *2*(4), 303–336. https://doi.org/10.1023/A:1009976227802

Turney, P. D. (2001). Mining the Web for Synonyms: PMI-IR Versus LSA on TOEFL. In *Proceedings of the 12th European Conference on Machine Learning* (pp. 491–502). London, UK, UK: Springer-Verlag. Retrieved from http://dl.acm.org/citation.cfm?id=645328.650004

Turney, P. D. (2002). 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* (pp. 417–424). Stroudsburg, PA, USA: Association for Computational Linguistics. https://doi.org/10.3115/1073083.1073153

Twitter in the Arab Region. (2011, January). Retrieved November 22, 2014, from http://www.arabsocialmediareport.com/Twitter/LineChart.aspx?&PriMenuID=18&CatID=25&mnu=

Urban Dictionary. (1999). Retrieved December 27, 2015, from http://www.urbandictionary.com/

Versteegh, K. (2014). *The Arabic Language*. Edinburgh University Press.

Volkova, S., Wilson, T., & Yarowsky, D. (2013). Exploring Sentiment in Social Media: Bootstrapping Subjectivity Clues from Multilingual Twitter Streams. In *ACL (2)* (pp. 505–510).

Wagner, K. (2016, November 15). Twitter will let you block tweets with nasty words in its latest attempt to combat abuse. Retrieved June 6, 2017, from https://www.recode.net/2016/11/15/13634504/twitter-safety-abuse-features-block-keywords

Wahsheh, H. A., Al-kabi, M. N., & Alsmadi, I. M. (2012). Evaluating Arabic Spam Classifiers Using Link Analysis. In *Proceedings of the 3rd International Conference on Information and Communication Systems* (pp. 12:1–12:5). New York, NY, USA: ACM. https://doi.org/10.1145/2222444.2222456

Wahsheh, H. A., Al-Kabi, M. N., & Alsmadi, I. M. (2013). SPAR: A system to detect spam in Arabic opinions. In *2013 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT)* (pp. 1–6). https://doi.org/10.1109/AEECT.2013.6716442

Wang, A. H. (2010). Don't follow me: Spam detection in Twitter. In *Proceedings of the 2010 International Conference on Security and Cryptography (SECRYPT)* (pp. 1–10).

Wang, D., Irani, D., & Pu, C. (2011). A Social-spam Detection Framework. In *Proceedings of the 8th Annual Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference* (pp. 46–54). New York, NY, USA: ACM. https://doi.org/10.1145/2030376.2030382

Wu, S., Hofman, J. M., Mason, W. A., & Watts, D. J. (2011). Who Says What to Whom on Twitter. In *Proceedings of the 20th International Conference on World Wide Web* (pp. 705–714). New York, NY, USA: ACM. https://doi.org/10.1145/1963405.1963504

Yoon, T., Park, S.-Y., & Cho, H.-G. (2010). A Smart Filtering System for Newly Coined Profanities by Using Approximate String Alignment. In *2010 IEEE 10th International Conference on Computer and Information Technology (CIT)* (pp. 643–650). https://doi.org/10.1109/CIT.2010.129

Zafarani, R., Abbasi, M. A., & Liu, H. (2014). Chapter 3: Network Measures. In *Social Media Mining: An Introduction* (pp. 51–76). Cambridge University Press.

Zarrella, D. (n.d.). Is 22 Tweets-Per-Day the Optimum? Retrieved April 7, 2017, from https://blog.hubspot.com/blog/tabid/6307/bid/4594/Is-22-Tweets-Per-Day-the-Optimum.aspx

BIOGRAPHY

Ehab Abozinadah received his bachelor degree of Computer Science on 2004 from King Abdul-Aziz University, Master of Science in Education-Information Technology on 2008 from Western Oregon University, Master of Science in Information System on 2013 from George Mason University, Graduate Certificate in Information Security on 2016 from George Mason University. He is the director of e-learning systems in King Abdelaziz University since 2010. He has been a researcher on Security lab and Forensic lab at George Mason University since 2013. He has been away from his home country since 2005, which gave him a great experience and knowledge.