Detection of Abusive Accounts with Arabic Tweets

Ehab A. Abozinadah, Alex V. Mbaziira, and James H. Jones Jr.

***Abstract*—Twitter is one of the most popular sources for disseminating news and propaganda in the Arab region. Spammers are now creating abusive accounts to distribute adult content in Arabic tweets, which is prohibited by Arabic norms and cultures. Arab governments are facing a massive challenge to detect these accounts. This paper evaluates different machine learning algorithms for detecting abusive accounts with Arabic tweets, using Naïve Bayes (NB), Support Vector Machine (SVM), and Decision Tree (J48) classifiers. We are not aware of another existing data set of abusive accounts with Arabic tweets, and this is the first study to investigate this issue. The data set for this analysis was collected based on the top five Arabic swearing words. The results show that the Naïve Bayes (NB) classifier with 10 tweets and 100 features has the best performance with 90% accuracy rate.**

***Index Terms*—Arabic text classification, machine learning, pornographic spam, social network abuse.**

I. Introduction

Twitter is a micro blogger provider where users compose messages of not more than 140 characters. These messages are called tweets, and may contain text, pictures, videos or hyperlinks. The usernames in Twitter start with a prefix (@). Twitter users create their social networks through *followers* and *following* relationships. Tweets will be posted on the user and the 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 including the username prefixed by @ in the tweet. The tweets' topics can be indexed using hashtags for each topic. All hashtags in Twitter are preceded with the hash (#) symbol and can also be searched through Twitter's search engine.

Since the 2011 Arab spring, the number of Twitter users in Arab nations has been escalating. Twitter has registered five million active users in Arab countries, who send on average 17 million tweets a day. Twitter, like other social media, is a popular medium for disseminating news and propaganda from anti-government groups and civil activists [1]–[3]. Consequently, spammers are exploiting Twitter's popularity in the Middle East to disseminate malicious content. These mal-actors have opened up Twitter accounts to launch spamming campaigns targeting Arabic speakers within the 22 nations in the Middle East. Some of the Arab nations have attempted, but failed, to censor Internet traffic to block malicious URLs and contents from abusive social media accounts. These attempts have failed because spam detection tools trained in the English language are being implemented on Arabic spam [4], [5]. Spammers are exploiting this loophole to launch successful spam campaigns.

In the meantime, the number of abusive accounts has been increasing over time by exploiting the simplicity of using emails as a verification mechanism to create accounts on Twitter. The users of these accounts exploit this loophole to remain anonymous while they post abusive content. These accounts use profanity, swearing words, insulting words, harassment, child pornography and exploitation. Most of these tweets are created using slang, misspelled words, or combining different words into one word to be undetectable and bypass internet censorship mechanisms.

A lot of research has been conducted on data mining and machine learning on English corpus [6]–[8], but little research has been conducted on Arabic text mainly due to its morphological complexity and limited availability of software that is compatible with the Arabic language. Also, Arabic words have different meanings and spellings, the structure of Arabic sentences is different from English, the letters in Arabic have different shapes based on the letter location in the word, and words in Arabic are either masculine or feminine, and come in three different formats: singular, dual or plural. Based on our knowledge, there is no research that has been conducted on detecting abusive accounts with Arabic tweets.

Arabic nations are facing challenges in detecting abusive accounts and spam in Arabic tweets. The state of the art for the current censorship systems implemented in Arab nations is the use of keyword lists to identify the abusive accounts [9]. In this paper we focus on classifying the content of abusive accounts with Arabic tweets using machine learning algorithms. We compare the result of three well known classifiers on our Arabic corpus, namely: Naïve Bayes (NB), Support Vector Machine (SVM), and Decision Tree (J48). After choosing the best classifier for detecting abusive accounts with Arabic tweets, we find the minimum number of tweets and features that can give the best results with short indexing. Comparatively, much research has been conducted on detecting spammers in the English language without giving attention to determining the minimum number of tweets that give approximately the same result.

In summary, our paper makes the following contributions:

- We present the first data set for abusive accounts with Arabic tweets, finding that abusive accounts are using more than three hashtags in their tweets, they tweet more frequently than legitimate users, using insulting words, using slang, and not using blacklisted URLs.
- We reprocess this bag of words by removing the sequence of letters instead of stemming, finding that most of the words are slang and removing the sequence of letters identifies each word uniquely and reduces the word indexing.
- We found the effectiveness of identifying three sets of features that are based on the tweets, profile content, and

Manuscript received March 12, 2015; revised June 9, 2015.

The authors are with Computer Science Department, George Mason University, Fairfax, VA 22030 USA (e-mail: eabozina@gmu.edu, ambaziir@gmu.edu, jjonesu@gmu.edu).

social graph, where the top 20 features are a mix of the three sets.
- We compare different numbers of features, finding that 100 features have a better performance than a larger number of features.
- We evaluate different numbers of tweets, finding that 10 tweets has a better performance than larger and smaller quantities.
- We evaluate the results from three different machine learning algorithms named above; we found the Naïve Bayes classifier to produce the best result.

The rest of the paper is organized as follows: Section II discusses the background and related work. Section III describes the data and how the dataset was constructed. Section IV describes features used in the dataset, Section V describes the classification methods, while Section VI provides an evaluation of the results. Finally, section VII presents the conclusions and future work for this study.

II. Background and Related Work

Twitter is a popular micro blogger social media in the Arab region. With over 5 million active users, the top three tweeting countries since March 2014 are Saudi Arabia, Egypt and Kuwait with 40%, 17%, 10% respectively of all Arab countries [10]. In this section we present the motivation and background related to detection of abusive accounts in Arabic tweets. We first present concepts about Arabic tweets and then discuss several spam detection mechanisms based on English language tweets. We also review literature on machine learning approaches to classify an Arabic corpus.

A. Arabic Tweets

Arabic is a language with 28 letters where each letter has a variety of shapes when it comes at the beginning, middle or at the end of a word. The direction of Arabic writing is from the right to left, compared to other languages that use the English alphabet. Arabic grammar also uses accent symbols to stress pronunciation and the meaning of words. Other forms of the language are Arabic slang which varies across age-group and place within the Middle East. Slang, though a form of verbal and short written message communication, does not follow grammatical rules typical of the Arabic language.

B. Spam Detection and Machine Learning

Some governments in the Middle East have issued guidelines for adoption and use of social media by the public and government agencies [11]. However, spammers have flaunted these guidelines by generating Twitter accounts with spam containing profanity, curse words, promotion of child pornography and exploitation, harassing and swearing words. Criminals are exploiting the weak spam detection mechanisms to send spam to Arab users. The ineffective spam detection mechanisms rely on white lists and blacklists [9].

Previous studies applied machine learning to various learning problems using Arabic content. Some researchers have investigated these approaches to classify Arabic content from websites [12]. For instance, techniques like weighting Arabic words from websites have been used to predict spammer behavior [13]. Weighting of Arab words is an attempt to identify some of the most popular words used by Arabic spammers on websites.

Other studies [14] investigate the popularity of trending Arabic news instead of focusing on the popularity of words by comparing three classifying 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 include the article source, website and the number of tweets that contain the article URL, where they extract the following elements from the news website: article category, source, language, and named individuals. The internal features were the title and the description of the article. The internal features were weak and didn't yield a good result because of the complexity of Arabic language and there are no lists that exist to indicate the popularity of a word.

In [15] two classification algorithms, Naïve Bayes and Support Vector Machines are used to classify Saudi Arabian newspaper content. The evaluation uses three metrics: recall, precision and F1, and both classifiers register good performance outcomes.

In [16], Arabic web data is classified into five categories, namely health, business, culture, science, and sport. The classification was based on a Naïve Bayes classifier and the average accuracy was 68.78%. This outcome reflects the challenges for a Naïve Bayes classifier to learn from Arabic text and successfully predict outcomes.

Also, [17] uses the Naïve Bayes algorithm based on the Chi square features selection method and evaluation based on comparing different Arabic text categorizations. The data contained 1000 features, however the classifier registered the best performance when the dataset was reduced to 800 features.

In [18] three classifiers, Naïve Bayes, k-Nearest Neighbors (kNN), and distance-based classifiers, were used to categorize 1000 Arabic text corpus documents into 10 categories. In this study, the Naïve Bayes classifier outperforms the other two classifiers based on the result of recall, precision, error rate, and fallout measures.

Other studies have evaluated machine learning classifiers and built frameworks for addressing spam detection. For instance [5] built a framework to detect spam in Arabic opinions of the user feedback and comments on the web content or news. The framework has two categories and subcategories. The first category is the spammer and contains two subcategorizes: high level spammers and low level spammers. The second category is non-spammer and contains three subcategorize: positive, neutral, and negative. The user is considered a spammer if he or she uses a URL or five consecutive numbers. So, if the user uses a legitimate URL to explain his or her opinions, it will count as a spammer, and this is considered a drawback to this study.

Wahsheh *et al.*, [4] use Naïve Bayes and Decision trees algorithms to detect the Arabic link spam that aims to have a higher rank in search engines. The Decision tree had better results in detecting the link spam. Most of the Arabic link spam were using many links on their page that point to the same destination.

In Benevenuto *et al.* [6] spammers take advantage of trending topics to have their tweet visible and have a higher chance to create more traffic to their malicious URL. 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 focuses on English language trending topics and ignores other languages. The evaluation was conducted by using a Support Vector Machine (SVM) that detects 70% of the spammers and 96% of non-spammers.

In [19] tweets are classified into five categories, namely: News, Event, Opinions, Deals, and Private Messages. This categorization gives the user the privilege to choose the category he or she is interested in instead of being overwhelmed with raw data. This categorization was based on eight features extracted from the author's profile and the tweet to overcome the limitation of using the Bag-of-words (BOW) method. The features contain the authorship information, presence of shortened words and slang, time event phrases, opinioned words, emphasis on words, currency and percentage sign, mention at the beginning of the tweet, and mention within the tweet. The eight features alone attain better results than BOW, but by combining them together got more accurate results. So, BOW is more useful when using it with other features.

In [7], the authors evaluate user-based and content-based features to filter between spammer and non-spammer accounts. This study uses four different classifiers that include Random Forest, SMO, Naïve Bayesian, and k-Nearest Neighbors. The Random Forest outperforms the other classifiers. One of the user-based features used in this study is the distribution of tweets over a 24 hour period, where the authors suggest 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 put true information about their location, therefore, they could be located somewhere in the world where it is early morning in their country and it is afternoon in the U.S., when most local U.S. users are active in social networks. In addition, this study uses 100 recent tweets to classify the spammers' accounts, which is too long to effectively deactivate and suspend these accounts.

In [20] spammers are detected who take advantage of top trending topics in twitter to spread malicious links. The detection mechanism is based on building a set of 12,000 features of the tweets and 500,000 features of the associated web pages. Information gain was used to reduce the features and have a more accurate classification measure. The features of the dataset were reduced to 1000 features and 5000 features for the associated web pages by over 91% and 99% respectively. Datasets of 100 and 1000 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 1000 features.

Wang *et al.*, [21] build a Social-Spam Detection framework that can be used across different social networks. This framework has three components: a mapping and assembly component that converts social network objects into standard framework objects, a pre-filtering component for cross-checking new spam against known spam blacklist, and a classification component that uses Naïve Bayes to classify spam objects.

Another study [22] focuses on detecting profanity words in an online community instead of detecting abusive accounts. This study uses approximate string element to detect profanity words with special characters and replace these characters with the matching letter. These characters have been used to bypass the filtering process in the online communities.

TABLE I: SUMMARY OF DATASET

Type of Content	Total
Accounts	350,000
Tweets	1,300,000
Hashes	530,000
Links (URLs)	1,150,000

In [23] decision logic with four classifiers is used to detect pharmaceutical spammers on twitter. The study shows improved results by using two set of words instead of one set, where the first set is the primary set that contains all the words of pharmaceutical products, and the second set is the secondary set that contains all the words that are associated with the pharmaceutical products, for example shipping, refill, etc.

III. DATASET

Table I summarizes the dataset used in this study. The data for this paper were collected from April 1, 2014 – June 30, 2014. As we are focusing on Arabic tweets, we used the top five Arabic swearing words as searching seeds that returned most of the tweets in Arabic language. The searching seeds were picked from a website with a catalog of Arabic swear words [24]. In each search result, we picked the most recent 200 tweets. From the total result we ended up having 255 unique users. For each user we scraped the follower, following, profile information, and the most 50 recent tweets that include Arabic words, English words, numbers, characters, hashes, mentions, and links, and we did the same for the follower and the following.

We manually selected 500 accounts, which we manually labeled. The labeling process was based on checking the most recent 50 tweets, profile pictures, and hashtags. In this sample, half of the instances were labeled as abusive accounts while the other half were labeled as non-abusive.

A. Data Preprocessing

We normalized the tweets using the following steps:

- All non Arabic words, symbols, were removed
- All the numbers were removed
- There are some letters that have a similar pronunciation, and leads to misspellings for some people. For instance one of most common misspelled words [25] is (apparently = apparently). To uniquely identify each word we convert (أ إ آ to ا - ة to ه - ي to ى)
- We removed all the stop words by using the stop word list in [26].
- Based on the observation we made in the collected tweets, we found common spelling mistakes using sequences of letters in Arabic words except the name of god (Allah-الله). As shown in Table II the number of sequence letters for each letter on the collected tweets is one of the techniques the spammers use to bypass the filtering and censorship mechanisms. Unlike English, Arabic does not capitalize

letters or words. However, in English full words can be capitalized to express sentiment like venting or anger within a sentence, but in Arabic, sequences of words emphasize the point. To this end, we identify the words uniquely by taking off all the sequence of letters except name of god (الله) as it originally contains a sequence of letters.

IV. Features

In this section we explain the features extracted from Twitter accounts. To build the abusive account detection features, we extracted three different sets of features which include: profile-based features, tweet-based features, and social graph features.

A. Profile-Based Features

Profile-based features are properties extracted from account information. The profile objects comprise the number of the tweets, followers, and following. Using these features, we applied statistics to obtain the ratio of followers to followings, ratio of number of tweets to the followers, ratio of number of tweets to the following and the reputation score [8].

$$\text{Reputation} = \frac{\text{Followers}}{(\text{Followers} + \text{Followings})} \quad (1)$$

The reputation score is the number of followers divided by the total number of the people on the user network.

B. Tweet-Based Features

Tweet-based features are properties of the content of each tweet that include the text, hashes, links, and mentions. We analyze all sets of tweets for each user with three methods.

Firstly, we obtain statistical measures like maximum, minimum, average, mode, standard deviation and median of the following terms: number of hashes in the set, number of Arabic words on the set, number of English words on the set, number of symbols on the set, number of links on the set, number of mentions on the set, number of pictures on the set. In addition, we used the total averages for the following matrix: number of hashes to the number of the links, number of the Arabic words to the number of English words, number of mentions to number of links, number of hashes to number of pictures.

Secondly, we also tokenize each set of normalized tweets to 1-gram and 2-gram indexing. N-gram indexing is the process of breaking up the text into N words. We observed abusive tweets comprising swearing and slang words that are formed using one or two words.

Lastly, we used VirusTotal [27] to analyze shortened links for malicious content. This tool identifies the final web page belonging to the shortened link, the web page type, and the page malicious score (clean <= 0 < malicious). From each shortened link and expanded link we used the domain names and added them to the feature set.

C. Social Graph Features

Social graph features are extracted from concepts of social graph theory. Since we are detecting abusive accounts, we are interested in Twitter social network influence because they attract high numbers of followers. We therefore measure the eigenvector, out-degree, and in-degree for each account. Eigenvector measures the user influence on the network [28].

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.

V. Classification Methods

A. Used Classifiers

Classifiers are data mining algorithms that classify the data into categories. In this paper we used three classifiers namely: Naïve Bayes, Support Vector Machine (SVM), and Decision Tree (J48).

TABLE II: Set of Sequence for Each Letter

Letter	Sequence Set	Letter	Sequence Set
ل	700	ش	10
ض	132	س	10
خ	22	ي	92
ه	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

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 [20].

Support Vector Machine (SVM) is a set of associated supervised learning methods that classify the data based on dimensional patterns [29].

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, leafs, nodes, and branches. The decision nodes are the questioner node that feed the leaf nodes with the data subset [30].

We evaluated the performance of each classifier based on average precision (*P*), average recall (*R*), average F-measure (*F*) and accuracy (*A*). All three measures are computed from the confusion matrix. The confusion matrix is presented on Table III where (*TP_R*) represents number of non-abuser correctly classified as non-abuser, (*FP_R*) represents number of non-abuser incorrectly classified as abuser, and (*TN_R*) represents number of abuser correctly classified as abuser, (*FN_R*) represents number of abuser incorrectly classified as non-abuser. The precision (*P*), and recall (*R*) are measures of completeness and exactness respectively.

$$P = \frac{TP\_R}{(TP\_R + TN\_R)} \quad (2)$$

And

$$R = \frac{TP\_R}{(TP\_R + FN\_R)} \quad (3)$$

F-measure (*F*) is based on the precision and recall values and computed as:

$$F = \frac{2PR}{(P+R)} \quad (4)$$

Accuracy (*A*) is the correct result compare to all results and computed as:

$$A = \frac{TP\_R}{(FP\_R+TN\_R+FN\_R)} \quad (5)$$

B. Information Gain

Each word is tokenized and vectored, where each word appear as score vector of term frequency-inverse document frequency (*TF-IDF*). The term frequency (*TF*) is identifying how important the word is on the document by calculating the frequency of the word t in the document d.

$$TF(t,d) = t/d \quad (6)$$

While the invert document frequency (*IDF*) identifies the importance of the word in all documents by getting the logarithmic of the all the documents *D* divided by total documents that contain the word *N*.

$$N = \{d \in D : d \in t\} \quad (7)$$

$$IDF = log(D/N) \quad (8)$$

The score vector is result of TF-IDF:

$$TF - IDF = TF(t,d) * log(D/N) \quad (9)$$

Information gain helps in ranking the features to determine the most useful features that have a high effectiveness on the classifier and have a lower rate of classifying error. Also, the information gained helps to reduce the number of the features based on a threshold that is established by the user or by the classifier performance.

To reach better results with our data set, we evaluated our data set in two steps. In the first step we used a dataset that has the most recent 10 tweets with reduced futures of 100 instead of using the full features of 3553 in Table IV. The 100 features were chosen based on the information gain threshold of 0.04 (chosen based on our observation) and used against the three classifiers. While in the second step we used the best performing classifier against different sets of features and tweets to ensure that we reach the best performance with the minimum number of tweets and features.

As shown in Fig. 1 the top 200 features contain the most frequent terms, and we start by using 100 features to evaluate the classifier performance and compare the outcome with results from data sets with more or less features.

VI. Evaluation

We evaluate the classifier performance using datasets with different feature-sets and tweet-sets. We compare the classifiers to find the suitable classifier that detects abusive accounts with Arabic tweets. We then evaluate the classifier that gives the best results with the minimum number of features and tweets.

A. Evaluating the Three Classifiers

10-fold cross validation was used to evaluate the classifiers and get their performance result using a dataset with 10 tweets and 100 features. The cross validation will divide the data into k sets randomly. Each set will be tested against the rest of the sets (*k*-1) and the performance result will be the average of all the tests. As we mentioned above, we used the accuracy, precision, and recall to evaluate the classifier performance.

Table V shows the average result of the classifiers performance. Using average F-measure, the Naïve Bayes classifier outperforms SVM by 22%, while for J48; Naïve Bayes registers a slightly higher difference of 1%. We also evaluate the classifiers using Average precision and Naïve Bayes outperforms SVM and J48 by 17% and 1% respectively. For average recall, Naïve Bayes outperforms SVM and J48 by 21% and 1%. Based on these results Naïve Bayes performed better than the other two classifiers.

TABLE III: Confusion Matrix

Type	Prediction	
	Normal	Abuser
Normal	True Positive (TP_R)	False Negative (FN_R)
Abuser	False-Positive (FP_R)	True-Negative (TN_R)

TABLE IV: Number of Features in Each Tweeting Set

Twitter Accounts	Features
5	2500
10	2911
15	3553

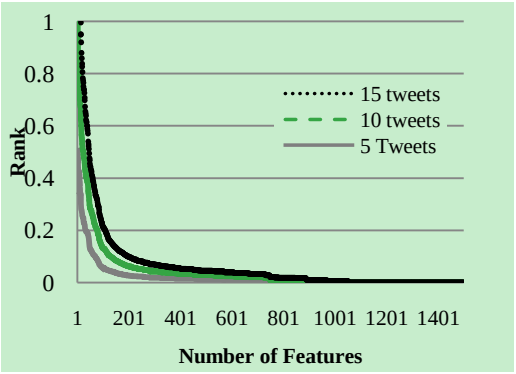

Fig. 1. Total features.

TABLE V: Classifiers Average Performance

Classifiers	P	R	F
NB	85%	85%	85%
SVM	67%	64%	63%
J48	84%	84%	84%

Furthermore, we evaluated classifier performance using classifier accuracy. Naïve Bayes out-performed Support Vector Machines and decision trees as shown in Fig. 2.

B. Evaluating Number of Tweets and Features

We further evaluated the Naïve Bayes classifier by using different sets of tweets and features for each labeled account. We started the evaluation by finding the best number of tweets with two feature sets. For each account we use three sets of recent tweets comprising 5, 10, and 15 tweets. These sets enable us determine the minimum number of tweets that can give us the best result for classifier performance.

To classify each set, we shuffled each set using a custom Python randomizer based on the native Python random number generator [31]. Then we divided each set into 80% training set and 20% testing set to predict the result. Based on

the result in

Fig. 3, the set of 10 tweets with 100 features has the best result. Also, as show in Table VI, 10 tweets with 100 features has the best average performance measure of precision, recall, f-measure and accuracy with 91%, 90%, 90%, and 90% respectively. From this result we make the following observations. Firstly, using the 10 tweets set has a better performance than 5 or 15 tweets. Secondly, the 100 features have better results than the 50 features for the 10 tweets. This finding led us to study larger numbers of features to determine which sets of features have better performance with the 10 tweets set.

Therefore, we studied larger feature sets which combined the 10 tweets set with 150 and 200 features as shown in Fig. 4 for comparison with the results from the 100 features set sample. Also, as shows in Table VII, the 100 features set outperforms 150 and 200 features based on the result of the average performance of precision, recall, f-measure, and accuracy.

In addition, with 10 tweets and 100 features we found 50% of the features came from the tokenized words of the normalized tweets and 50% based on the other features. This reflects the effectiveness of analyzing the tweets as a bag of words. Also, we picked the top 20 features as shown in Table VIII and found the top ranked feature (word) is from the tokenized words, and the 30% of top 20 features are from the tokenized words. In fact, all the presented words are abusive words. The rest of the features except feature 15 (i.e., num_follower) are based on the tweets characters. This shows how the number of hashtags on the tweet, number of URLs on tweets, and the number of the tweets could identify the account type uniquely.

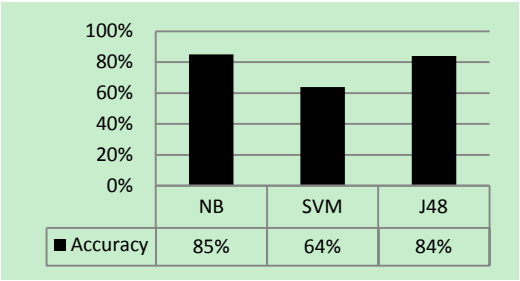

Fig. 2. Classifiers accuracy.

TABLE VI: AVERAGE PERFORMANCE OF 5, 10, AND 15 TWEETS

	P	R	F	A
5 tweets - 50 features	86%	85%	85%	85%
5tweets - 100featuers	88%	87%	87%	87%
10tweets - 50features	83%	81%	81%	81%
10tweets - 100features	91%	90%	90%	90%
15tweets - 50featuers	85%	84%	84%	84%
15tweets - 100featuers	87%	86%	86%	86%

The result of the top 100 features shows that the abusive accounts do use more hashtags than legitimate accounts. The hashtags that have been used are variety of unrelated topics that included name of countries, cities, top trending topics, profanity, slang words, and swearing words. Also, the hashtags contain a number of words instead of one word, to bypass the blacklisted hashtags. Based on our observation the accounts with variety of hashtags do have more followers than the accounts with related hashes. The abusive accounts use this technique to have their accounts appear in the search result when the user searches for any topic related or unrelated to their account activities.

TABLE VII: 100, 150 AND 200 FEATURES AVERAGE PERFORMANCE

	P	R	F	A
100 features	91%	90%	90%	90%
150 features	89%	89%	89%	89%
200 features	83%	82%	82%	82%

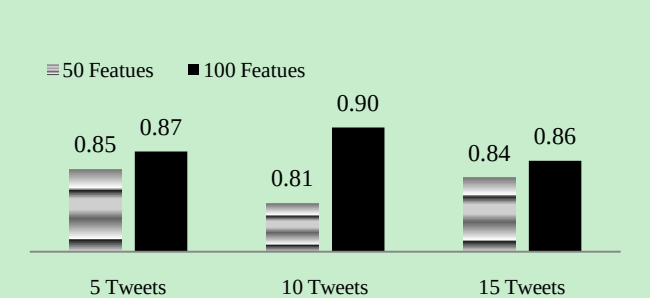

Fig. 3. Features selection.

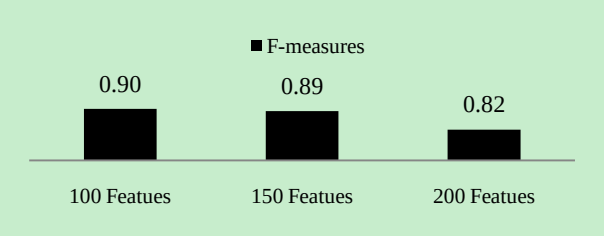

Fig. 4. 10 tweets set with different features sets.

TABLE VIII: TOP 20 FEATURES FROM THE 100 TOP FEATURES

Rank	Score	Features	Rank	Score	Features
1	0.4446	سكس	11	0.3145	av_hashes_link
2	0.4147	num_tweets	12	0.309	sd_hash
3	0.3889	sum_hash	13	0.2854	ورعان
4	0.3869	mean_hash	14	0.2825	محارم
5	0.3823	max_hash	15	0.2758	num_follower
6	0.3802	URL_total	16	0.2581	av_hash_words
7	0.3667	نيك	17	0.251	Technology
8	0.3509	median_hash	18	0.2425	طيز
9	0.3251	min_hash	19	0.2425	سكس عربي
10	0.3169	Reputation	20	0.2407	mode_hash

VII. CONCLUSION AND FUTURE WORK

In this paper we approached the problem of detecting abusive accounts with Arabic tweets by using text classification. To identify each word uniquely, we preprocess the data set by replacing some letters and removing the sequence of letters from each word, and to reduce the word indexing. Through the experiments we show how Naïve Bayes classifier outperforms Support Vector Machine and Decision Tree classifiers by using the same set of 100 features and 10 tweets, where it reaches 90% accuracy. Furthermore, we compared results of this evaluation using different sets of tweets and features to determine the minimum set of tweets and features that can achieve the highest classifier performance. The 10 tweets with 100 features have the best result with Naïve base classifier.

Also, we show the effectiveness of using the normalized tweets with the other features to identify the abusive accounts, where we found half of the features from the tokenized words and the other half based on the statistical view of the twitter accounts. In future work we will attempt to identify translated tweets from tweets that have been written by an Arabic writer. We also plan to build a framework that uses more features to detect abusive spam in other languages.

ACKNOWLEDGMENT

Ehab Abozinadah would like to thank King Abdul-Aziz University (KAU) - Saudi Arabia Culture Mission (SACM) for funding his PhD scholarship.

REFERENCE

[1] A. Bruns, T. Highfield, and J. Burgess, "The Arab spring and social media audiences English and Arabic twitter users and their networks," *Am. Behav. Sci.*, vol. 57, no. 7, pp. 871–898, Jul. 2013.
[2] C. Christensen, "Twitter revolutions? Addressing social media and dissent," *Commun. Rev.*, vol. 14, no. 3, pp. 155–157, Jul. 2011.
[3] Z. Harb, "Arab revolutions and the social media effect," *MC J.*, vol. 14, no. 2, Apr. 2011.
[4] H. A. Wahsheh, A.-K. Mohammed, and I. M. Alsmadi, "Evaluating Arabic spam classifiers using link analysis," in *Proc. the 3rd International Conference on Information and Communication Systems*, New York, NY, USA, 2012, pp. 12:1–12:5.
[5] H. A. Wahsheh, M. N. Al-Kabi, and I. M. Alsmadi, "SPAR: A system to detect spam in Arabic opinions," in *Proc. 2013 IEEE Jordan Conference on Applied Electrical Engineering and Computing* Technologies *(AEECT)*, 2013, pp. 1–6.
[6] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida, "Detecting spammers on twitter," presented at Collaboration, Electronic messaging, Anti-Abuse and Spam Conference, 2010.
[7] M. McCord and M. Chuah, "Spam detection on twitter using traditional classifiers," in *Autonomic and Trusted Computing*, J. M. A. Calero, L. T. Yang, F. G. Mármol, L. J. G. Villalba, A. X. Li, and Y. Wang, Eds., Springer Berlin Heidelberg, 2011, pp. 175–186.
[8] A. H. Wang, "Don't follow me: Spam detection in twitter," in *Proc. the 2010 International Conference on Security and Cryptography (SECRYPT)*, 2010, pp. 1–10.
[9] A. Chaabane, T. Chen, M. Cunche, E. D. Cristofaro, A. Friedman, and M. A. Kaafar, "Censorship in the Wild: Analyzing Internet Filtering in Syria," *ArXiv14023401 Cs*, Feb. 2014.
[10] Twitter in the Arab region. [Online]. Available: http://www.arabsocialmediareport.com/Twitter/LineChart.aspx?&PriMenuID=18&CatID=25&mnu=
[11] I. Elbadawi, "Social media usage guidelines for the government of the United Arab Emirates," in *Proc. the 6th International Conference on Theory and Practice of Electronic Governance*, New York, NY, USA, 2012, pp. 508–510.
[12] R. Belkebir and A. Guessoum, "A hybrid BSO-Chi2-SVM approach to Arabic text categorization," in *Proc. 2013 ACS International Conference on Computer Systems and Applications (AICCSA)*, 2013, pp. 1–7.
[13] H. A. Wahsheh, I. M. Alsmadi, and M. N. Al-Kabi, "Analyzing the popular words to evaluate spam in Arabic web pages," *Res. Bull. JORDAN ACM*, vol. 11, pp. 22–26.
[14] N. A. Rsheed and M. B. Khan, "Predicting the popularity of trending Arabic news on twitter," in *Proc. the 6th International Conference on Management of Emergent Digital EcoSystems*, New York, NY, USA, 2014, pp. 3:15–3:19.
[15] S. Alsaleem, "Automated Arabic text categorization using SVM and NB," *Int Arab J E-Technol*, vol. 2, no. 2, pp. 124–128, 2011.
[16] M. E. Kourdi, A. Bensaid, and T. Rachidi, "Automatic Arabic document categorization based on the naive bayes algorithm," in *Proc. the Workshop on Computational Approaches to Arabic Script-Based Languages*, Stroudsburg, PA, USA, 2004, pp. 51–58.
[17] F. Thabtah, M. Eljinini, M. Zamzeer, and W. Hadi, "Naïve bayesian based on chi square to categorize Arabic," *IBIMA*, vol. 10, pp. 158–163, 2009.
[18] D. Rehab, "Arabic text categorization," *Int Arab J Inf Technol*, vol. 4, no. 2, pp. 125–132, 2007.
[19] B. Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu, and M. Demirbas, "Short text classification in Twitter to improve information filtering," in *Proc. the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, NY, USA, 2010, pp. 841–842.
[20] D. Irani, S. Webb, and C. Pu, "Study of trend-stuffing on twitter through text classification," presented at Collaboration, Electronic messaging, Anti-Abuse and Spam Conference, 2010.
[21] D. Wang, D. Irani, and C. Pu, "A social-spam detection framework," in *Proc. the 8th Annual Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference*, New York, NY, USA, 2011, pp. 46–54.
[22] T. Yoon, S.-Y. Park, and H.-G. Cho, "A smart filtering system for newly coined profanities by using approximate string alignment," in *Proc. 2010 IEEE 10th International Conference on Computer and Information Technology (CIT)*, 2010, pp. 643–650.
[23] R. Shekar, K. J. Liszka, and C. Chan, *Twitter on Drugs: Pharmaceutical Spam in Tweets*.
[24] How do I swear in aribic from insults.net. [Online]. Available: http://www.insults.net/html/swear/arabic.html
[25] Common misspellings - Oxford dictionaries. [Online]. Available: http://www.oxforddictionaries.com/words/common-misspellings
[26] Stop-words - Stop words - Google Project Hosting. [Online]. Available: https://code.google.com/p/stop-words/
[27] VirusTotal - Free online virus, malware and URL scanner. [Online]. Available: https://www.virustotal.com/
[28] D. Easley, *Networks, Crowds, and Markets: Reasoning about a Highly Connected World*, New York: Cambridge University Press, 2010.
[29] S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," *J Mach Learn Res*, vol. 2, pp. 45–66, Mar. 2002.
[30] M. A. Friedl and C. E. Brodley, "Decision tree classification of land cover from remotely sensed data," *Remote Sens. Environ.*, vol. 61, no. 3, pp. 399–409, Sep. 1997.
[31] 9.6. random — Generate pseudo-random numbers — Python 2.7.9rc1 documentation. [Online]. Available: https://docs.python.org/2/library/random.html

Ehab A. Abozinadah earned a MSc in information systems from George Mason University, Fairfax VA in 2013, the graduate certificate in information security assurance from George Mason University in 2012, a MEd information technology at Western Oregon University 2008 and a BSc in computer science in 2004. He is currently pursuing a PhD in information technology (information security and assurance concentration) at George Mason University, Fairfax, VA. He was previously the director of quality assurance of e-learning systems in King Abdulaziz University, Saudi Arabia. His research interests are cybercrime, machine learning and social networks.

Alex V. Mbaziira earned a MSc in information security and assurance (advanced cybersecurity concentration) from George Mason University, Fairfax VA in 2013, a MSc in information systems at Uganda Martyrs University in 2004 and a BSc in computer science and Economics in 2000

He is currently pursuing a PhD in information technology (information security and assurance concentration) at George Mason University, Fairfax, VA as a fulbright scholar. He was previously the dean for information technology at St Lawrence University in Uganda where he also served on different committees and boards. He also served as the director for Information and Communication Technology Kampala International University, and also as a business analyst for MOGAS/Castrol Lubricants East and South Africa. His research interests are cybercrime, machine learning and cybercriminal networks.

James H. Jones Jr. earned a PhD in computational sciences and informatics from George Mason University in Fairfax, Virginia, USA, in 2008, a MS in mathematical sciences from Clemson University in Clemson, South Carolina, USA, in 1995, and a BS in industrial and systems engineering from Georgia Tech in Atlanta, Georgia, USA, in 1989.

He is currently an associate professor at George Mason University in Fairfax, Virginia, USA. He was previously on the Faculty of Ferris State University, served as a principal scientist, chief scientist, and chief technology officer for business units of SAIC, was a research scientist at the Georgia Tech Research Institute, and was a scientist for the US Department of Defense. His research interests are focused on digital artifact extraction, analysis, and manipulation, and on offensive cyber deception in adversarial environments.

Dr. Jones is a member of IAFIE, ACM, and IEEE.