Detecting Arabic Spammers and Content Polluters on Twitter

Nour El-Mawass\textsuperscript{1} and Saad Alaboodi\textsuperscript{2}
College of Computer and Information Sciences, King Saud University (Riyadh, Saudi Arabia)
\textsuperscript{1}nour.elmawass at ieee.org, \textsuperscript{2}salaboodi at ksu.edu.sa

Abstract—Spam is thriving on Arabic Twitter. With a large online population, a mounting political unrest, and an undersized and unspecialized response effort, the current state of Arabic online social networks (OSNs) offers a perfect target for the spam industry, bringing both abuse and manipulation to the scene. The result is a ubiquitous spam presence that redefines the signal to noise ratio, and makes spam a de facto component of the online social platforms. English spam on online social networks has been heavily studied in the literature. To date however, social spam in other languages has been largely ignored. Our own analysis of spam content on Arabic trending hashtags in Saudi Arabia results in an estimate of about three quarters of the total generated content. This alarming rate, backed by independent concurrent estimates, makes the development of adaptive spam detection techniques a very real and pressing need. In this study, we present a first attempt at detecting accounts that promote spam and content pollution on Arabic Twitter. Using a large crawled dataset of more than 23 million Arabic tweets, and a manually labeled sample of more than 5000 tweets, we analyze the spam content on Saudi Twitter, and assess the performance of previous spam detection features on our recently gathered dataset. We also adapt the previously proposed features to respond to spammers evading techniques, and use these features to train a new highly accurate data-driven detection system.

Keywords—Online Social Networks, Social Spam Detection, Machine Learning, Supervised Classification, Twitter, Arabic Spam

I. INTRODUCTION

The rise of online social networks (OSNs) has marked the beginning of a new spam era. Whether it is on Facebook, Twitter or other similar platforms, spam on OSNs is characterized by its invasiveness, ubiquity, and interference with the website’s main functionalities. The spammers’ aggressive abuse of OSNs threatens of reducing the value and credibility of these platforms, which may destroy their premise of an alternative, secure, and enjoyable online existence.

In our investigation, we found that about three quarters of the tweets in trending hashtags in Saudi Arabia are spam messages. This estimation is backed by independent reports placing Saudi Arabia as the number two “global target for online spam and other forms of cyber-violation”, and as the number one “most spammed country in the world for three years in a row with a 83.3\% spam rate”\cite{1}. A deeper analysis shows that the percentage of tweets generated automatically is even higher. This not only means that Twitter’s resources are being consumed by malicious accounts rather than the intended users, but it also implies the need to doubt any statistics or opinion mining results based on these trends. In particular, it becomes legitimate to question the reports that rank Saudi Arabia as the Arab nation with the highest number of active Twitter users \cite{2}, or that show a “booming usage” with a penetration higher than 51\% \cite{3}.

While the first generation of spammers on Twitter was generally naive and had obvious characteristics that helped separate it from the rest of the population, spammers nowadays recur to cheap automated techniques to gain trust and credibility and go unnoticed in the larger crowd.

In this work, we study the contemporary population of spammers on Twitter, specifically those that tweet in Arabic and that mainly hijack Saudi trends, and target the Arabic speaking population. We develop new detection features that adapt to the current evading techniques, and we use these features to train a Machine Learning-based system that has the goal of detecting spammers. We show that our approach outperforms older widely-accepted approaches, and is therefore more adapted to the current detection needs.

A. Contributions and outline

We summarize the contributions of this work in the following points:

- We provide unique statistics on the prevalence of spam and automation in trending hashtags written in Arabic. Previous work has mainly targeted English and Chinese spam \cite{4}, and we hope that this work will contribute to a greater understanding of non-English spam.
- We evaluate the performance of a widely-known, English-customized spam detection system on our recently gathered dataset.
- We evaluate previously proposed detection features, and assess the impact of evasion techniques on these features.
- We propose a set of features that are most suitable to the detection of the current Twitter’s Arabic spammers population and develop a highly accurate detection system that outperforms other models on our dataset.

The rest of this paper is organized as follows. In part II we introduce Twitter and cover related work on social spam detection. Part III is dedicated to data crawling and labeling. In part IV we explain in detail how the features are computed and how they respond to different evasion techniques. The architecture of the system is discussed in part V. In part VI, we show the results and performance of our classifier and compare it to the performance of an English-customized detection system. We discuss the obtained results in section VII and conclude the paper in part VIII.
II. Background and Related Work

A. The Twitter social network

Twitter is an online social network and microblogging platform allowing its users to write and read tweets, which are texts limited by 140 characters that can optionally contain a URL or an image. Accounts are public by default, and the friendship relationship is unidirectional: A user \( i \) can follow user \( j \) without \( j \) following \( i \) back. In this scenario, \( i \) is a “follower” of \( j \) and \( j \) is a “friend” (or alternatively a “followee”) of \( i \).

A user \( i \) can mention another user \( j \) (not necessarily a friend of \( i \)), by including its user screen name preceded by the “@” symbol in the tweet. The same convention also appears in any reply to a tweet by \( j \). A user can also “favourite” or “retweet” other users’ tweets. A hashtag is a special text entity that may be formed by one or several words preceded by the “#” symbol. Hashtags link the tweets that contain them, allowing communities and discussions to grow around specific topics. A hashtag that gains a lot of popularity in a given region becomes a “trend”, and will show on the Twitter page of users in that region.

B. Related work

The supervised learning model is extensively used in the social spam detection literature. While details of the implementation vary, the essence of the model relies on applying statistical modeling on a labeled dataset representing accounts of the targeted OSN. This trend started early in the field with applications targeting spammers on Youtube [5].

There are several approaches to collect the data. One method is to attract spammers using social honeypots. These are accounts created with the aim of mimicking the average user of the studied social network [6]. Another, more active approach, is to collect users directly either by brute force [7] or by sampling the Twitter sphere. The latter approach involves choosing a small sample of users as a seed, and expanding the network by following the social networks of these users up to a given degree [8]. Yet another fast collection method is to sample accounts that are tweeting in Twitter’s public timeline [6] (now discontinued), or in trending hashtags [8], or from a sample collected from Twitter’s streaming API [9].

In building the labeled ground-truth data, some approaches infer the spamming status of an account by manually assessing its tweets [7], [10]. An alternative is to assess the safety of URLs in tweets using safe browsing services as a mean to detect spam tweets [11].

More recent work has investigated the relationship between automation and spamming. In [12], a system for automated spammers detection is described. Features related to automation have been exploited to adapt to the changing structure of Twitter’s spammers population [10]. An analysis of automated activity in Twitter is presented in [13], and a system that detects the automation of an account is described in [14]. Similar to our work, authors in [15] target Arabic content on Twitter, their work focuses on detecting automatically generated Arabic tweets, using crowdsourcing to build their labeled collection.

The dominating methodology in the work we discussed so far is the supervised classification methodology. Other approaches require little human expert involvement and use graph and time-based parameters to cluster and detect malicious activity on social networks [16], [17].

III. Data Collection

A. Crawling Twitter

We used two main ways to collect our dataset. First, we used Twitter API to crawl 7 trending hashtags in Saudi Arabia in the period between 26 November and 15 December 2014. This resulted in a dataset containing 319,390 unique Arabic tweets, and 102,131 unique account identifiers. In addition, we used the search API on December 13, 2014 to collect 22,771,358 unique tweets generated by 1,816,668 unique account identifiers. Table III-A summarizes the information of the crawled dataset.

B. Building the labeled dataset

To ensure that the labeled dataset contains a wealth of spammers (both in numbers and variation), we had to recur to a source that is known to attract spamming activity. We selected a trending hashtag related to an important sport event (the 22nd Arabian Gulf Cup final match) and pulled all the tweets that contained the chosen related hashtag (see table III-B for details).

We randomly chose 10% of the 55,239 obtained tweets, and classified them by mean of manual examination. 76.3% of the texts were classified as spam and 23.7% were classified as legitimate texts relevant to the hashtag context. For each tweet, we pulled the related account IDs (including the original account ID that generated the tweet and the retweeter account ID if existing), and further investigated the content of the obtained accounts, classifying them into spammers and non-spammers.

C. Manual account classification

Building on both our first hand experience and Twitter’s definition of spam, we give here our definition of a spam tweet. A tweet is considered a spam tweet if it satisfies the following conditions:

1) The tweet is not composed purely of text. That is, it contains a hashtag, a mention, a URL or an image.
2) And it is out of context, e.g. the tweet’s content or sentiment is irrelevant to the context in which the tweet is embedded. This can apply to the following overlapping definitions:

1In the aim of reproducibility, we intend to make the labeled features matrix and the trained models available online.
2Twitter’s definition is rather loose, allowing the platform a freedom in interpreting spamming behavior https://support.twitter.com/articles/18311-the-twitter-rules
3We developed a manual labeling algorithm to guide human annotators labeling both tweets and accounts. For an extensive discussion of the algorithm, we refer the interested reader to our technical report [18].
### TABLE I
CRAWLED DATASET INFORMATION

<table>
<thead>
<tr>
<th>Source</th>
<th>Period</th>
<th>Nb. of unique tweets</th>
<th>Nb. of unique account IDs</th>
</tr>
</thead>
<tbody>
<tr>
<td>7 trending hashtags in KSA</td>
<td>26 Nov - 15 Dec 2014</td>
<td>319,390</td>
<td>102,131</td>
</tr>
<tr>
<td>Twitter Streaming API</td>
<td>Dec 13 2014 (12:00 noon - 12:00 night UTC)</td>
<td>22,71,358</td>
<td>1,816,668</td>
</tr>
</tbody>
</table>

### TABLE II
INFORMATION ON THE SELECTED TRENDING HASHTAG

<table>
<thead>
<tr>
<th>Hashtag</th>
<th>Nb. Of tweets</th>
<th>Period</th>
<th>Nb. of classified tweets</th>
<th>% of spam tweets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Who’s behind the failure of the Saudi football team</td>
<td>55,239</td>
<td>26 to 27 Nov 2014</td>
<td>5255</td>
<td>73.6%</td>
</tr>
</tbody>
</table>

Fig. 1. A spam tweet that promotes a followers selling service by hijacking two unrelated trending hashtags. The tweet text says: “[This is an] automatic retweet, Khaliji (Gulf-based) accounts. Call [phone number]”. Note the high number of retweets and the random number added to the end of the text.

- The tweet topic is not related to the hashtag/trend it contains (the topic can be inferred from the text or the image).
- The URL redirects to a page not related to the tweet text or the tweet hashtag.
- The URL redirects to a malicious/phishing site.
- The tweet is advertising a product or a service by hijacking hashtags or mentioning users in an automated way.

In addition to this definition, and as of Twitter’s rules, we consider any tweet advertising a paid retweet/favorite service or selling followers to be a spam as well.

Fig. 1 shows an example of a spam tweet, which is misusing two unrelated hashtags. Its goal is to promote a followers selling service, and it provides a local phone number for interested customers.

### IV. FEATURES

#### A. Profile attributes
Using features related to the social network of a profile is a legacy of the earlier work on spam detection. This explains why a lot of evading techniques have the exact aim of evading these features. These features include the number of followers, the number of friends, and the relationship between these two variables captured by the followers per friends ratio: $\frac{nb(followers)}{nb(friends)}$.

We also looked at these measures in time by computing the number of followers (resp. friends) acquired per day:

$$nb_{/day}(followers) = \frac{nb_{total}(followers)}{age_{days}(account)}$$

Other general features of the account include the total number of tweets, the number of lists containing the account, the number of tweets favorited by the account, the age of the account, and its “global tweeting frequency”, computed as:

$$\frac{nb_{total}(tweets)}{age_{days}(account)}$$

We noticed, however, that some spamming accounts trick both the age and the tweeting frequency features by remaining relatively idle for an extended amount of time after the creation of the account, or by exhibiting spamming behavior only recently in a newly compromised account. Thus, we introduced the “recent tweeting frequency”, which computes the frequency over the last 200 tweets only.

#### B. Content attributes
Content attributes are based on the recent activity of the profile, namely the last 200 tweets.

The most general descriptors of a profile’s content are the rates of retweets, replies and of original tweets (they sum to 1). These attributes assess the interaction of the account with its environment and how much of its content is self-generated. On the tweet level, we computed the minimum, maximum, median and average number of words per tweet.

1) Entities-related attributes: We use “entity” to denote the non-natural language parts of the tweet, namely hashtags, URLs and mentions. It is natural to expect that the distributions of usage of these entities is different between spammers and non-spammers. We therefore use several features formulations to capture this difference.

On the global level (over the most recent 200 tweets), three features can be extracted: the proportion of tweets with URLs, hashtags and mentions, respectively. In addition, since spammers tend to repeat the same URL, hashtag or mention in their tweets for advertisement and visibility reasons, we compute the number of unique entities and the average number of times a unique entity is used. For URLs for example, a normal user that is not topic-focused is expected to have an average usage of 1 per url, meaning that he used each unique URL only once. A spammer on the other hand tends to have a much higher average usage per unique URL$^4$.

We use the following expression to compute this value:

$$\frac{nb_{total}(URLs)}{nb(unique(URLs))}$$

$^4$Note that the unshortened final URL is used, since the shortened URL can be different each time the tweet is generated.
The existence of these entities can be also assessed on the individual tweet level by computing the minimum, maximum, median and average number of occurrences of each type of these entities in a tweet. For example, we can compute the average number of hashtags per tweet (computed by averaging this measure over the most recent 200 tweets). The number of features obtained this way is 12 ($4 \times 3$). We also computed the same measures per word per tweet, increasing the number of features to 24 ($12 \times 2$).

A technique used by spammers to mask their heavy use of only a specific set of URLs (or entities in general) is to mimic a normal user’s behavior by introducing other entities and using them only once. This decreases the average number of uses of an entity, gearing it toward a value close to that of a legitimate user. To account for this behavior, we introduce the “diversity index” measure. This index is usually used in ecological and social sciences to measure the adjusted number of species in a population, while taking into account that these species must be evenly distributed. We use the following “True Diversity” expression to compute the diversity index $1/\sum_i p_i^2$, where $p_i$ denotes the rate of usage of an entity $i$. For example, a spammer who has used a hashtag $h_1$ 147 times, and 20 other hashtags once each, will have the number of unique hashtags equal to 21, while his hashtag diversity index will be close to one (precisely 1.29), thus reflecting the true diversity of hashtags in his tweets, i.e. that the account content is not as diverse as the na"ive raw measure suggests. Another example is a legitimate user who used 9 URLs, once each. This user will have his URLs diversity index equal to the number of unique URLs (i.e. 9), since the probability of usage of each URL is the same. We compute the diversity index for URLs, hashtags and mentions. And we use the result to compute an adjusted value of the average number of uses of each of these entities.

2) Content replication: One of the most prominent features of spamming accounts is the replication of the tweet text (possibly with different URL/hashtag/mention each time). This feature can be evaded by masking the replication with automatically generated legitimate content.

Another evasion technique aims at tricking the similarity measure by adding artificial (meaningless) variation to the tweet text (such as the random three letters word “ibf”). As such, we do not rigidly check if two texts are completely identical but rather use our own definition of similarity based on the Levenshtein edit distance (the minimum number of character edits needed to transform a string of characters into another). To ensure that the similarity is computed on the core text of the tweet and doesn’t take into account variables, we filtered the tweet text by removing any hashtag, URL, or mention. A tweet text is considered a duplicate of another if the similarity is higher than 90%. This threshold misses some duplicates that are carefully twisted (especially if the tweets are semantic duplicates\(^5\) rather than syntactic replicates [10]) but is still able to detect the duplicates that evade the rigid (exact) similarity measure. Techniques to add artificial difference to the tweet include adding a random or iterated number to the tweet (incremented by one each time the same tweet is generated), or adding randomly generated short words (such as xjl, jz1, qic).

The basic features related to similarity measures are:
1) the average similarity computed as:
$$\frac{\sum_{t_i,t_j \in T, i<j} similarity(t_i,t_j)}{|T||T-1|/2}$$
where $T$ represents the tweets extracted from the account.
2) the number of replicates (the number of tweets tweeted before by the account).

In practice, the average similarity measure is of little help, it does not clearly capture the replication aspect of an account. The number of replicates is relatively robust, since the existence of replicated content cannot be wiped by the automated masking activity.

3) Content reputation: The reputation of a specific tweet can be assessed using two measures, the number of times it was retweeted, and the number of times it was favourite. A tweet acquires a reputation if its content is relevant to a lot of people, or if the account that generated it is famous/important [19]. When neither of these conditions explain the high reputation of a tweet, this reputation can be alternatively explained by an automated retweeting/favouriting activity, which is generally closely related to spam content.

To capture tweet reputation we extracted features that measure the statistics over 200 tweets of an account, namely the minimum, maximum, median and average number of retweets and favorites per tweet (8 features in total).

4) Spam dictionary: A feature that is directly inspired by spam detection in emails, measures the proportion of tweets containing a word from a spam dictionary. Since the masking behavior can dramatically decrease the proportion of spam tweets in a spamming account, applying this feature on an account content may not be helpful in detecting complex spamming accounts. An Arabic dictionary for spam terms is not readily available. In addition, spam on Twitter does not appear to have the same lexicon of emails’ spam. Therefore we created a dictionary based on the spam tweets in our labeled dataset. The obtained list was manually filtered, and only the most frequent and relevant words were kept.

V. CLASSIFICATION

The targeted classifier is a binary classifier that outputs one of two results: spammer or non-spammer, using features extracted from the account.

A. Training methodology and evaluation metrics

We use Weka’s implementation of three Machine Learning algorithms, namely Naive Bayes, Random Forests and Support Vector Machines with Radial Basis Function (RBF) kernel. Due to the limited size of the dataset, we perform a 10-fold cross validation training. We assess the performance in terms of the precision, the false positive rate, the recall

\(^5\)We have not encountered semantic duplicates in our Arabic tweets dataset. We think the cause is the lack of services similar to Spinbot (a text rewriting service, http://spinbot.com) in Arabic.
and the F1-measure of the spammers class. We also report the area under the ROC curve and the global accuracy over the respective datasets.

B. Features selection

We use the information gain and the Chi squared selection techniques with the Ranker search method in Weka to select the features to use in our classifier. By running the search in a 10-fold cross validation manner, we obtain the same set of top 20 features from the two techniques. After removing correlated features that refer to very similar measures\(^6\), we are left with the set in Fig. 2, to which we add the “average number of retweets per tweet” feature.

VI. Results

We report the performance over three tuples of datasets and features sets/models:

1) Our set of selected features over the Arabic spammers dataset (denoted AS-dataset) described in this paper. We denote the obtained model AS-Model.
2) The set of 10 features used in an English-customized detection system [7] (denoted TS-Model) and applied over the authors’ original dataset\(^7\) (denoted TS-dataset).
3) The model trained in 2) (TS-Model) is then tested on the arabic spammers dataset (AS-dataset).

The performance results are reported in Table III. Three main observations can be made:

- Except with the Naive Bayes classifier, the TS model that has a high performance over the TS-dataset, shows a significant overall degradation over the AS-dataset.
- Our AS-Model yields a consistently higher performance over the AS-dataset than the TS-Model.
- Our model offers better recall than precision in detecting spammers (more completeness than exactness). Depending on the desired compromise between the cost of missing a spammer and that of misclassifying a benign user as a spammer, the parameters of the model can be adjusted to meet the required performance balance.

VII. Discussion

The performance degradation of the previously proposed detection model [7] can be explained by one of two – non mutually exclusive – hypotheses:

- The English and Arabic spammers populations differ significantly, making a system that is built for one population unsuitable for the other.
- Spammers’ evasion techniques are intrinsically evolving. They continuously change their behavior, drifting away from their old behavioral models, making old detection systems ineffective.

Moreover, some important observations concerning features formulated in the previous literature can be made. First, none of the naïve features related to the social network of an account (namely number of friends/followers and their ratio) were first suggested in [7], are still valid in detecting spammers. Nevertheless, a general observation is that, given the elaborated evasion techniques developed and used by the spamming community, these content features alone are not enough to achieve an adequate accuracy, and new features are therefore needed.

In particular, the used spam dictionary, being static in nature, cannot prove beneficial beyond the time-limited dataset from which it was extracted. Having a spam dictionary that can be effectively used as a detection feature requires building this dictionary in a dynamic, evolving and environment-responsive way. This constitutes a major finding since spam dictionaries are the main difference between spam uttered in different languages. A direct consequence is that spammers accounts are characterized by interlingual features that are useful in both English and non-English spam. This can stem from the underlying spam generation process, which – due to the spam-as-as-service economy – is governed by a predefined set of tools and techniques, and differ only in the content and language of the final product. In other terms, spam-generating accounts use similar tweeting and masking techniques, yielding target entities that can be detected using the same detection methods (although due to evasion and cultural differences, we suspect the statistical parameters of the detection models cannot be universal).

VIII. Conclusion and Future Work

In this paper, we offered a first contribution toward characterizing and modeling Arabic spammers on Twitter,
a subject that presents a pressing economic, political and computational problem. We proposed a high performance spammers detection system that is customized to the current Arabic spammers population on Twitter. We also presented the evasion techniques developed by spammers and how detection features can be adapted to account for the evasion effects. Using these features, we assessed the performance of our approach with newly proposed features compared to a widely-known, English-customized detection system. We compared both approaches over our recent Arabic spammers dataset and showed that a small subset of features proposed in the literature is still valid. With the wealth of data offered by an invasive and ubiquitous Arabic spam presence on Twitter, we plan to further investigate the subject, especially in terms of detection and characterization of spam content, techniques and campaigns (as opposed to spamming accounts).

### References


