An Analysis of 14 Million Tweets on Hashtag-Oriented Spamming[*]

Surendra Sedhai Aixin Sun
School of Computer Science and Engineering
Nanyang Technological University, Singapore

surendra001@e.ntu.edu.sg axsun@ntu.edu.sg

Abstract

Over the years, Twitter has become a popular platform for information dissemination and information gathering. However, the popularity of Twitter has attracted not only legitimate users but also spammers who exploit social graphs, popular keywords, and hashtags for malicious purposes. In this paper, we present a detailed analysis of the HSpam14 dataset, which contains 14 million tweets with spam and ham (*i.e.,* non-spam) labels, to understand spamming activities on Twitter. The primary focus of this paper is to analyze various aspects of spam on Twitter based on hashtags, tweet contents, and user profiles, which are useful for both tweet-level and user-level spam detection. First, we compare the usage of hashtags in spam and ham tweets based on frequency, position, orthography, and co-occurrence. Second, for content-based analysis, we analyze the variations in word usage, metadata, and near-duplicate tweets. Third, for user-based analysis, we investigate user profile information. In our study, we validate that spammers use popular hashtags to promote their tweets. We also observe differences in the usage of words in spam and ham tweets. Spam tweets are more likely to be emphasized using exclamation points and capitalized words. Further, we observe that spammers use multiple accounts to post near-duplicate tweets to promote their services and products. Unlike spammers, legitimate users are likely to provide more information such as their locations and personal descriptions in their profiles. In summary, this study presents a comprehensive analysis of hashtags, tweet contents, and user profiles in Twitter spamming.

1 Introduction

Twitter, one of the most popular micro-blogging platforms for information sharing, has not only attracted ordinary people but also politicians and celebrities. Moreover, news agencies, organizations, and marketers are actively reaching out to millions of customers via Twitter. Like other online platforms, Twitter is also aggressively exploited by spammers for various malicious purposes. Grier *et al.* reported that the clickthrough rate of URLs in tweets is 0.13%, which is two orders of magnitude higher than the clickthrough rate of spam emails [16]. Unlike email services, Twitter does not limit the frequency of tweet posts which is a desirable feature for spammers. The popularity of its active user base and its high clickthrough rate make Twitter an attractive platform for spammers.

In Twitter, users are exposed to the tweets from the other users they follow. Furthermore, Twitter users are likely to follow their followers in return for the sake of courtesy [47]. Exploiting such user behavior, spammers follow random users to obtain more followers [25]. Similarly, spammers exploit trending topics to spread their messages beyond their followers through hashtags. Thomas *et al.* reported that 17% of spam users exploit hashtags to make their tweets visible in search and trending topics [40]. A similar spamming strategy has been reported for website spam and blog spam [49]. In addition to using popular hashtags and social graphs, spammers also create fake accounts and hijack accounts to promote their products and services. The issue becomes more complicated when third-party applications post tweets on behalf of legitimate users with the permission granted by the users. As a result, tweets posted by a user may not be solely composed by the user herself. The tweets from third-party applications may contain spam information.

In earlier studies, social graph, life span of spammers, and users involved in campaigns were analyzed to better understand *user-level* spam on Twitter [39, 40]. However, because of the intrinsic difficulty in spam tweet detection and the lack of a benchmark dataset of spam tweets, there have been very limited studies on *tweet-level* spamming activities. In this paper, we report a tweet-level analysis based on hashtags and tweet content, and a user-level analysis based on user profiles. Our analysis is based on the HSpam14 dataset consisting of 14 million tweets with spam or ham labels [37].

Grier *et al.* [16] analyzed the URLs in the tweets to understand spamming activities on Twitter, whereas we provide a detailed analysis of spam and ham tweets based on the usage of the hashtags, tweet content, and user profiles. We also present a comparison of the observations in this study and previous studies [10, 40, 15, 48]. The key findings of our analyses are summarized as follows:

---

[*] This paper is an extended version of a SIGIR'15 paper (Sedhai & Sun, 2015) [37], which describes the construction of the HSpam14 dataset.

- Spam tweets contain more hashtags compared with ham tweets. Position and orthography of hashtags used in spam and ham tweets are also significantly different.
- Spammers exploit globally popular hashtags to promote their content.
- Spam tweets are emphasized to grab the attention of users by using exclamation signs and capitalized words.
- Spammers post similar content multiple times and distribute the spamming activity among many accounts to promote their content.
- There are significant differences in the profile information of spammers and legitimate users. Legitimate users are likely to provide more information in their profiles compared with spammers.
- To remain undetected, spammers have changed their strategy from having many followers to no or few followers. This evolving strategy is observed by comparing our observations with the observations from earlier studies.

2 Related Work

Spam is a common problem in online media, and can be found in the form of emails [12], websites [38], videos [2], microblogging [1], comments [33], and reviews [22], to name a few. However, spammers use platform specific techniques to elude spam detection systems. Recently, crowdsourcing systems have also been exploited to post malicious content [26, 45]. Moreover, spammers continue changing their strategies to remain undetected from spam detection systems [7]. Hence, to detect such spam types, a wide range of information such as content, links, user behavior, and HTTP sessions have been utilized [38].

Spam on Email, Web, Comment, and Product Review. Spam emails contain unwanted and malicious information such as advertisements, fraud schemes, promotions, or malwares. Spam emails are identified by utilizing machine learning techniques as well as whitelisting and blacklisting senders, domains, or IP addresses [12]. Those techniques have also been used to detect website spam, which could be in the form of link spamming, cloaking, click spamming, and comment spamming [7, 33, 38]

Online product reviews provide opinions from users about products/services. As reviews are important information referred to by potential customers, spammers are attracted to promote their products and defame the products of competitors. User behavior as well as content information have been utilized to detect spam reviews [22, 30, 34]. Topic-model-based approach has also been used for spam product review detection [29]. Spam review detection techniques have also used features derived from part-of-speech tags, n-gram, and the sentiment of the reviews [28, 36]. As spammers tend to promote their products by posting similar content multiple times, near-duplicate detection techniques have been exploited to identify such similar content and also the spammers [3, 8, 46]. Duplicate detection and classification techniques have also been used for spam review detection [22, 23]

Spam on Twitter. Spam detection on Twitter is an active but challenging research topic because of the short and noisy nature of tweets. Moreover, it is difficult to discriminate spammers from legitimate users because they may also have many followers and post legitimate content [47]. Social honeypots are used to harvest such deceptive spam profiles on Twitter [25]. Users captured by the social honeypots are classified using standard classification techniques. Features derived from user demographics, social graphs, tweet contents and temporal aspects of user behavior have been analyzed and used to identify content polluters [26]. Network and content information of users have been effectively used to detect social spammers on micro-blogs [20]. Similarly, the credibility of tweets on trending topics is estimated based on the tweet content, user profile, topic, and propagation-based features [6].

In addition to creating multiple fake accounts and manually posting spam tweets, social bots have been exploited to spam on Twitter. Social bots are programs that automatically produce content and interact with people on social media. Social bots post tweets about popular and focused topics and follow back the users who follow the bots to elude spam detection systems [32]. Using such a simple strategy, bots can gain high influence on Twitter and may pollute timelines of users. These spam bots can be detected by social honeypot traps [25] and by using features derived from temporal behavior, tweet content, and user profile [9, 13].

Spammers also include unrelated links with trending words (*e.g.,* hashtags) in tweets [1]. To detect such spams, Hu *et al.* proposed a matrix-factorization-based model that learns lexical information from external spam resources. They also proposed online learning algorithms to cope with evolving spam activities [19]. Similarly, optimization approaches incorporating sentiment information have also been used to detect spammers on Twitter [18].

Crowdsourcing platforms such as Mechanical Turk, Freelancer, and Innocentive are open platforms to assign work to people who are willing to perform certain tasks for compensation. However, such platforms have been exploited by spammers to generate and propagate malicious campaigns and rumors. A study on the crowdsourcing sites reveals that approximately 90% of the tasks are for crowdturfing [45]. The types of malicious tasks and properties of requesters and workers on crowdsourcing sites have been analyzed in [27]. Although spammers try to mimic legitimate tweets using crowdsourcing platforms, such malicious tweets are significantly different from legitimate tweets; thus, machine learning techniques remain effective to detect malicious tweets generated from crowdsourcing systems [44].

Twitter Dataset Analysis. There are studies that focus on detailed analysis of spams on Twitter [16, 40]. Similarly, some spam-related studies on Twitter [10, 15, 48] performed brief analyses on their own datasets. It would be useful to validate those findings using another dataset. Further, it would be interesting to compare the findings from our dataset and those from earlier datasets to study the evolving behaviors of spammers. However, most existing studies are focused on analyzing different aspects of spammers and legitimate users (*i.e.,* user-level) rather than tweet contents. To this end, we present a detailed analysis of a dataset based on hashtags, tweet contents, and user profiles, and also compare our findings with those from previous studies.

Spammers exploit short URLs to camouflage their spam URLs. Characteristics of such short URLs on Twitter are analyzed using click traffic data [43], and it is reported that links shared by legitimate users and spammers are significantly different [5]. Grier *et al.* [16] analyzed spamming activities on Twitter with the primary focus on URL usage and reported that 8% of the URLs posted on Twitter are indicative of phishing, malware, and scams. Analysis of the clickthrough rate shows that the click rate of spam tweets is orders of magnitude higher than that of spam emails. Their study also shows that blacklists are too slow, allowing 90% of visitors to view a page before the page is blacklisted.

Another study shows that spammers exploit spam-as-service to post content and URLs on Twitter [40]. Their analysis indicates that 77% of spam accounts are identified by Twitter within the first day of creation, and 92% within the first three days of creation. Grier *et al.* [16] showed that spammers exploit hashtags and trending topics heavily, and a significant portion of trending hashtags are from spammers. There is a similar phenomenon on websites and blogs, where trending topics and popular search terms are hijacked for spamming purposes [17, 49].

Topological properties of Twitter social network are studied in [21]. Their analysis showed that the indegree and outdegree of the network follow the power-law distribution, which is similar to many other social networks [4]. Kwak *et al.* analyzed a Twitter dataset to understand topological characteristics of Twitter [24]. They ranked users based on their follower counts, PageRank and retweet counts. It was reported that rankings based on follower counts and PageRank are similar whereas the ranking based on retweet counts is different from these two rankings. This result shows that the influence based on follower count is different from the influence based on tweet popularity [24].

A user-based analysis has been conducted to understand differences in the behavior of spammers and legitimate users [48]. The study shows that spammers have more followers and followees than legitimate users. Moreover, it is observed that spammers use more hashtags compared with legitimate users probably to make the tweets more visible in search results [48]. Followers and followees of the fraudulent accounts and randomly selected accounts are also found to be significantly different [41]. Randomly selected accounts are likely to have more followers and followees, while fraudulent accounts are likely to have fewer followers and followees. Particularly, 50% of spammers have fewer than 10 followers and followees [41], which is different from the finding of Yardi *et al.* [48].

3 Overview of HSpam14

In this paper, we present analysis based on the HSpam14 dataset. Before conducting analysis on the collection of tweets, we briefly describe the dataset and the data annotation process. A detailed explanation and discussion related to the annotation process were reported in our earlier work [37].

3.1 Dataset

We collected the HSpam14 dataset via Twitter's streaming API using hashtags in the *trending up* and *trending down* categories reported in Hashtags.org as keywords. On average, 135 hashtags were used as query keywords each day. The data crawling process lasted for about two months, May and June 2013. In total, we collected 24.36 million tweets, which were published by 11.97 million unique users. The collected tweets contain 20.21 million hashtags and 6.97 million hyperlinks. Among the collected tweets, 14.07 million tweets are in English[1] and were labeled spam and ham to generate the HSpam14 dataset as discussed in [37].

3.2 Tweet Annotation

In this section, we briefly discuss the tweet annotation process as reported in [37]. First, such as other studies [16], we labelled tweets about quick money gain and adult content as spam tweets. Similarly, we marked tweets focused on immediate follower gains as spam tweets.

With the assumption that the majority of tweets are ham tweets, we select a subset of tweets using heuristics-based keywords for manual annotation. More specifically, three sets of tweets are selected: (i) tweets containing any of the top 100 most popular hashtags in our dataset, (ii) tweets containing keywords related to adult content, and (iii) tweets containing keywords related to quick money gain, lucky draw, free gift, etc. Based on these criteria, 7.10 million tweets out of the 14.07 million tweets are selected. However, in this step we may have missed spam tweets which can be captured during the later stage of the labeling process. The selected tweets are then grouped into near-duplicate clusters obtained by the *MinHash* algorithm [3]. Two tweets are grouped into a same cluster if their MinHash codes for 1-gram, 2-gram, and 3-gram are the same. Analysis of the clusters shows that intra-cluster similarity is more than 0.94 with both Jacquard coefficient and cosine similarity measures, and the inter-cluster similarity is less than 0.04. The top 1000 largest clusters among all clusters and the top 10 largest clusters for each keyword used for tweet selection are then selected for manual labeling.

Using the labeled clusters as ground truth, k-nearest-neighbor approach is adopted to 'grow' the labels, as in [11]. A cluster is labeled by kNN if the prediction of the label is confident based on the nearest neighbors, and if the newly predicted label would not cause misclassification of the manually labeled groundtruth clusters if these manually labeled clusters

---

[1] We used the language detection library for Java available at `http://code.google.com/p/language-detection/`

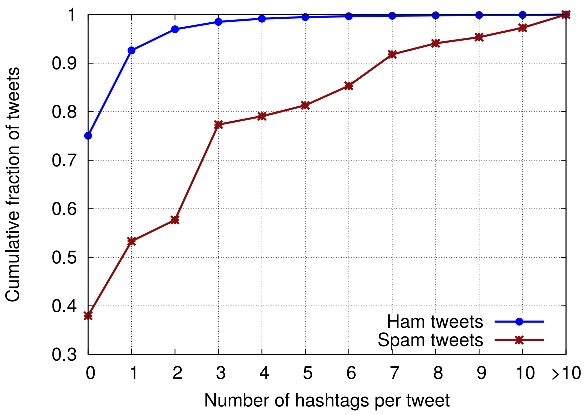


Figure 1: Cumulative fraction of spam/ham tweets with different number of hashtags per tweet

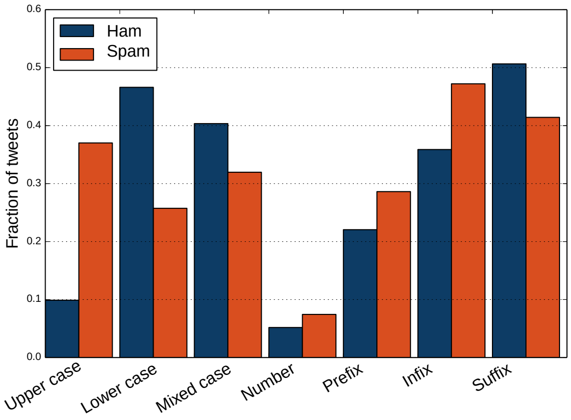


Figure 2: Orthography and position of hashtags in spam and ham tweets

were to be labeled by the kNN classifier with additional newly labeled clusters. If the cluster cannot be labeled confidently by the kNN classifier, the cluster is manually annotated. Further, we use MinHash clustering algorithm to detect potential spam users who post a large number of tweets with similar content. Similarly, based on the domain names of the links embedded in tweets, we cluster the tweets that contain links from the same domain and manually label the clusters.

To label the remaining tweets that are not labeled in previous steps, we adopt the approach for *learning from positive examples only* for classification [31] to determine reliable ham tweets. After this step, 2.386 million spam and 4.898 million ham tweets are identified. The precision of reliable ham tweet detection is 0.968, evaluated on 1000 randomly selected ham tweets. We then adopt the expectation-maximization (EM) algorithm by [35] for learning from labeled and unlabeled data to predict the labels of the remaining tweets. The precisions for spam and ham labels are 0.94 and 0.96, respectively, for the EM annotation step, evaluated on 500 randomly sampled ham and another 500 randomly sampled spam tweets.

At the end of the labeling process, there are 3.338 million spam tweets and 10.676 million ham tweets labeled in the dataset. Table 1 reports the number of tweets labeled as spam and ham in each step.

4 Hashtag Usage in Spam and Ham

In this section, we make a thorough comparison of hashtag usage patterns in spam and ham tweets. We analyze hashtags from three different perspectives: number of hashtags in each tweet, orthographic features of the hashtags, and hashtag co-occurrence. Some of the features have been used to predict hashtag usage in early studies [42].

Number of Hashtags in Tweets. We first report the number of hashtags per tweet. Figure 1 shows the cumulative fraction of tweets containing different numbers of hashtags, ranging from 0 to 10, for spam and ham tweets. Here, we observe that spammers use many hashtags per tweet, probably with the aim of reaching more users through hashtags. We can also observe that nearly 80% of legitimate tweets have no hashtags, and very few ham tweets have more than two hashtags. In contrast, more than 60% of spam tweets have one or more hashtags, and approximately 40% of them have two or more hashtags. This indicates that spam tweets tend to have more hashtags compared with ham tweets.

In our dataset, we observed that popular hashtags such as #android, #ipad are heavily exploited by spammers to promote their applications and games. Table 2 lists the top 20 most frequently used hashtags in spam and ham tweets in our dataset. The frequent hashtags in spam tweets are intended to quickly gain followers, whereas the frequent hashtags in ham tweets are about horoscope, music, love and news. We observe that the frequencies of hashtags in spam tweets are at least one order of magnitude higher than in ham tweets. Interestingly, many of the hashtags that are frequently used in spam tweets are written in capital letters.

Orthographic Features of Hashtags. In this section, we compare spam and ham tweets by analyzing three aspects, namely, (i) the usage of uppercase letters in hashtags, (ii) the usage of numbers, and (iii) the position of hashtags in tweets.

Capitalization is probably the simplest way to emphasize a tweet. In the first three columns of Figure 2, we observe that more than 35% of spam tweets contain hashtags written in all capital letters. For instance, #MUSTFOLLOW and #IPADGAGMES are listed in the top 20 most frequent hashtags in spam tweets (See Table 2). In contrast, nearly 90% of ham tweets contain hashtags written in lowercase letters or mixed cases. The fourth column in Figure 2 shows that hashtags in spam tweets contain slightly more numbers than those in ham tweets.

Now we discuss the position of hashtags, *i.e.,* whether a hashtag appears as the first, middle or last word in a tweet. In Figure 2, we observe that hashtags are more likely to appear in the middle of the spam tweet. This observation suggests that spammers are likely to use hashtags as a part of a message. In ham tweets, hashtags are more likely to appear at the end

Table 1: Number of spam and ham tweets (in millions) annotated in each step in the HSpam14 dataset.

Annotation step	Spam	Ham
Manual annotation of duplicate clusters	1.644	0.226
kNN-based annotation of clusters	0.501	0.445
User-based cluster annotation	0.019	0.002
Domain-based cluster annotation	0.221	0.121
Reliable ham tweets detection	-	4.093
EM-based annotation	0.951	5.789
Total	3.338	10.676

Table 2: Top 20 most used hashtags in ham/spam tweets. Hashtags are in descending order of their frequencies. The most frequent ham hashtag #FF appears in 43K tweets, while the 20th ham hashtag #SS5INADay2 appears in 9K tweets. The most frequent spam hashtag #TEAMFOLLOWBACK appears in 661K tweets, and the 20th spam hashtag #HITFOLLOWSTEAM appears in 127K tweets

Ham	#FF, #NowPlaying, #NP, #SoundCloud, #jobs, #news, #BELIEVEtour, #tbt, #MUFC, #Oomf, #Love, #Music, #occupygezi, #Taurus, #QnA, #Gemini, #Leo, #Virgo, #RT, #SS5INADay2
Spam	#TEAMFOLLOWBACK, #TFBJP, #gameinsight, #androidgames, #OPENFOLLOW, #androidgames, #FF, #RETWEET, #IPADGAMES, #RT, #SougoFollow, #ipad, #FOLLOWBACK, #THF, #FOLLOWNGAIN, #500aday, #AUTOFOLLOW, #MUSTFOLLOW, #TEAMHITFOLLOW, #HITFOLLOWSTEAM

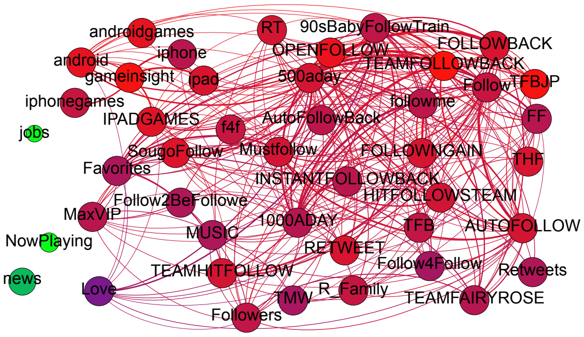

androidgames
RT
90sBabyFollowTrain
android
iphone
OPENFOLLOW
FOLLOWBACK
gameinsight
ipad
500aday
TEAMFOLLOWBACK
Follow
TFBJP
iphonegames
followme
IPADGAMES
f4f
AutoFollowBack
FF
jobs
SougoFollow
Mustfollow
FOLLOWNGAIN
THF
Favorites
INSTANTFOLLOWBACK
HITFOLLOWSTEAM
Follow2BeFollowe
MaxVIP
1000ADAY
MUSIC
TFB
AUTOFOLLOW
NowPlaying
RETWEET
TEAMHITFOLLOW
Follow4Follow
Retweets
news
Love
TMW
R_Family
TEAMFAIRYROSE
Followers

Figure 3: Hashtag co-occurrence graph. Green node is the least spammy hashtag, blue is moderately spammy, and red is the most spammy hashtag.

of the tweet content, often used as a topical indicator of the tweet.

Hashtag Co-occurrence. Nearly 90% of ham tweets contain zero or one hashtag while more than 40% of spam tweets have two or more hashtags (see Figure 1). This suggests that many co-occuring hashtags are likely to be exploited by spammers.

We propose the *Spammy Index* to quantify the extent a hashtag is used in spam tweets [37]. Given a hashtag t, its spammy index, denoted by $si(t)$, is defined in Equation 1.

$$si(t) = \log_2(df(t, D)) \times \frac{df(t, D_s)}{df(t, D)} \tag{1}$$

Where $df(t, D)$ denotes hashtag t's document frequency (*i.e.,* the number of tweets containing hashtag t) in tweet collection D; and D_s is the collection of spam tweets and $D_s \subset D$

A hashtag has a high spammy index if (i) its document frequency is high, and (ii) its probability of being used in spam tweets is high. Figure 3 plots the hashtag co-occurrence graph. Hashtags that appear more than 25,000 times in our dataset are included in this graph. The size of a node is proportional to its frequency in our dataset and width of an edge is proportional to co-occurrence frequency. The spammy index of a node is represented by the color of the node, where green is the least spammy, blue is moderately spammy and red is the most spammy.

In Figure 3, we observe that spammy hashtags mostly co-occur with spammy hashtags. Figure 3 shows that spammy hashtags such as #500aday, #TEAMFOLLOWBACK, #FOLLOWGAIN co-occur in many tweets. High co-occurrence of spammy hashtag is explained by the observation that more than 20% of spam tweets have more than four hashtags. Further, we notice that popular hashtags such as #android, #ipad, and #iphone are highly exploited in spam tweets. In contrast, less spammy hashtags are less likely to co-occur with other hashtags. This can be partially explained by the observation that 97% of ham tweets have at most two hashtags. Less spammy hashtags such as #news, #jobs, and #NowPlaying co-occur with less frequently used hashtags, such that they appear aloof in the co-occurrence graph.

Semantic Distance of Co-occurring Hashtags. In this section, we analyze the semantic distance between hashtags in spam and ham tweets. A virtual document representing a hashtag profile is created by merging all the ham tweets containing only this hashtag. Note that a tweet containing two or more hashtags is not used in the profile of any of the hashtags. By using tweets containing exactly one hashtag, we aim to more precisely define the meaning of each hashtag. The se-

Table 3: Ten co-occurring hashtag pairs in spam and ham tweets with low and high Jensen-Shannon divergence

	Low Divergence	**High Divergence**
Spam	(#followback, #TFB)	(#threewords, #NBAFinals)
	(#BlackBerry, #Apple)	(#Shoutout, #MUFC)
	(#ipadgames, #iphone)	(#BETAwards, #SingleBecause)
	(#Entertainment, #Travel)	(#ff, #androidgames)
	(#Get, #followback)	(#TeenageBio, #TheOlderIGet)
	(#life, #dead)	(#music, #Retweet)
	(#win, #competition)	(#New, #deal)
	(#Entertainment, #fashion)	(#nbafinals, #BornSinner)
	(#London, #USA)	(#MUSIC, #ANDROID)
	(#MustFollow, #FollowMe)	(#IWishIWas, #GRIZZLIES)
Ham	(#nature, #green)	(#thankyousiralex, #mufc)
	(#party, #happy)	(#heartbreaker, #musicjournals)
	(#crazy, #love)	(#OccupyGezi, #CNN)
	(#lovely, #cute)	(#CFC, #AFC)
	(#yum, #food)	(#photography, #music)
	(#CNN, #BBC)	(#24seven, #Rushers)
	(#MONEY, #business)	(#Photo, #beautiful)
	(#cute, #lol)	(#Best, #friends)
	(#tech, #technology)	(#beautiful, #Art)
	(#food, #foodporn)	

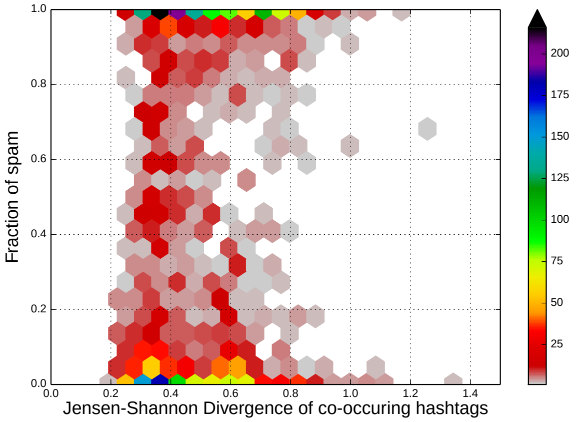


Figure 4: Semantic similarity of co-occurring hashtags and the fraction of times the pair of hashtags appears in spam tweets. The number of hashtag pairs in each hexagon is indicated by its color.

mantic distance between a pair of hashtags is estimated using the Jensen-Shannon divergence between hashtag profiles.

There are 3204 hashtags in our dataset that each appear at least 50 times in the tweets having only one hashtag. We profile these 3204 hashtags. Among these hashtags, there are 3531 pairs of hashtags that co-occur more than 50 times. The mean Jensen-Shannon divergence of these co-occurring hashtag pairs is 0.490 with standard deviation 0.169.

Figure 4 plots the relation between the semantic distance of a co-occurring hashtag pair (*i.e.,* Jensen-Shannon divergence on x-axis) and the fraction of times the pair appears in spam tweets (y-axis). We observe that both spam and ham tweets contain semantically similar as well as semantically dissimilar hashtags. In a legitimate tweet, semantically dissimilar hashtags may provide more precise meaning by providing the intersecting topics of two hashtags. For example, #mufc is a hashtag about football club Manchester United, which is used in many contexts, whereas two hashtags #mufc and #thankyousiralex together make it easier to understand that it is a farewell tweet thanking the manager of the club Manchester United, Sir Alex. In contrast, in spam tweets, spammers exploit popular hashtags without considering their semantics, probably to reach a wider audience. For example, #threeword is a popular hashtag used with three-word sentences such as ‘i love you’, and #NBAFinals is another popular hashtag; manual inspection of the tweets containing the hashtags shows that these hashtags do not provide any additional meaning to the tweets.

Popular hashtags such as #ANDROID, #music, and #NBAFinals have been heavily exploited in spam tweets, even though these hashtags are not only semantically different from the co-occurring hashtags but also found to be unrelated to the tweets from manual inspection. This observation is consistent with earlier studies on spam datasets [16]. However, legitimate tweets may also contain redundant hashtags to emphasize a particular topic. Hashtag pairs such as (#party, #happy), (#nature, #green) are used to emphasize the topic. Spam tweets also use semantically similar hashtags with the same motivation and malicious intention. In Table 3, we list more examples of hashtag pairs with low and high Jensen-Shannon divergence used in spam and ham tweets. In short, both spam and ham tweets may contain semantically similar and seman-

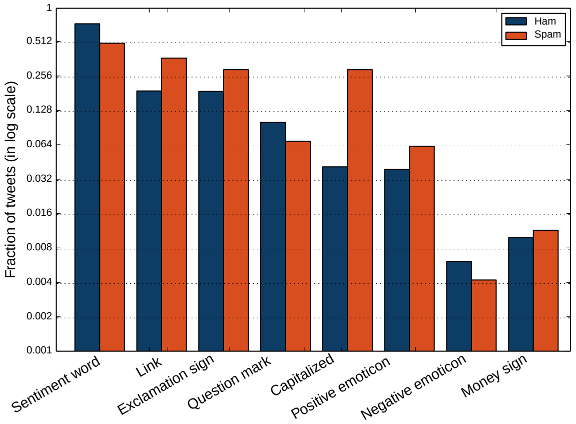


Figure 5: Differences in content metadata in spam and ham tweets

tically dissimilar hashtags. However, hashtags present in a spam tweet act as a channel to propagate the tweet to a wider audience but do not provide any additional information to the tweet.

5 Content Analysis

In this section, we perform content based analysis by analyzing frequent words and the usage of sentiment words, links, exclamation points, question marks, capitalization, positive emoticons, negative emoticons and the money sign in spam and ham tweets. These content-based metadata have also been used in many applications such as spam detection [1], social bot detection [13], and credible tweet detection [6].

5.1 Words and Content-based Metadata

Frequent Words. The usage of frequent words in spam tweets is different from that in ham tweets. Table 4 reports that spam tweets in the dataset often use spammy words such as 'follow', 'teamfollowback' and 'openfollow', whereas these words are less likely to appear in ham tweets. Furthermore, spam tweets exploit globally popular words such as 'android' and 'androidgames', whereas legitimate tweets mostly contain words commonly used in day-to-day conversations.

Content-based Metadata. In this section, we analyze content-based metadata: sentiment words,[2] links, exclamation points, question marks, capitalization, positive/negative emoticons[3] and the money sign. Figure 5 shows the presence of these attributes in spam and ham tweets. The figure indicates that ham tweets are more likely to contain sentiment words compared with spam tweets. More than 35% of spam tweets contain links whereas only 18% of the legitimate tweets contain links. The significantly greater usage of links in spam tweets is probably to promote the associated Web pages.

We also observe that exclamation points are more likely to appear in spam tweets than in ham tweets. In contrast, question marks are more likely to appear in ham tweets. We refer to a tweet containing more than 25% of characters in uppercase as a capitalized tweet. Approximately 30% of the spam tweets in our data set are capitalized tweets and only 4% of ham tweets are capitalized. Usage of money signs is slightly greater in spam tweets than in ham tweets. Spammers use more exclamation points, capitalized words, and money signs, probably to draw the attention of users. For emoticons, spam tweets are more likely to contain positive emoticons than ham tweets. Ham tweets contain relatively more negative emoticons. One reason for the lesser usage of negative emoticons in spam tweets could be that negative emoticons are less pleasing to users.

5.2 Near-duplicate Tweets

To understand the collective behavior of spammers, we analyze spam and ham tweet clusters formed by near-duplicate tweets. Near-duplicate clusters are manually labeled when annotating the HSpam14 dataset (see Table 1). The near-duplicate tweets are identified by using the MinHash based algorithm [37], and only those near-duplicate clusters that are manually labeled are used in this analysis and not the clusters labeled by kNN. There are 11,130 manually labeled clusters, of which 6,225 are spam clusters, and the remaining are ham clusters. One of the reasons for having near-duplicate spam tweets could be the collective spamming activities of users (*e.g.,* similar tweets promoting the same product/service that are posted by many spammers). Ham tweet clusters, however, are probably due to the intrinsic feature of Twitter, *e.g.,* using few words to express a statement or report events.

Previous studies show that there are spam campaigns on social networks [14, 16]. Spammers distribute spamming tasks among spam accounts to remain hidden from spam detection systems [10]. In our dataset, we also observe that spammers post content aggressively, and they distribute such activities among many users.

Aggressive Content Promotion. In our dataset, we observe that spam content is posted aggressively, probably to draw the attention of users. Of the top 10 largest clusters based on the number of tweets, there is only one cluster of ham tweets. The one cluster of ham tweets is from automated posts of Facebook on behalf of its users. The largest cluster contains 127,559 promotional tweets posted over 42 days. During this time span, there were two days in which more than 10,000 tweets were posted; on most of the remaining days, more than 1,000 tweets were posted. Similar aggressive posting of tweets is observed in the other top 10 largest clusters. Furthermore, of the top 10 clusters of spam tweets, 9 clusters contain links, and each cluster contains thousands of

---

[2] We used the sentiment word list available at `http://neuro.imm.dtu.dk/wiki/A_new_ANEW:_evaluation_of_a_word_list_for_sentiment_analysis_in_microblogs`

[3] The emoticons are listed at `http://www.datagenetics.com/blog/october52012/index.html`

Table 4: The top 20 most frequently used words in ham and spam tweets

Ham	you, my, me, so, your, love, just, have, like, i'm, all, go, up, when, get, out, new, lol, we, now
Spam	follow, followers, retweet, you, teamfollowback, new, tfbjp, gameinsight, me, want, android, i've, openfollow, my, androidgames, ff, gold, coins, collected, more

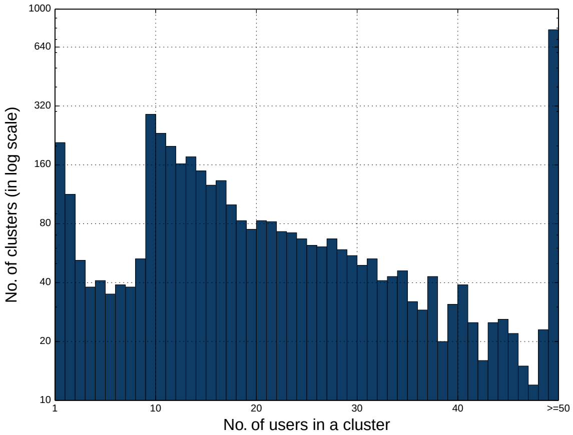


Figure 6: The distribution of the number of clusters against the number of users posting these tweets in each cluster. Tweets in each cluster contain links to one particular domain.

near-duplicate tweets with links from one particular domain. Such aggressive posting of tweets containing links from a domain is probably to promote the website.

In the HSpam14 dataset, there are 3.3 million spam tweets, of which 71.5% are detected by the near-duplicate detection approach. Such a significant fraction of near-duplicate tweets is due to the aggressive posting of similar tweets by spammers. In contrast, only 7% of ham tweets in this dataset are labeled based on near-duplicates. Near-duplicate ham tweets are due to the retweet of popular posts or tweets from legitimate online services such as Facebook, Foursquare, or Instagram. Hence, aggressively posted promotional tweets could be effectively detected by the near-duplicate detection approach.

Distributed Activity. We now analyze the near-duplicate clusters of spam tweets based on domains. There are 4398 clusters of spam tweets, where tweets in each cluster contain links to one particular domain. In other words, all the tweets in the same cluster are promoting the same website. Figure 6 plots the number of such clusters against the number of users (or tweet accounts) in each cluster who posted the tweets. The figure shows that approximately 800 clusters contain near-duplicate spam tweets that are posted by at least 50 Twitter accounts. It is less likely that legitimate users post such spam tweets. Manual inspection of the users posting these tweets shows that these accounts are likely to be created solely for posting promotional tweets. In most of these clusters, the tweets are posted by more than 10 users, whereas only approximately 5% of the clusters contain tweets by a single account. Previous studies show that multiple accounts are exploited to promote spam content which is consistent with our observation [10, 16]. Because creating a Twitter account is relatively easy, spammers may have created multiple accounts for posting spam tweets.

6 User-based Analysis

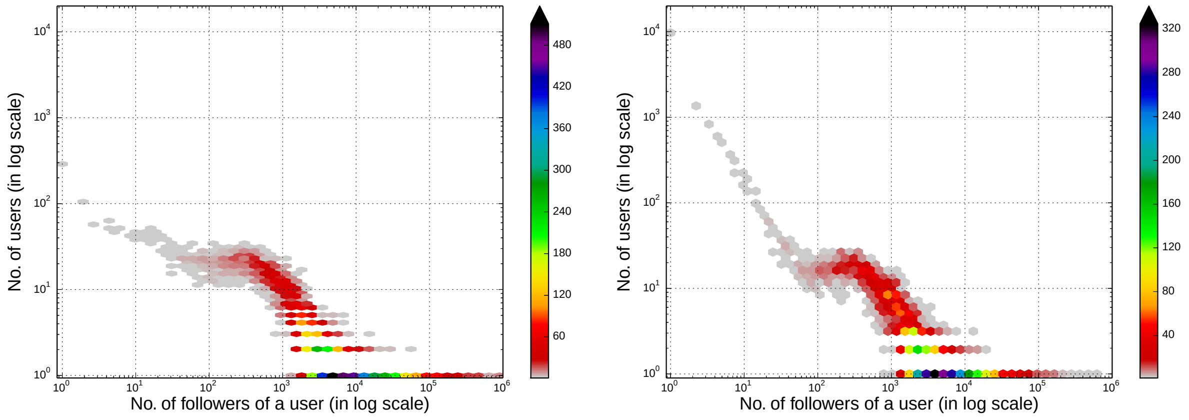
Yardi *et al.* [48] reported that the follower-to-followee ratio of spam and legitimate users is not significantly different. However, the total number of followers and followees of spammers is three times that of legitimate users. In contrast to their findings, we observe that a large number of spammers have fewer than 10 followers and/or followees, which may be a strategy of spammers to remain undetected.

We consider a user to be a *spam user* if the user has posted at least five tweets and all of them are spam tweets. Similarly, a user is considered as a *ham user* if the user has posted at least five tweets and all of them are ham tweets. There are 32,581 ham users and 36,662 spam users in the HSpam14 dataset. With these two sets of users, we analyze followers, followees and the information provided in user profiles (*e.g.,* number of tweets, location, URL, time zone, and profile description).

6.1 Followers and Followees

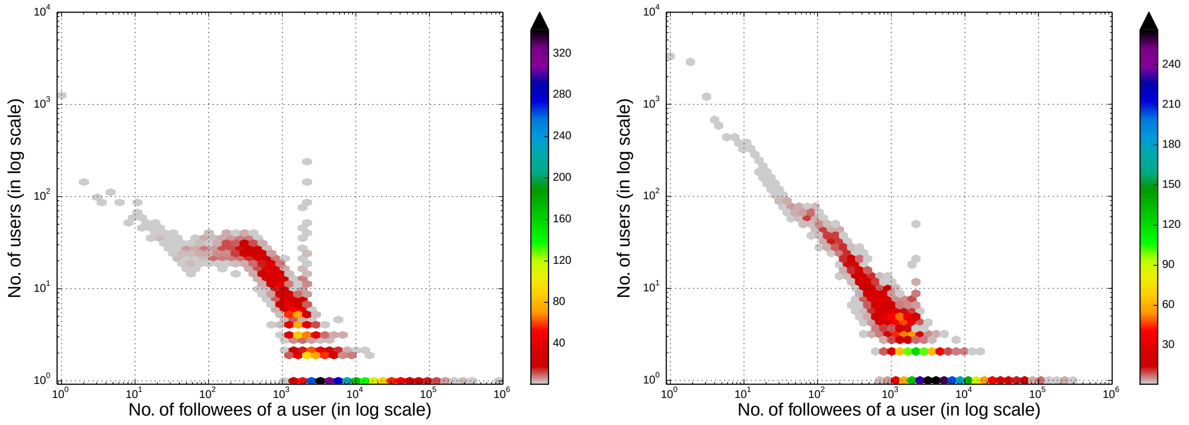
Previous studies reported that spammers try to have as many followers as possible. In our dataset, we observe that there are spammers who have thousands of followers. However, in Figure 7(b) we notice that almost 40% of spam users have fewer than 10 followers. In general, spammers with fewer followers and followees have less chance to be detected or reported as spammers, even though they post many spam tweets. Spammers may intentionally create many accounts each with a small number of followers/followees to promote their content. These spammers are difficult to identify by graph-based or social-honeypot-based spammer detection systems because they are disconnected from the social graph and do not follow many users, but content-based methods could identify these spammers.

In an earlier analysis based on a dataset of 2009 [48], the authors reported that spammers are likely to have more followers and followees compared with legitimate users. However, recent studies on a spam dataset of 2013 (which was collected around the same time as our dataset) show that approximately 40% of spammers have no follower and followee, which is consistent with our result [41]. This indicates that spammers may have changed their strategy from having more followers to hiding under the radar. Understanding such



(a) Follower distribution of ham users

(b) Follower distribution of spam users



(c) Followee distribution of ham users

(d) Followee distribution of spam users

Figure 7: Distributions of followers and followees of spam and ham users

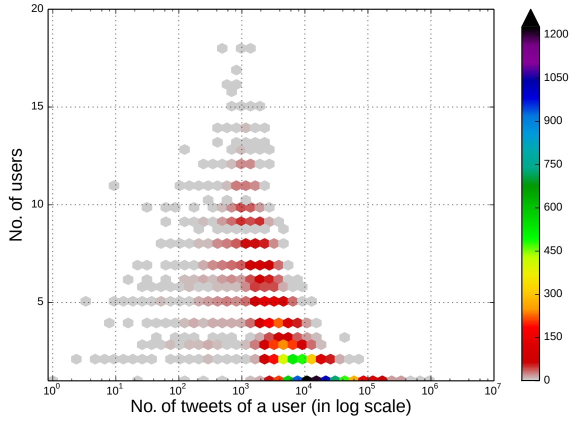
evolving behavior of spammers is crucial for the effective detection of Twitter spam.

Figures 7(a) and 7(b) show that there is an order of magnitude more spam users than ham users having fewer than 10 followers. Figures 7(c) and 7(d) show a similar pattern for the number of followees of spam and ham users. From Figures 7(a) and 7(b), it can be inferred that, in the range of having 0 to 70 followers, the number of spam users drops rapidly compared with ham users. A similar pattern is observed for the number of followees of the spam and ham users. A user having very few followers and followees could be a spammer, although a new or inactive legitimate user may also have a small number of followers/followees. In Figures 7(c) and 7(d), we observe a sudden increase in the number of users having approximately 2000 followees; similar behavior was observed in a previous study [40]. A Twitter user following other users is likely to be followed back by those users. Using this trick, a user may obtain a large number of followers. However, Twitter has a restriction on the number of users that a user can follow. If a user does not have enough followers, the user will not be able to follow more than 2000 users. The abrupt increase in the number of users having approximately 2000 followees is probably due to this restriction. Unexpectedly, Figures 7(c) and 7(d) show that both spammers and legitimate users are affected by this restriction. Further analysis of such users utilizing their profile information may give more insights.

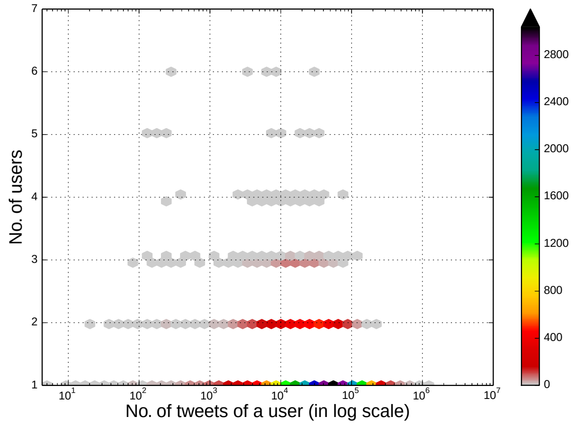
6.2 User Profile-based Analysis

In this section, we perform user profile analysis from different perspectives, namely, the number of tweets, location, URL, time zone, and description, that are commonly listed in user profiles. It is reported that spammers try to mimic the behavior of legitimate users to avoid the spam detection system [26]. However, we observe that spam and ham users have significant differences based on the information provided in user profiles .

Figures 8(a) and 8(b) report the distributions of spam and ham users, which are grouped by the number of tweets posted per user. We observe that there are a significantly large number of spammers posting 500-5000 tweets. The large number of tweets may have been posted for content promoting purposes. It can be observed that spammers post more tweets than legitimate users. However, Figure 8(b) shows that there are also legitimate accounts that post significantly more



(a) Tweets of spam users



(b) Tweets of ham users

Figure 8: Distributions of number of tweets of spam and ham users. The number of tweet-user pairs in a bin is indicated by its color.

tweets than spam users. Manual inspection of these accounts reveals that most of them are associated with news agencies, radio stations, or weather reports. Such accounts usually post the most up-to-date information, resulting in a large number of tweets.

Furthermore, Figure 9 shows that more than 70% of legitimate users provide location information, while less than 50% of spam users provide location information in their profiles. About half of ham users provide URLs in their profiles and only approximately 20% of spam users do the same. Similarly, approximately 80% of ham users provide time zone information, and the percentage of spam users doing so is less than 60%. It is also observed that more than 35% of legitimate users provide complete user profile information (*i.e.,* location, URL, time zone and description), whereas only 10% of spam users provide full information in their user profile pages. Legitimate users are more likely to provide profile descriptions than spammers. Figure 9 shows that more than 90% of legitimate users provide profile descriptions, which is much higher than the 56% of spam users. In Table 5, we list the top 20 most commonly used words in their user profiles, which shows that our observation is consistent with the observations of a previous study [15]. Although there are many words in common among the user profile descriptions of spam and ham users, the two words ‘follow’ and ‘followback’ are certainly more frequent in the descriptions of spam user profiles.

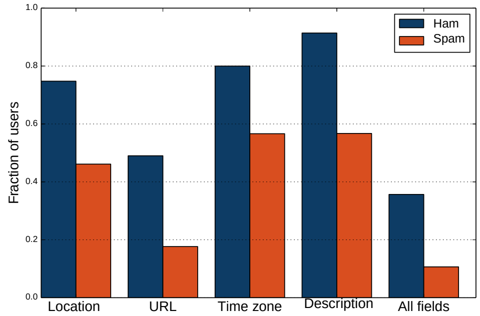


Figure 9: Fraction of spam/ham users providing information in user profile page

Ghosh *et al.* [15] reported that spammers provide significantly more information than randomly selected users, which is opposite of what we observe in our dataset. However, their observation about the word usage patterns in the profile descriptions of spammers and legitimate users is consistent with our findings. Specifically, they report that promotional and spammy words such as market, online, free, and money are frequently used by spammers in their profile descriptions; in contrast, legitimate users use words such as love, life, music, and friend in their profile descriptions [15].

7 Case Study: Spam Detection

The analyses of the various aspects of spam and ham tweets provide insights about the differences in the behavior of spammers and legitimate users. The next question is, *are the observed differences useful in identifying spam tweets from ham tweets?* Based on our analysis, in this section, we design features and evaluate their effectiveness in the spam detection task. Note that, as a case study, the spam detection task here is formulated as a simple binary classification task for the purpose of evaluating the effectiveness of the features.

We first evaluate the features based on Gini coefficients with the tweets posted on a randomly selected day (17th May 2013) in the HSpam14 dataset. There are 22,185 ham tweets (only manually labeled and reliable ham tweets are used here) and 48,849 spam tweets on the selected day. We evaluated 39 features based on our analysis, such as the usage of hashtags and user profile information. Table 6 reports the top 15 features ranked by their Gini coefficients. We observed that spammers exploit many hashtags, which are found to be discriminative features for spam tweet detection (Feature 1). As shown in Figure 1, we observe that having more than two hashtags is an important feature for identifying spam tweets. We also observe in our analysis that to remain undetected, the spammer does not follow many other users (Features 4 and

Table 5: Top 20 most commonly used words in ham and spam users' profile descriptions, ordered by their frequencies.

Ham	love, follow, news, all, life, from, just, music, like, about, don't, fan, one, world, get, twitter, have, who, more, instagram
Spam	follow, love, back, mention, life, follback, teamfollowback, roleplayer, instagram, followers, twitter, one, music, always, all, more, followback, family, que, god

Table 6: The 15 features ranked by Gini coefficient

Rank	Tweet feature	Gini
1	Contains more than 2 hashtags	0.0405
2	Contains spammy hashtag	0.0175
3	Has less than 5 percentile followers	0.0147
4	Followers-followees ratio	0.0114
5	User profile contains description	0.0110
6	Has less than 5 percentile followeees	0.0096
7	Contains capitalized hashtag	0.0062
8	Fraction of upper case characters	0.0060
9	Contains URL in the tweet	0.0046
10	Contains exclamation sign	0.0037
11	Percentile of followers of the user	0.0034
12	User profile contains time zone info	0.0032
13	Contains negative sentiment words	0.0029
14	Contains URL in user profile	0.0023
15	Suffix hashtag	0.0019

6). This insight is found to be useful for spam detection on Twitter using the Gini coefficients of the features. Similarly, having user profile information such as a profile description (Feature 5), time zone information (Feature 12) and a URL on the profile page (Feature 14), are found to be discriminative features. The presence of exclamation piont (Feature 10), URL (Feature 9), and negative sentiment words (Feature 13) in tweet contents are also found to be important attributes in identifying spam tweets. It is observed that the orthographic features of hashtags are discriminative features for spam detection. We observe that spammers use significantly more capitalized hashtags compared with legitimate users, shown in Figure 2. Interestingly, a 2012 study [42] reported that the orthography of hashtags plays an important role in better spreading ideas in the microblog community. In our dataset, where the tweets were collected in 2013, the same technique has been heavily exploited by spammers.

We further computed the Gini coefficients of the terms (*i.e.,* words and bi-grams and tri-grams) in our dataset. Table 7 lists the top 15 terms ranked by Gini coefficients and the top 15 hashtags ranked by spammy index. In the two lists, six of them are exact matches, highlighted in boldface in the table. This observation suggests that the spammy index does capture discriminative hashtags for spam detection.

To evaluate the effectiveness of these features, we trained a logistic regression classifier using the tweets posted on the selected day (17th May 2013), and then used the classifier to detect spam tweets from the tweets posted on the following day (18th May 2013). In the training dataset, we used manually labeled tweets and reliable ham tweets. There are 0.071 million training tweets and 0.371 million test tweets in this case study. From the test tweets, we randomly selected 400 tweets for manual annotation. The simple classifier achieves precision, recall, and F_1 of 0.96, 0.77, and 0.86, respectively. The high precision and relatively good recall show that spam detection at the tweet-level can be achieved with reasonable accuracy.

8 Discussion

In this section, we compare the observations from our analyses with those reported in previous studies. We observe that spammers use popular hashtags in their tweets, which is consistent with the findings reported in [16]. Tsur *et al.* [42] report that hashtag capitalization helps to promote tweets, and we observe that spammers heavily exploit capitalized hashtags to attract attention. Based on a dataset of 2009, Yardi *et al.* [48] report that spammers are likely to have more followers and followees than legitimate users. However, a recent study [41] based on a dataset of 2013 shows that approximately 40% of spammers have no followers or followees, which is consistent with our observation. The change in the behavior of spammers over time could be due to their strategy to evade graph-based spam detection system. Similar to the finding reported in [40], we also observe that both spammers and legitimate users are affected by the limit of 2,000 friends on Twitter.

Our analysis shows that spammers provide less profile information compared with legitimate users. However, Ghosh *et al.* [15] reported that link farmers make heavy use of their profile information, which is inconsistent with our observation. It would be interesting to conduct a separate behavior analysis for different types of spammers such as promoters, link farmers, crowdturfers, and social bots. However, Ghosh *et al.* [15] also reported that spammers are likely to use more promotional words, which is also observed in our study. Likewise, as in the previous studies [10, 15], our analysis also shows that spammers exploit multiple accounts to promote their content.

Our observations on the dataset may lead to more effective spam detection in Twitter. Our analysis on the features and case study on spam detection show that spam detection at tweet-level can be achieved with reasonably good accuracy by using an off-the-shelf classifier. By combining with the user-level spam detection, a more robust spam detection system could be developed. Particularly, we observe that many

Table 7: The top 15 terms (words and bi-/tri-grams) ranked by Gini coefficients and the top 15 hashtags ranked by spammy index. The matched terms and hashtags in both lists are highlighted in boldface.

Terms	follow, **follow back**, back, ipad ipadgames, please follow, please, followers, **retweet**, **tfbjp**, **teamfollowback**, **ipad**, follow me, collected, followback, **gameinsight**
Hashtags	**#TEAMFOLLOWBACK**, **#TFBJP**, **#gameinsight**, #OPENFOLLOW, #androidgames, #android, #IPADGAMES, #SougoFollow, **#FOLLOWBACK**, **#ipad**, #THF, #FOLLOWNGAIN, #500aday, **#RETWEET**, #TEAMHITFOLLOW

spammers have very few followers and may not follow a large number of accounts. Furthermore, many spammers do not provide much information on their user profile pages. The lack of user information (*e.g.,* URL) and social graph participation make the detection of user-level spam more challenging. However, spammers tend to attract the attention of users to reach a large audience by including more hashtags, capitalized words, and so on. Spammers also collectively post near-duplicate tweets because of the recency ranking of the tweets that appear in user timelines. These observations suggest using features derived from tweet content and also spam detection at tweet cluster level (*e.g.,* near-duplicates of tweets). Based on the comparison with earlier studies, spammers have indeed changed strategies to avoid detection. Therefore, a semi-supervised spam detection system should be designed to detect such evolving spammers effectively.

9 Conclusion

In this paper, we analyze various aspects of spam and ham tweets based on the HSpam14 dataset. We observe that spam tweets contain many hashtags whereas a majority of ham tweets do not contain any hashtags. Furthermore, spammers tend to use popular hashtags to reach a wider audience beyond their followers. Orthographic features of spam and ham hashtags are also found to be significantly different. Spammers tend to highlight their tweets by using capitalized words and exclamation points. Spammy words such as 'teamfollowback' and 'openfollow' are common in spam tweets, whereas such spammy words are rare in ham tweets. We observed that spammers aggressively post near-duplicate tweets to promote their tweets. Moreover, multiple accounts are engaged to promote the same content. Spammers may also remain disconnected from the social graph to avoid social graph-based spam detection. We also observed that legitimate users tend to provide detailed information on their profile pages compared with spammers. In summary, this study provides detailed insights into spamming behaviors on Twitter based on the analyses of hashtags, tweet contents, and user profile information.

We note that the HSpam14 dataset does not contain the full tweet history of all users. Because of the limited user information, spam and ham users are defined based on the labels of their tweets available in the dataset. Hence, performing analysis using the complete timeline and social graph of the users may give better insights regarding spammers and legitimate users. Because of the limitation of the dataset, we could not conduct analysis on the social graphs, and thus our analyses are mainly focused on the content of tweets.

References

[1] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida. Detecting spammers on twitter. In *Proceedings of the 7th Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS)*, 2010. Available online: http://www.decom.ufop.br/fabricio/download/ceas10.pdf.

[2] F. Benevenuto, T. Rodrigues, V. Almeida, J. Almeida, and M. Gonçalves. Detecting spammers and content promoters in online video social networks. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 620–627, New York, NY, USA, 2009. ACM.

[3] A. Broder. On the resemblance and containment of documents. In *Proceedings of the Compression and Complexity of Sequences*, pages 21–29. IEEE Computer Society, 1997.

[4] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the web. *The International Journal of Computer and Telecommunications Networking*, 33:309–320, June 2000.

[5] C. Cao and J. Caverlee. Detecting spam urls in social media via behavioral analysis. In *ECIR*, pages 703–714, 2015.

[6] C. Castillo, M. Mendoza, and B. Poblete. Information credibility on twitter. In *Proceedings of the 20th International Conference on World Wide Web*, pages 675–684, New York, NY, USA, 2011. ACM.

[7] D. Chinavle, P. Kolari, T. Oates, and T. Finin. Ensembles in adversarial classification for spam. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM)*, pages 2015–2018, Hong Kong, China, 2009.

[8] A. Chowdhury, O. Frieder, D. Grossman, and M. C. McCabe. Collection statistics for fast duplicate document detection. *ACM Transactions on Information Systems (TOIS)*, 20(2):171–191, April 2002.

[9] Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia. Who is tweeting on twitter: Human, bot, or cyborg? In *Proceedings of the 26th Annual Computer Security Applications Conference*, pages 21–30, New York, NY, USA, 2010. ACM.

[10] Z. Chu, I. Widjaja, and H. Wang. Detecting social spam campaigns on twitter. In *Proceedings of the 10th International Conference on Applied Cryptography and Network Security*, pages 455–472, Berlin, Heidelberg, 2012. Springer-Verlag.

[11] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng. Nus-wide: A real-world web image database from national university of singapore. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, pages 48:1–48:9. ACM, 2009.

[12] G. V. Cormack. Email spam filtering: A systematic review. *Foundations and Trends in Information Retrieval*, 1(4):335–455, April 2008.

[13] E. Ferrara, O. Varol, C. A. Davis, F. Menczer, and A. Flammini. The rise of social bots. *CoRR*, abs/1407.5225, 2014.

[14] H. Gao, J. Hu, C. Wilson, Z. Li, Y. Chen, and B. Y. Zhao. Detecting and characterizing social spam campaigns. In *Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement*, pages 35–47, New York, NY, USA, 2010. ACM.

[15] S. Ghosh, B. Viswanath, F. Kooti, N. K. Sharma, G. Korlam, F. Benevenuto, N. Ganguly, and K. P. Gummadi. Understanding and combating link farming in the twitter social network. In *Proceedings of the 21st International Conference on World Wide Web*, pages 61–70. ACM, 2012.

[16] C. Grier, K. Thomas, V. Paxson, and M. Zhang. @spam: The underground on 140 characters or less. In *Proceedings of the 17th ACM Conference on Computer and Communications Security*, pages 27–37, New York, NY, USA, 2010. ACM.

[17] P. Heymann, G. Koutrika, and H. Garcia-Molina. Fighting spam on social web sites: A survey of approaches and future challenges. *IEEE Internet Computing*, 11(6):36–45, 2007.

[18] X. Hu, J. Tang, H. Gao, and H. Liu. Social spammer detection with sentiment information. In *Proceedings of the 2014 IEEE International Conference on Data Mining*, pages 180–189, 2014.

[19] X. Hu, J. Tang, and H. Liu. Online social spammer detection. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, pages 59–65, Québec City, Québec, Canada., 2014.

[20] X. Hu, J. Tang, Y. Zhang, and H. Liu. Social spammer detection in microblogging. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, IJCAI '13, pages 2633–2639, 2013.

[21] A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: Understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis*, pages 56–65, New York, NY, USA, 2007. ACM.

[22] N. Jindal and B. Liu. Review spam detection. In *Proceedings of the 16th International Conference on World Wide Web*, pages 1189–1190. ACM, 2007.

[23] N. Jindal and B. Liu. Opinion spam and analysis. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, pages 219–230. ACM, 2008.

[24] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web*, pages 591–600, New York, NY, USA, 2010. ACM.

[25] K. Lee, J. Caverlee, and S. Webb. The social honeypot project: Protecting online communities from spammers. In *Proceedings of the 19th International Conference on World Wide Web*, pages 1139–1140. ACM, 2010.

[26] K. Lee, B. D. Eoff, and J. Caverlee. Seven months with the devils: A long-term study of content polluters on twitter. In *Proceedings of the 5th International Conference on Weblogs and Social Media (ICWSM)*, Barcelona, Spain, 2011.

[27] K. Lee, P. Tamilarasan, and J. Caverlee. Crowdturfers, campaigns, and social media: Tracking and revealing crowdsourced manipulation of social media. In *Proceedings of the Seventh International Conference on Weblogs and Social Media (ICWSM)*, Cambridge, Massachusetts, USA, 2013.

[28] F. Li, M. Huang, Y. Yang, and X. Zhu. Learning to identify review spam. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume 3*, pages 2488–2493. AAAI Press, 2011.

[29] J. Li, C. Cardie, and S. Li. Topicspam: a topic-model based approach for spam detection. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL), Volume 2: Short Papers*, pages 217–221, Sofia, Bulgaria, 2013.

[30] E.-P. Lim, V.-A. Nguyen, N. Jindal, B. Liu, and H. W. Lauw. Detecting product review spammers using rating behaviors. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, pages 939–948. ACM, 2010.

[31] B. Liu, Y. Dai, X. Li, W. S. Lee, and P. S. Yu. Building text classifiers using positive and unlabeled examples. In *Proceedings of the Third IEEE International Conference on Data Mining*, pages 179–186. IEEE Computer Society, 2003.

[32] J. Messias, L. Schmidt, R. Oliveira, and F. Benevenuto. You followed my bot! transforming robots into influential users in twitter. *First Monday*, 18, 2013.

[33] G. Mishne, D. Carmel, and R. Lempel. Blocking blog spam with language model disagreement. In *First International Workshop on Adversarial Information Retrieval on the Web, co-located with the WWW conference*, pages 1–6, Chiba, Japan, 2005.

[34] A. Mukherjee, B. Liu, J. Wang, N. Glance, and N. Jindal. Detecting group review spam. In *Proceedings of the 20th International Conference Companion on World Wide Web*, pages 93–94. ACM, 2011.

[35] K. Nigam, A. K. McCallum, S. Thrun, and T. Mitchell. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39(2-3):103–134, May 2000.

[36] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, pages 309–319. Association for Computational Linguistics, 2011.

[37] S. Sedhai and A. Sun. HSpam14: A collection of 14 million tweets for hashtag-oriented spam research. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 223–232. ACM, 2015.

[38] N. Spirin and J. Han. Survey on web spam detection: Principles and algorithms. *SIGKDD Explorations Newsletter*, 13(2):50–64, May 2012.

[39] E. Tan, L. Guo, S. Chen, X. Zhang, and Y. Zhao. Unik: Unsupervised social network spam detection. In *Proceedings of the 22nd ACM International Conference on Conference on Information & Knowledge Management*, pages 479–488. ACM, 2013.

[40] K. Thomas, C. Grier, D. Song, and V. Paxson. Suspended accounts in retrospect: An analysis of twitter spam. In *Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference*, pages 243–258. ACM, 2011.

[41] K. Thomas, F. Li, C. Grier, and V. Paxson. Consequences of connectivity: Characterizing account hijacking on twitter. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, CCS '14, pages 489–500, 2014.

[42] O. Tsur and A. Rappoport. What's in a hashtag?: Content based prediction of the spread of ideas in microblogging communities. In *Proceedings of the 5th ACM International Conference on Web Search and Data Mining*, pages 643–652. ACM, 2012.

[43] D. Wang, S. B. Navathe, L. Liu, D. Irani, A. Tamersoy, and C. Pu. Click traffic analysis of short url spam on twitter. In *Proceedings of the 9th Conference on Collaborative Computing: Networking, Applications and Worksharing (Collaboratecom)*, pages 250–259. IEEE, 2013.

[44] G. Wang, T. Wang, H. Zhang, and B. Y. Zhao. Man vs. machine: Practical adversarial detection of malicious crowdsourcing workers. In *Proceedings of the 23rd USENIX Conference on Security Symposium*, pages 239–254. USENIX Association, 2014.

[45] G. Wang, C. Wilson, X. Zhao, Y. Zhu, M. Mohanlal, H. Zheng, and B. Y. Zhao. Serf and turf: Crowdturfing for fun and profit. In *Proceedings of the 21st International Conference on World Wide Web*, pages 679–688. ACM, 2012.

[46] Z. Wang, W. K. Josephson, Q. Lv, M. Charikar, and K. Li. Filtering image spam with near-duplicate detection. In *Proceedings of the 4th Conference on Email and Anti-Spam*, Mountain View, California, USA, 2007.

[47] J. Weng, E.-P. Lim, J. Jiang, and Q. He. Twitterrank: Finding topic-sensitive influential twitterers. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, pages 261–270. ACM, 2010.

[48] S. Yardi, D. M. Romero, G. Schoenebeck, and D. Boyd. Detecting spam in a twitter network. *First Monday*, 15(1), 2010.

[49] L. Zhu, A. Sun, and B. Choi. Detecting spam blogs from blog search results. *Information Processing and Management*, 47(2):246–262, March 2011.