Spam Detection and Spammer Behaviour Analysis in Twitter Using Content Based Filtering Approach

B.Mukunthan\textsuperscript{1} and M.Arunkrishna\textsuperscript{2}

\textsuperscript{1} Associate professor, School of Computing, Sri Ramakrishna College of Arts and Science (Autonomous) Coimbatore-641006, Tamilnadu, India.
\textsuperscript{2} Research Scholar, Department of Computer Science, Jairams Arts and Science College(Affiliated to Bharathidhasan University), Karur - 639003, Tamilnadu, India

mukunthan@srcas.gmail.com, arunkrishna.murugan@gmail.com

Abstract. Twitter is one of the most popular social media networks and therefore it is prone to misuse. One of the ways in which people misuse Twitter is by spamming. Spam becomes an issue once a communication medium especially one, which enables global communication and handle huge amount of online data. Since Twitter is popular among so many people, it makes it easy for spammers to thrive. Spammers are people who send unwanted messages to people to either advertise a product or lure the victims into clicking malicious links, which may affect their user systems. The main objective of these spammers is usually to make money from their victims. In the last years, several systems has been made with the aim of determining whether a user is a spammer or not. However, these systems cannot filter each spam message and a different account can be created by a spammer and used to send other messages. This paper proposes a content-based approach, which can be used to filter spam tweets. The approach involves using tweets in machine learning and compression algorithms in order to filter the undesired tweets.

1. Introduction

This is technology driven era, online social media is a key aspect of communication. Global connectivity has been made possible by online platforms such as Facebook, Instagram and Twitter. In particular, this study focuses on Twitter. Twitter is a very popular social media platform where people can connect and learn about what is happening all over the world. The main objective of Twitter as an interactive platform is to allow its users to stay connected at all times through tweets. Regrettably, spammers who make use of Twitter to post malicious content \cite{12} and unwanted messages have hampered this objective. More often than not, this tends to divert attention from the trending topics and lowers the quality of the Twitter experience legitimate users have. Spam has definitely become a major problem on Twitter. In light of this, Twitter has come up with a raft of measures in a bid to tackle this growing problem. This includes the setting up of elaborate Twitter spam detection systems \cite{15}. These systems make use of machine learning and deep learning techniques to spot and flag potential spam Twitter accounts. In addition, Twitter also has a feature that allows its users to report suspected spam accounts. Machine learning algorithms use statistics \cite{4} to find patterns in data such as images, words and even numbers. Only data that can be digitally stored can be fed into a machine-learning algorithm. Many of the recommendation systems that we use today like those on Netflix, YouTube and Spotify: search engines like Google and Baidu: social-media feeds like Facebook and
Twitter are all powered by machine learning. Each of these platforms collect as much data about you as possible. Such data includes what you like watching, the statuses you react to and even the links you click. The platforms then use machine learning to make highly educated guesses about what you might want next. Honestly speaking, this process is quite basic. Simply find the pattern and then apply it [1] [3]. On the other hand, deep learning is machine learning on steroids. It uses a technique that gives machines an enhanced ability to find and amplify even the smallest patterns. This technique is called a deep neural network since it has many layers of computational nodes that work together to analyze data and deliver a result. The result is usually in the form of a prediction. One other important thing you need to know about machine learning comes in three flavors: supervised, unsupervised and reinforcement.

**Supervised learning:**

In supervised learning, the data is labelled to tell the machine exactly and specifically what patterns to look for. By Mapping an input to an output function, we can train the system based on the given sample input-output pairs.

**Unsupervised learning:**

In unsupervised learning, the data has no labels and the machine just looks for whatever patterns it can find. Unsupervised techniques are greatly used in cyber security [22]. Based on the similarities the system groups the unsorted information without any prior training.

**Reinforcement learning:**

This is the latest frontier of machine learning. A reinforcement algorithm learns by trial and error to achieve a clear objective. It tries out lots of different things and is rewarded or penalized depending on whether its behaviors help or hinder it from reaching its objective. These types of machine/deep learning have been used in Twitter to differentiate between spammers and genuine users [5]. This study sought to determine the effectiveness of using machine learning and deep learning techniques to spot spams. Spam is any form of activity that disrupts other users.

2. **Preliminary Data**

2.1 **Twitter Spam**

Most social media platforms are vulnerable to privacy and security problems since they process large amounts of personal user data every day. Twitter just like any other social media platform is no exception. The rise of Twitter as a leading global online platform has exposed it to malicious user who pollutes the information given by authentic Twitter users. This poses a great security threat to its users. The study established that there are several categories of Twitter spammers.

The first set of spammers is fake users. This refers to Twitter users whose identity cannot be authenticated. These users make use of impersonation to steal the identities of genuine Twitter users. They then use their fake profiles to send spam content targeting followers of the user whose profile they just impersonated or other Twitter users. Another category of spammers is phishers. These refer to Twitter accounts run by people pretending to be normal users but they have no interest in enjoying the Twitter experience. Their main goal is to acquire private personal data of other genuine Twitter users for malicious intentions.

The last category of Twitter spammers are promoters. These spammers masquerade as genuine online marketers and promoters but their main aim is different. They send product links and advertisements links to twitter users not to market a product or service but in order to obtain the user’s personal information.

In general, Twitter spammers are motivated by the need to spread viruses, circulate pornography and conduct phishing attacks. This ultimately compromises Twitter’s reputation, which discourages potential new users and turns off existing users. The #robotpickuplines hashtag was started by @grantimahara, a robot builder, model maker and television host on MythBusters (http://dsc.discovery.com/fansites/mythbustersmythbusters.html). As a well–known figure, he had over 20,000 followers as of June 2009. On the morning of 5 June, along with @cybernortis and others, he tweeted a series of robot jokes. Users retweeted his tweet and added their own new tweets. At around
11 am, he retweeted. Traffic spiked quickly and contained a mix of retweets and original posts, mostly sexual jokes of varying quality (e.g., references to floppy drives and hard drives). Most activity occurred in the first 24 hours of the hashtag’s lifecycle and then trailed off over the next few days. After about two hours, the hashtag was placed on Twitter’s top 10-trend list and stayed there for three - four hours, at which point it was replaced by other topics. Using whitelisted Twitter accounts, they tracked the hashtag over the next four days and captured 17,803 tweets from 8,616 unique users. User participation followed a power law distribution where 6,021 users tweeted one time, 2,595 tweeted two or more times, and a dedicated 205 tweeted 10 or more times using the #robotpickuclines [8]. Spammers started to use the hashtag when it became a trending topic, and the spam lifecycle mirrored the meme lifecycle, with a slight lag. They subsequently collected the entire 1st degree network of the 8,616 users in the dataset, which contained all their followers and friends. Followers are users who follow a given user (in-degree), and friends are users who the given user follows (out-degree). The 1st degree network contained 3,048,360 directed edges, 631,416 unique followers, and 715,198 unique friends. They used Network Workbench (Network Workbench Team, 2006) and GUESS (Adar, 2006), to analyze network properties and create the visualizations.

2.2 Literature survey

After a systematic review, a survey of existing methods for detecting spam profiles in Online Social Networks was done. Major research databases for Computer Science such as IEEE Xplore, ACM Digital Library, Google Scholar, SpringerLink and Science Direct have searched for the Twitter spamming topic. This research focused on papers after 2009. since it was after this period that social networks became popular. Even though these Online Social Networks were developed before 2009, it took some time for people to familiarize themselves with the platforms and therefore there were few cases of spamming. This is also another reason as to why the study is based on the papers after the year 2009. The only papers that used were those that I found suitable and related to the study. 21 papers were selected for review after excluding the papers whose titles and abstracts talked about irrelevant topics [22]. Mainly, the papers were categorized based on features used to detect spammers. Through this paper we are trying to compile a list of social networking papers on detection of spam profiles in Twitter that we have read. The list may likely be incomplete, but gives shape to the current research surrounding social networks spam detection. To make spam detection more accurate, new researchers can go through this paper and evaluate the work that has been done.

3. Methodology for Identifying Spammers

3.1 Features used for Spam Detection

| Table 1. Features used for Spam Profile Detection on Twitter |
|-----------------------------------------------------------|
| User based features | Which include demographic features like profile details, number of followers, number of followings, followers/following ratio, reputation, age of account, average time between tweets posting time behavior, idle hours, tweet frequency. |
| Content based features | Includes number of hashtags (#), number of URLs in tweets, @ mentions, retweets, spam words, HTTP links, trending topics, duplicate tweets. |
| User based and content based | Any other feature like graphical distance, graph connectivity: Markov clustering method, URL rate, interaction rate, social relations, social activities, and graph-based features, neighbor-based features, automation-based features |

The features mentioned above play a pivot role in determining spammer as outlined in the Twitter policy. In regards to the number off followers, spammers have a smaller number of followers because
of their poor Twitter engagements. Spammers also tend to follow a large number of people in order to increase their reach and amplify their malicious intentions. Because of this, they have a poor Twitter reputation. Twitter reputation refers to the ratio of followers to the sum of followers and following. The Twitter policy further suggests that spammers have a peculiar time schedule that includes early morning or late night. They also tend to send duplicate tweets with different user names. They also tweet more frequently at odd times in order to catch unaware unsuspecting users.

3.2 AI/ML Techniques

Using machine learning and deep learning technique [7], Twitter has been able to outsmart spammers. It has helped monitor all Twitter users and track their patterns to ensure they are in line with the Twitter policy. Machine learning has enabled the formation of Twitter alert systems that identify any spam messages and red flags spammers as shown in figure 1.

![Figure 1. Spam alert system.](image1)

With deep learning technique, Twitter systems are able to identify techniques used by spammers to trick users into clicking malicious links. These techniques include multiple mentions to followers and non-followers, hijacking top trends and intersecting famous trends. Changes in behavior on users are also picked up using machine learning. This includes variation in the interval of tweets, number of tweets in a particular time and sources of tweets. In addition, URL features such as duplicate domain names and duplicate URLs are also closely monitored. Twitter content entropy features track content of the users signaling dissimilarity or similarity between tweets. On the other hand, profile features help keep track of one’s profile description language and flag off any dissimilarities.

![Figure 2. ML Based Spam Detection Techniques.](image2)
Once all this information is collected from a malicious Twitter user, it is subjected to supervised machine learning algorithms for analysis, which in turn identifies suspected spam accounts. Currently, Twitter uses new spam detection features to outgun evasive spammers. This includes graph-based features, timing-based features, automation features among others. These features are hard to evade since they are requiring a lot of money, resources and time for a spammer to effectively evade.

4. Experimentation and Results

Some spam tweets can be easily identified because they include links to an external URL. In some cases, the links are transparent (e.g., http://www.nomoredatingpigs.com); In others, they are masked by URL shortens (e.g., http://www.bit.ly/KLYbHo), and the burden is placed on host sites like http://www.bitly.com to help detect illegitimate links. Other patterns we have observed spammers exhibit are more than one hashtag on disparate topics, letter number patterns in usernames, and suggestive keywords (e.g., “naked”, “girls”, “webcam”). A simple algorithm to detect spam in #robotpickuplines is based on these properties:

1. searches for URLs;
2. username pattern matches; and,
3. keyword detection.

Manually coded 300 randomly sampled tweets from #robotpickuplines as spam or not spam and ran this algorithm on the set. The given algorithm matched 91 percent of the time, with 27 missed spam tweets and 12 false positives. So we can use this baseline algorithm to mark all the tweets in #robotpickuplines. In an attempt to counter this and encourage more users, Twitter makes use of various spam detection methods. Twitter has upgraded their systems to ensure easy identification of spam messages. Spammers who use spam tweets can be easily flagged since they include links to external URLs as shown in the figure 3.

![Figure 3. Spam Detection Based on URL.](image-url)

The study notes that according to the Twitter policy, there are several indicators of a spam profile, which are the metrics, used in determining whether one is a spammer or not. Following a large number of Twitter users in a short period and posting tweets comprising of links or using popular trending hashtags to post unrelated content. Regularly posting other people’s tweets as yours is also used as a metric to detect a spammer. This research examined behavioral patterns among spam accounts.

Figure 4 shows traffic over the first 24 hours of use of the #robotpickuplines hashtag. The meme started at 11am CST and spiked around 3pm when it became a trending topic. It dropped around 4am and picked up again, although less heavily, the next day. Figure 5 shows the spam tweets using #robotpickuplines. We find that 14 percent of tweets are spam, and spam lags slightly behind the meme trend. Spam picks up about five hours after @grantimahara’s first post and decays quickly. Figure 5 has been scaled up to compare to figure 4, the raw number of spam tweets is actually 14 percent of legitimate accounts for this hashtag. The uptake slightly lags the meme itself. The drop–spike–drop pattern from 2–4am is unexpected; larger and more case studies may reveal broader trends.
4.1 Useful Findings and Discussion

Finding 1: Does age of account differ between spammers and legitimate users?

We anticipated that spam accounts would be newer than legitimate accounts because spammers can easily drop used accounts and create new ones; however, the results show that they were not significantly different (legitimate: mean=258 days, standard deviation=170 days; spam: mean=269 days, standard deviation=128). The distribution of both spammers and legitimate users by age of accounts showed that most accounts were created 100–200 days ago, which maps to a date created between February–April 2009 (when Twitter was heavily publicized by celebrities and the media) (see Figures 6 and 7).

Finding 2: Do spammers tweet more frequently than legitimate users?

With the calculated average distribution of tweets for each user by subtracting the date of first tweet from the date of last tweet and divided by total number of tweets, we can find the average
number of tweets per day was higher among spam accounts than legitimate accounts (legitimate: mean=6.7; spam: mean=8.66). The study also compared retweet and @reply behavior from legitimate accounts versus spammers. Results showed that 19 percent of legitimate tweets and 21 percent of spam tweets contained other hashtags within the set of #robotpickuplines tweets. Use of @replies was slightly higher with 26 percent legitimate uses and 24.8 percent spam uses. A chi-square test was performed and there was no significant difference in retweets or replies between spammers and legitimate users, but there was significant different in number of tweets $\chi^2(1,n=300)=4.464, \rho<0.05$, and use of hashtags $\chi^2(1,n=300)=3.847, \rho<0.05$.

**Finding 3: Do spammers have more friends than followers?**

This paper then examined structural properties of the network and hypothesized that follower–to–friend ratio would be higher for legitimate accounts than for spammers because spam bots may auto-follow Twitter users en masse. The calculations show the ratio to be not significantly different (1.38 for legitimate, 1.12 for spammers). However, the total number of followers and friends for spammers was three times that of legitimate users.(see Figures 8 and 9).

![Figure 6. Average number of spam follower and friends.](image)

![Figure 7. Number of legitimate follower and friends.](image)

**Finding 4: Are spammers clustered?**

This research also examined the 2nd degree network of #robotpickuplines users for overlapping links. It was must to examine two questions: 1) are there local clusters of users who might have learned about the meme from one another? and, 2) are there local clusters of spammers? (e.g., do spammers follow one another to boost their follower count?). For both spammers and legitimate users, the result shows some overlap in follower and friend ties; 3,236 accounts (out of 8,616) followed one or more other people who tweeted about #robotpickuplines. These accounts had an average out-degree of 2.28. The out-degree distribution revealed that 1,718 followed only one other user who tweeted #robotpickuplines and 10 users had an out-degree of greater than 10. There was little variation between spammers and legitimate users in the 1st degree network distribution. However, the spam graph is relatively small; larger sample sizes might reveal differences in network structures.
5. Conclusion

In conclusion, spamming is a problem facing social networks and needs keen attention. This paper has made efforts towards the problem of spam in OSNs (Online Social Networks). The study is primarily focused in the area of spam detection in twitter. This research has come up with approaches, which can be applied to help curb spamming in Twitter. Machine-learning Deep-learning is the secret to tackling this spam problem since it has a wide scope. Using these learning techniques, Twitter has been able to weed out unscrupulous twitter users. The future of spam detection in Twitter is mainly based on AI / ML techniques. The technology will continue to develop further helping to flush out spammers and improve Twitter space. Spam detection has become an integral part of Twitter in providing security to twitter users against cyber criminals and other malicious entities.

References

[1] Liu, S., Wang, Y., Zhang, J., Chen, C., & Xiang, Y. (2017). Addressing the class imbalance problem in twitter spam detection using ensemble learning. *Computers & Security*, 69, 35-49.
[2] Sun, N., Lin, G., Qiu, J., & Rimba, P. (2020). Near real-time twitter spam detection with machine learning techniques. *International Journal of Computers and Applications*, 1-11.
[3] Wu, T., Liu, S., Zhang, J., & Xiang, Y. (2017, January). Twitter spam detection based on deep learning. In *Proceedings of the australasian computer science week multiconference* (pp. 1-8).
[4] Chen, C., Wang, Y., Zhang, J., Xiang, Y., Zhou, W., & Min, G. (2016). Statistical features-based real-time detection of drifted twitter spam. *IEEE Transactions on Information Forensics and Security*, 12(4), 914-925.
[5] Sedhai, S., & Sun, A. (2017). Semi-supervised spam detection in Twitter stream. *Transactions on Computational Social Systems*, 5(1), 169-175. Spam detection on social networks using cost-sensitive feature selection and ensemble-based regularized deep neural networks. *Neural Computing and Applications*, 1-19.
[6] Gupta, Himank, Mohd Saalim Jamal, Sreekanth Madisetty, and Maunendra Sankar Desarkar. "A framework for real-time spam detection in Twitter." In *2018 10th International Conference on Communication Systems & Networks (COMSNETS)*, pp. 380-383. IEEE, 2018.
[7] Alsaaffar, D., Alfahhad, A., Alqhtani, B., Alamri, L., Alansari, S., Alqhtani, N., & Alboaneen, D. A. (2019, October). Machine and deep learning algorithms for Twitter spam detection. In *International Conference on Advanced Intelligent Systems and Informatics* (pp. 483-491). Springer, Cham.
[8] Jain, G., Sharma, M., & Agarwal, B. (2018). Optimizing semantic LSTM for spam detection. *International Journal of Information Technology*, 11(2), 239-250. Detecting the Spread of Online Fake News using Natural Language Processing and Boosting Technique (Doctoral dissertation, Dublin, National College of Ireland).
[9] Alom, Z., Carminati, B., & Ferrari, E. (2020). A deep learning model for Twitter spam detection. *Online Social Networks and Media*, 18, 100079.
[10] Ahmad, T., Ahmed, S. U., Ali, S. O., & Khan, R. (2020). Beginning with exploring the way for rumor free social networks. *Journal of Statistics and Management Systems*, 23(2), 231-238.
[11] Murugan, N. S., & Devi, G. U. (2019). Feature extraction using LR-PCA hybridization on twitter data and classification accuracy using machine learning algorithms. *Cluster Computing*, 22(6), 13965-13974. Measuring social media influencer index-insights from Facebook, Twitter and Instagram. *Journal of Retailing and Consumer Services*, 49, 86-101.
[12] Alorini, D., & Rawat, D. B. (2019, February). Automatic spam detection on gulf dialectical Arabic Tweets. In *2019 International Conference on Computing, Networking and Communications (ICNC)* (pp. 448-452).
[13] Herzallah, W., Faris, H., & Adwan, O. (2018). Feature engineering for detecting spammers on Twitter: Modelling and analysis. *Journal of Information Science*, 44(2), 230-247.
[14] Aslan, Ç. B., Sağlam, R. B., & Li, S. (2018, July). Automatic detection of cyber security related accounts on online social networks: Twitter as an example. In Proceedings of the 9th International Conference on Social Media and Society (pp. 236-240).

[15] Ameen, A. K., & Kaya, B. (2017). Detecting spammers in twitter network. International Journal of Applied Mathematics Electronics and Computers, 5(4), 71-75.

[16] Wardani, D. W., & Wardhani, Y. (2018). Detecting Spammers on Twitter by Identifying User Behavior and Tweet-Based Features. Journal of Telecommunication, Electronic and Computer Engineering (JTEC), 10(2-4), 81-84.

[17] Aswani, R., Kar, A. K., & Ilavarasan, P. V. (2018). Detection of spammers in twitter marketing: a hybrid approach using social media analytics and bio inspired computing. Information Systems Frontiers, 20(3), 515-530.

[18] Lee, S., & Kim, J. (2013). Warningbird: A near real-time detection system for suspicious urls in twitter stream. Transactions on Dependable and Secure Computing, 10(3), 183-195. Nadeem, S. K., Rao, N. K., Reddy, V. S., & Nikhil, K. SPAMMER DETECTION AND FAKE USER IDENTIFICATION ON SOCIAL NETWORKS.

[19] Lee, K., Eoff, B. D., & Caverlee, J. (2011, July). Seven months with the devils: A long-term study of content polluters on twitter. In Fifth international AAAI conference on weblogs and social media.

[20] Vanmathi, C., & Mangayarkarasi, R. (2020). An Analysis of Machine Learning Approach for Detecting Automated Spammer in Twitter. In Advances in Distributed Computing and Machine Learning (pp. 443-451). Springer, Singapore.