Detection of Real Time Spam Tweet on Twitter

Saima Sayed
Information TechnologyVidyavardhini's College
of Engineering and TechnologyVasai, Palghar

Shweta Sawant
Information TechnologyVidyavardhini's College
of Engineering and TechnologyVasai, Palghar

Kasturi Redkar
Information TechnologyVidyavardhini's College
of Engineering and TechnologyVasai, Palghar

Swati Varma
Asst. prof Dept of Information Technology
Vidyavardhini's College
of Engineering and TechnologyVasai, Palghar

***Abstract*—Social Network web channels such as Facebook, Twitter, Instagram has bagged a massive demand in today's time. Twitter is one such platform that allows individuals to express their view on certain topics and get updates about the current trending topics. But along with this increasing popularity the number of spammers has also increased. Today one out of every**

300 messages and one out of every 31 tweets is estimated tobe spam. Spammers find these online media an easy catch to catch users in their malicious activities by spam messages. Their motive is to steal important information from users, exploit their privacy or antagonize the users. In this paper, we are developing a Twitter Based Spam Detection Model which will classify the tweet as spam or not using various features and classifiers.

Keywords*—*Social networks, micro-blogging twitter, classification, fake user detection, machine learning, spammer's identifi- cation.

I. INTRODUCTION

Social spam is progressively influencing person to person communication sites, for example, Facebook, Pinterest and Twitter. As indicated by an investigation by the Web based life security firm, online networking stages encountered a massive development of social spam during the main portion of 2013.Online Social Platforms such as Twitter, give users the authority irrespective of their characteristic to indepen- dently create tweets and also devour great amount of data. According to survey while the amount of data generated is being utilized by individuals and organizations to gain competitive advantage, an essential part of data is generatedby spam or fake users. The fast development in the bulk of worldwide spam is relied upon to bargain research works that utilization online media information, accordingly address-ing information validity.One of the most prominent informal communication administrations, Twitter, has distributed their meaning of spamming as a component of their "The Twitter Rules" and gave a few strategies to clients to report spam, for example, tweeting "@spam @username" where @username will be accounted for as a spammer.

Our work on social spam is roused by the underlying endeavors at collecting a Twitter corpus around a particular point with a lot of predefined catchphrases. This prompted the recognizable proof of a lot of spam inside those datasets. The way that specific points are drifting and consequently many are following its substance urges spammers to infuse their spam tweets utilizing the catchphrases related with these themes to boost the perceivability of their tweets. These tweets produce alot of commotion both to end clients who follow the point just as to instruments that mine Twitter information. To distinguish the user as spam or not we are going to implement it by accessing the tweets from twitter. After extracting the tweets we will be processing them. Once the tweets are processed we will be performing feature extraction followed by feature matching. Once these steps are done we will be using different classifiers to give us accurate results of spam accounts. The results are evaluated and compared which will let us know which classifier yields the best result.

II. RELATED WORK

Distinctive way work has been done in separating spam what's more, spam profiles in Twitter or other casual as- sociations. Duty in each spam disclosure investigate follow same path as referenced in following decrees. Either make their own dataset or used publically available dataset for spam or non-spam game plan [5]. Each paper gives a significant examination on different kind of extraction of features. Au- thentic selection of features makes an unbelievable impact on recognizable proof model. A significant assessment is done on authentic getting of classifier that fits the area criteria and yield higher throughput.

In the accompanying, we present the proposition for rec- ognizing spammers' profiles and spam tweets, which are based on the client's profile and practices. Most of past

examinations [1], [3], [6], [7], [8], [11] depend on client conduct and tweet content-based spam discovery. Alom et al.
[1] utilized 6 AI classifiers and 12 features for distinguishing spammers on Twitter. Lee et al. [6] utilized 10 AI classifiers and two extraordinary informational indexes. Ameen et al. [7] utilized four AI order calculations and 13 content-based traits. Likewise, Ala'M et al. [8] utilized four AI classifiers and the absolute generally normal client based and content-based highlights for recognizing spammers on Twitter. Benevento et al. [11] thought about two methodologies for identifying spam profiles and spam tweets. At first, they assembled their model to differentiate spam profiles dependent on user based highlights. They considered 23 user based traits, for example number of devotees/followings, number of tweets, age of a record, and so forth. By utilizing the SVM classifier their work accomplishes 84.5 percent exactness. At that point, they utilized both user based furthermore, content-based highlights to characterize the tweets into spam what's more, non-spam classifications, by accomplishing 87.6 percent precision. Hai Wang et al. [12] utilized a social chart model, by utilizing on the followings and adherents connections. They removed from Twitter around 25K clients, and 20 late tweets for every client, alongside 49M companions/supporters relationship. To survey the discovery strategy, they utilized four distinct classifiers (i.e., DT, NN, SVM, NB) to group clients into spammers and authentic clients. In the tests, Naive Bayesian classifier gives the best execution: 91.7 percent accuracy and 91.7 percent F1- score. Wang et al. [13] concentrated on spam tweets location as opposed to recognizing spammers accounts. They utilized two hand-labelled informational indexes (i.e., Social honeypot and 1KS-10KN) and four capabilities, i.e., client based, content- based, n-gram, and feeling highlights.

Some crossover approaches, as Herzallah et al. [10], utilizedclient practices, chart based and tweet content-based highlightsto recognize spammers on Twitter. They utilized well known AI calculations for characterizing clients into spammers and non-spammers classifications. Sing et al. [9] utilized threecapabilities, to be specific trust score, content-based and user based highlights and four AI order calculations for grouping the clients into spammers and non-spammers.

III. METHODOLOGY

The block diagram of Proposed System is shown in Fig. 1A detailed explanation of flowchart is given below:

- The proposed model uses publically available dataset for Training Testing purpose which consists of labelled spam and non-spam tweets or manual labelling of tweets are done with class spam and non-spam. Then we crawl Twitter using its streaming API to collect the data.
- Pre-processing of this data is done for cleaning the data and make it appropriate for machine learning models
- In the next step features are extracted from the dataset. Various types of features can be used in spam account detection. Not all features are useful. Some of the featuresselected are mentioned in section 5.2. Features that show more effectiveness in yielding correct results are selected for spam account detection.

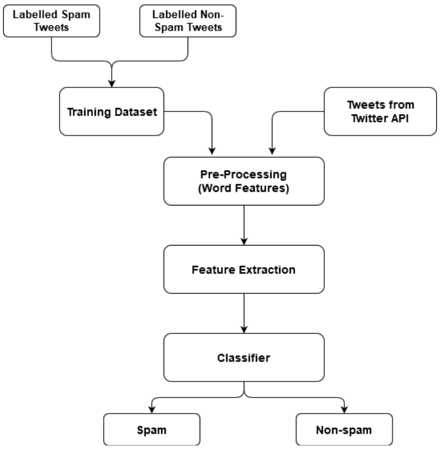

Fig. 1. Block diagram of Spam Detection Model

- Machine Learning based detection models are trained with labeled samples and then tested to identify classesof particular data instances.
- Finally detection models are evaluated with evaluation parameters like accuracy, detection rate, true positive,false negative, recall, precision, f-measures, etc.

IV. PROJECT IMPLEMENTATION

A. Pre-Processing

Natural Language Processing: Natural Language Processing (NLP) is a method which empowers a machine to deal with natural language (like English) and do all the things that a human can do.

1) *Tokenization*: The goal behind tokenization is to delete all marks of punctuation such as commas, full stop, hyphen, and parts. The process of dividing the whole text into distinct tokens is done, which helps in easy traversing of words inthe document.

2) *Stop word removal*: The motivation behind this process is utilized to wipe out conjunction, prepositions, articles and other frequent words such as adverbs, verbs and adjectives from textual data. The reason to eliminate these words isthat these words are normal and once in a while contain any significant data. Thus, it decreases textual information and increases device efficiency.

3) *Stemming*: Stemming is being used to simplify terms to their source words. These keywords are compared with pre- defined set of spam words. If the words match, then the useris regarded as spam. At this stage, if the user is not found as spam, then the third technique of Machine Learning is used.

B. Content Based Features

1) Mention Ratio: Users of the Twitter social media network can be tagged by "@" symbol followed by twitter handler. Spammers of the network can misuse this feature.The spammers motivate and tempt the benign users to know the sender of the message. The mention ratio for the user is calculated as the ratio between number of mentions in tweets and number of tweets posted by users. Naturally the benign users mention ratio is low compared to spammers.

Mention ratio = no. of mentions in tweets / no. of tweetsby the user

2) Ratio of URL: In Twitter, users generally post their ideas, opinions about a specific topic and share articles through tweets. The tweets include URLs, these refer source pages that contain detailed information. Some of the users include high amount of URLs into tweets continuously, sowe can suspect them as spammers. The Ratio of URL for a user is the ratio of number of URLs used in his tweetsto total number of tweets posted by that user. Generally, large numbers of URLs are used by spammers in the tweetsto share their intention to users. Spammers use enormous number of URLs whereas legitimate users use less number of URLs in tweets. The spammers URL ratio is nearer to one or more than one whereas for benign users the URL ratio is very small or closer to zero.

URL ratio = no. of urls / total no. of tweets

3) Unique mention ratio: Generally benign users contact with friends and colleagues and at the time of sending tweets they can use this group of the people or set of the people regularly but spammers tag the unknown persons randomly within their tweets. Generally, the spammers unique mention ratio is very high and low for genuine users.

4) Unique URL ratio: Enormous number of URLs used by the spammers in the tweets to fulfill their intention but at the same time some of the spammers use the same URL many times for the same user. The genuine user sees the same URL many times and gets tempted and ends up clicking and traversing to malicious sites. The unique URL ratio is the ratio of number of unique URLs to number of URLs used in the tweets.

Unique URL ratio = no. of unique urls / total no. of urls

5) Hashtag ratio: To group the tweets related to specific topics, Hashtags are used. A group is created by hashtag to discuss specific topics. Top trending hashtags regularly display on a user's wall. These trending hashtags are hijacked by spammers and they inject them into their tweets. Whenever genuine users search for these trending hashtags, tweets by the spammers are also shown in the search result.

C. User Based Features

1) Number of Tweets: The tweets posted by a user regularly tell us about the activity frequency of the user on a daily basis.

The Activity frequency of Spammer is high as compared tolegitimate users.

2) Number of Following (FI): Since spammers tend to follow too many legitimate accounts in order to attract attention, the number of following is expected to be high compared to legitimate users.

3) Number of Followers (FE): Spammers follow many legitimate users but since they are not connected to the legitimate users, therefore they do not get followed back, asa result the no. of followers of spammers are very less as compared to legitimate users.

4) Number of likes: Spammers tweets usually consist of irrelevant content because of which it doesn't get likes equalto a legitimate user.

5) Number of retweets: Since spammer's tweets are spontaneous, the quantity of retweets for their tweets are required to be less contrasted with authentic clients.

6) Age of Account (AU): According to study, spammers don't use the same account for a prolonged time and tend to change accounts for creating spams and remaining unnoticed. Since they continuously keep changing accounts, the age of their account is less.

7) Reputation of User: It is a feature that depicts overall impact of the user.It is given as the ratio of No. of followersto the total of users he is in contact with FE / (FI + FE).

D. Spam Detection Model

1) Naive Bayes: It is an efficient classifier that is used to classify the text message as spam message or ham message. The Naive Bayesian classifier is based on probability theory. This model is used because it gives good performance and requires less computational time for training the model. The main assumption of this algorithm is that the features of a dataset are independent, it means that the probability of one attribute does not affect the probability of the other. This classifier is used to classify the tweet based on posterior probability of the tweets belonging to different classes.

2) Random Forest: Random Forest is a very flexible machine learning classifier that consists of a collection of tree structured classifiers. It randomly selects the features to construct a collection of decision trees. As we realise that a forest is composed of trees, more trees mean stronger forests. Similarly, a random forest algorithm produces decision trees on data samples and eventually selects the best arrangement through methods for casting a vote on each of them after the prediction. It is a gathering strategy which is superior toa single decision tree since it diminishes the over-fitting by averaging the outcome. Due to the simplicity it provides, it can be used for classification as well as regression tasks.

3) SVM: Support Vector Machine is a supervised learning algorithm which can be used for both classification as well as regression. The heart of SVM is the linear separating hyperplane. SVM also hold up kernel method which is also known as kernel SVM, allows us to hoist non-linearity. It trains on a labelled data. It studies the labelled data and classifies the new data acording to what it learned in the training phase. The advantage of SVM is it provides high dimensionality, memory efficiency and versatility.

4) K-Nearest Neighbors(KNN): KNN can be used to solve many problems, in classification for example we can classify a new point just by examining the class of its nearest neighbors. We can also use KNN to find the most similar documents to a given document for plagiarism, finding mirrors, etc. In recommender systems we can use KNN to find the itemsthat are most similar to an item a user hasn't reviewed and then calculate if the user will like it or not. We can use it in clustering algorithms and there are many many more applications. We need some form of metric distance for this, there are several options the most common is the traditional euclidean distance but you can use Manhattan, Hamming, Jaccard and Levenshtein. A naive solution is to linearly search all the original points computing the distance to the query point.

V. RESULTS AND DISCUSSION

If there is a tweet 't' and the spam class 'S'. The output of the classifier is whether 't' belongs to 'S' or not. A common way to evaluate the classifier's performance is to use True Positives, False Positives, False Positives, False Negatives.

TABLE I
EVALUATION METRICS

Class	**Predicted Spam**	**Predicted Non-Spam**
True-Spam	True Positive	False Negative
Non-Spam	False Positive	True Negative

In Evaluation, we present the details for evaluation of our proposed model for detecting spam in tweets. Classification, Association and Clustering algorithms are used for mining the unseen patterns in huge amounts of data. To evaluate the proposed approach, we used standard metrics called precision,recall and F-measure.

1) Precision: Precision is a measure to determine rate of False Positive with respect to the Total Positive Predicted. Itis a measure which determines the classifiers exactness. A low precision indicates that there are a large number of false positives.

2) Recall: Recall is a measure to determine rate of False Negative with respect to the Total Actual Posotive. Recall is used as a measure of classifier completeness. A low recall indicates that there are many false negatives.

3) F measure: This metric, measure the association betweenprecision and recall.

4) Accuracy: Accuracy is used for evaluating the classification models. Accuracy is the fraction of predictions our model got right.

We Implemented a series of experiments with different classification models and assessed them using various metrics, as discussed above in Evaluation Metrics. These metrics are calculated for spammer and non-spammer seperately and then we compare these values with the value of metrics obtained by other classifiers. The below figure shows the classification report of each classifier.

TABLE II
COMPARISON OF CLASSIFIERS

	Naive Bayes		**SVM**		**Random Forest**		**KNeighbours**	
	spam	NS	spam	NS	spam	NS	spam	NS
Precision	0.28	0.79	0.82	0.96	0.88	0.96	0.86	0.96
Recall	0.29	0.79	0.84	0.95	0.88	0.96	0.84	0.97
F1-score	0.29	0.79	0.83	0.95	0.88	0.96	0.89	0.96
Accuracy	0.67 = 67%		0.92 = 92%		0.93 = 93%		0.94 = 94%	

From the above figure, it is clear that KNeighbours Classifier gives us highest values of Precision, Recall, F1-score and Accuracy.

VI. CONCLUSION

Spammers are the problem in any online social networking sites. Once a spammer is detected it is easy to suspend his/her account or block their IP address. This research deals with the study of spam classification techniques in twitter. Twitter API is developed to collect real data sets from the information which is publically available on Twitter. The approach of Nat- ural Language processing and Machine Learning techniques, can successfully do the classification. Also, here we identify different sets of expressions, tweets, words and other features which can show that a user is spam or legitimate.

REFERENCES

[1] Z. Alom, B. Carminati, E. Ferrari, "Detecting Spam Accounts on Twitter", in ACM International Conference on Advances in Social Networks Analysis and Mining. IEEE, 2018.

[2] S. Gheewala, R. Patel, “Machine Learning Based Twitter Spam Account Detection: A Review”, in proceedings of Second Interna- tional Conference on Computing Methodologies and Communica- tion.IEEE,2018.

[3] F. Masood, G. Ammad, A. Almogren, A. Abbas, H. A. Khattak, I.

[4] U. Din, M. Guizani, M. Zuair, “Spammer Detection and Fake User Identification on Social Networks”, IEEE,2019.

[5] K. S. Adewole, N. B. Anuar, A. Kamsin, K. D. Varathan and S. A. Razak, “Malicious accounts: Dark of the social networks”, Elsevier,2017, pp. 41-67.

[6] M. Mateen, M. Aleem, M. A. Iqbal and M. A. Islam, “A Hybrid Approach for Spam Detection for Twitter”, IEEE, 2017, pp. 466- 471.

[7] K. Lee, J. Caverlee, and S. Webb, “Uncovering social spammers: social honeypots+ machine learning,” in Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval. ACM, 2010, pp. 435–442.

[8] A. K. Ameen and B. Kaya, “Detecting spammers in twitter network,” International Journal of Applied Mathematics, Electronics and Computers, vol. 5, no. 4, pp. 71–75, 2017.

[9] A.-Z. Ala’M, H. Faris et al., “Spam profile detection in social networks based on public features,” in Information and Commu- nication Systems (ICICS), 2017 8th International Conference on. IEEE, 2017, pp. 130– 135.

[10] M. Singh, D. Bansal, and S. Sofat, “Who is who on twit- ter– spammer, fake or compromised account? a tool to reveal true identity in real-time,” Cybernetics and Systems, pp. 1–25, 2018.

[11] W. Herzallah, H. Faris, and O. Adwan, “Feature engineering for detecting spammers on twitter: Modelling and analysis,” Journal of Information Science, p. 0165551516684296, 2017.

[12] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida, “Detecting spammers on twitter,” in Collaboration, electronic messaging, anti- abuse and spam conference (CEAS), vol. 6, no. 2010, 2010, p. 12.

[13] A. H. Wang, “Don’t follow me: Spam detection in twitter,” in Security and cryptography (SECRYPT), proceedings of the 2010 international conference on. IEEE, 2010, pp. 1–10.