Spam Diffusion in Social Networking Media using Latent Dirichlet Allocation

Poonam Tanwar, Priyanka

***Abstract** Like web spam has been a major threat to almost every aspect of the current World Wide Web, similarly social spam especially in information diffusion has led a serious threat to the utilities of online social media. To combat this challenge the significance and impact of such entities and content should be analyzed critically. In order to address this issue, this work usedTwitter as a case study and modeled the contents of information through topic modeling and coupled it with the user oriented feature to deal it with a good accuracy. Latent Dirichlet Allocation (LDA) a widely used topic modeling technique is applied to capture the latent topics from the tweets' documents. The major contribution of this work is twofold: constructing the dataset which serves as the ground-truth for analyzing the diffusion dynamics of spam/non-spam information and analyzing the effects of topics over the diffusibility. Exhaustive experiments clearly reveal the variation in topics shared by the spam and non-spam tweets. The rise in popularity of online social networks, not only attracts legitimate users but also the spammers. Legitimate users use the services of OSNs for a good purpose i.e., maintaining the relations with friends/colleagues, sharing the information of interest, increasing the reach of their business through advertisings.*

Keywords: Spam detection ,SVM,LDA , Social Networking, Twitter

I. INTRODUCTION

Online Social Networks (OSNs), like Twitter and Facebook, have become increasingly popular in the last few years. Internet users spend more hours on social network sites than any other website [1]. In addition social network sites have become the main news source for around 30% of population according to the survey carried out by Pew Research Center [2]. Moreover, the social networking apps in smart phones make users' access to such sites become ubiquitous. These large social networks have attracted many researchers' interest.

Despite the interest of researchers the rich information in OSNs has also attracted the attention of spammer.

The aim of these attacks either misleads public opinion, Spread false information, or disrupt the conversations of legitimate users.

This false, irrelevant or unsolicited messages are sent over the internet are called spam. Spammer can send spam messages to a large number of users for a variety of use cases such as advertising, phishing, spreading malware, etc. [3] ."Spam" is invented by Monty Python in year 1970. Python explain in the sketch where two customers are lowered by wires into a café and try to order from a menu, which includes spam in almost every dish. Social spam is unwanted spam content appearing on social networking services, Social bookmarking site and any website with user generated content like comments, chat etc. It can be manifested in many ways, including bulk messages, malicious link, fraudulent reviews, fake friends, and personal information.

Spam is divided into four categories with different behaviors. [4]

1. Email spam
2. *Web spam (Cloaking and redirects)*
3. *Opinion spam (Hidden text and keyword stuffing)*
4. *Social spam (Hacked websites and malware)*

Electronic spamming is to send an electronic message and to send an unwanted message (spam), especially for advertising purpose, and for spreading the news, and send those messages repeatedly on the same site. The most widely recognized form of the spam is email spam. Opinions, comments, tweets, are the informative part of any post, tweet, and reviews. Spam refers to any extraneous or voluntary information which is attached to the tweets for advertisement, promotion, for information spread, or even for gaining the financial profit. Two main types of spam are Cancellable Usenet and Email where in case of first one a single message will be sent to 20 or more Usenet newsgroups and in case of later one spam targets individual users with direct mail messages. Email spam lists are often created by scanning Usenet postings, stealing Internet mailing lists, or searching the Web for addresses. Email spam typically cost users money out-of-pocket to receive. Opinion has been categories in to various

types of spam opinion like Email spams, Web spam, Social spam and Opinion spam are shown in table 1.1. The Social spam has various categories that are Bulk message spam and fraudulent review spam, Blog spam which are more harmful than the original spam. Deceptive opinion spam is the untruthful review that is positive opinion spam (hyper spam) and negative spam can (defaming spam).Whereas Disruptive opinion reviews are advertisement, announcements and random text types.

A. Social Spam

Social spam-It is a spam when some undesirable contents are displayed on the social /popular networking site using user generated information / contents like videos, advertisements, pictures, comments, smiles, chat, etc.

It can be used in many ways, including bulk messaging, job & other lucrative offers, advertisements, profanity, liking, insult, fraudulent reviews, hate speech, malicious links, false friends on social networking websites[5].

Revised Manuscript Received on October 30, 2019.
* Correspondence Author

Poonam Tanwar*, Manav Rachna International Institute of Research and Studies, Faridabad, Haryana, India. Email: Poonam.tanwar@rediffmail.com

Ms. Priyanka, Manav Rachna International Institute of Research and Studies, Faridabad Haryana, India. Email: priyanka17_engg@rediffmail.com

SPAM				
Email Spam	Web Spam	Opinion Spam		Social Spam
		Deceptive opinion	Disruptive opinion	Bulk message
		Hyper Spam	Advertisement Spam	Fraudulent reviews
		Defaming Spam	Announcement Spam	Blog spam
			Random Text Spam	

Table 1.1 Categorization of spam [4]

II. BACKGROUND STUDY

A research topic required good amount of literature survey to really make a good start of research work with effective evaluation of the work already done in similar field.Survey is a most important building block of the research process, that may content a research project in itself.Previous research papers or thesis the literature survey is an essential collection of previous work done and its outcomes. It is an essentially collection of scholarly papers & publications which includes the technological advancements and their uses in industry & society.

A. Spam classification in social media

OSN's are flooded with spam messages, videos and pictures. Analyzing and understanding of spam pattern should be carefully dealt before detection. The objective of this section are classifications of spam in OSNs

B. Analyzing spam with blacklisted URLs

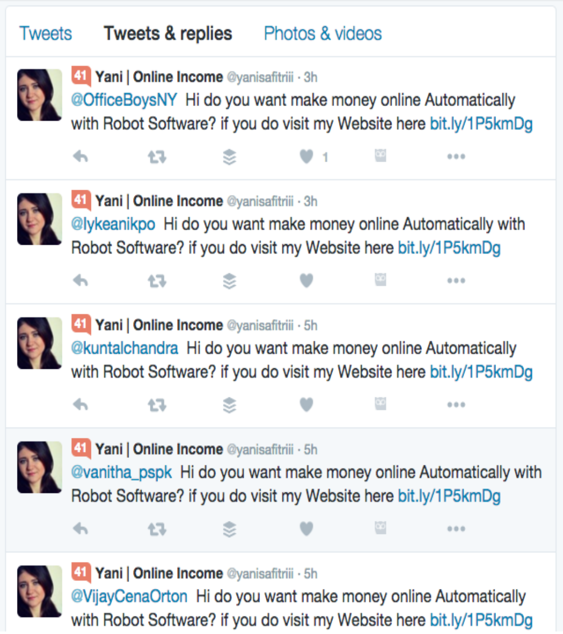
C. Grier et.al in 2010 analyzed 200 million public tweets from 25 million URLs, which were collected in short span of 30 days. After using three blacklists (Google Safe browsing, URIBL, and Joewein), 3 million tweets and 2 million URLs were reported as spam (5% were malware, phishing and 95% directing victims to11 scams), which is the 8% of all crawled unique URLs. 26% of URLs were detected as spam manually using available sample dataset, that represent the performance of blacklisting was moderate and there is huge scope for improvement.

C. Analyzing suspended accounts on Twitter

K.Thomas et.al in 2010 have worked on suspended accounts, a dataset of 1.8 billion tweets (sent by 32.9 million accounts in the period of seven months), with 80 million (from the 1.1 million accounts suspended by Twitter itself) are spam. approx. 3.3% accounts in the dataset were suspected by twitter. Researcher decided to validate sampled of 100 suspended accounts and they observed that most of these accounts are fake account instead of compromised ones.Twitter designed detection algorithm after deep examination and can only detect 37% of spam accounts, 77% spam accounts are suspended within immediately in a day of their first tweet and 92% spam accounts could only last for three days [6,14,15].

D. Characterizing spam campaigns in OSNs

H. Gao et.al. in 2010 has worked for Facebook spam issue. They retrieved about 187 million messages (specially post) from approximately 3.5 million end users from eight regional networks in Facebook.Initially they detected the users whoever spreading the spam messages and combined the post shared by the same user. All posts from the same user/URL has been clustered based on similarity index. On the basis of experiments, they found that 70.3% spam posts directing victims& general public to phishing sites and 35.1% spam posts lead to malware downloading of videos and pictures, which is much different to, with only 5% spam direct to phishing and malware [18][19][20]. Figure 2.1 is example of twitter spams for iTunes gift card, which motivate a number of victims get into their trap [18-20].



Tweets Tweets & replies Photos & videos

41 Yani | Online Income @yanisafitriii · 3h
@OfficeBoysNY Hi do you want make money online Automatically with Robot Software? if you do visit my Website here bit.ly/1P5kmDg
1

41 Yani | Online Income @yanisafitriii · 3h
@lykeanikpo Hi do you want make money online Automatically with Robot Software? if you do visit my Website here bit.ly/1P5kmDg

41 Yani | Online Income @yanisafitriii · 5h
@kuntalchandra Hi do you want make money online Automatically with Robot Software? if you do visit my Website here bit.ly/1P5kmDg

41 Yani | Online Income @yanisafitriii · 5h
@vanitha_pspk Hi do you want make money online Automatically with Robot Software? if you do visit my Website here bit.ly/1P5kmDg

41 Yani | Online Income @yanisafitriii · 5h
@VijayCenaOrton Hi do you want make money online Automatically with Robot Software? if you do visit my Website here bit.ly/1P5kmDg

Figure 2.1 Example of Twitter Spam

III. PROPOSED SYSTEM

Data Collection

This work aims to analyze the diffusion of spam/not spam information. We collected the IDs of spam/ham tweets from HSpam14. This dataset is generated by collecting the tweets on the trending topics. Twitter generally provides two APIs for data collection, REST API and streaming API. The data of past week is fetched through REST API, whereas streaming API is used for collecting the live tweets. The keywords on trending topics are passed to the streaming API and the tweets containing those key-words are filtered and stored.After collecting 14 million tweets, the annotation process is done through heuristic and finding the near duplicate clusters.

In heuristic, it is assumed that most popular hashtags are likely to contribute for spamming. The tweets containing popular hashtags are labelled as spam. HSpam14 dataset only provides the tweet IDs and their label as spam or not spam.

But to study the diffusion of spam/non spam information there is a need of tweets' text and user related properties like number of followers, followees, account age, and content posted. We overcome this limitation by going through the steps of algorithm 1, which takes the tweet IDs as input and return the tweets with complete information i.e., author ID, account creation date, tweet creation date, tweet text and the retweet count. In this work, we collected 10,000 tweets that matched some trending hashtags in six months' time and then conducted systematic annotation of the tweets being spam and ham (i.e., non-spam). We annotated dataset on the basis of HSpam14.

Our annotation process includes four major steps:

(i) Heuristic-based selection to search for tweets that are more likely to be spam,

(ii) Near-duplicate cluster based annotation to firstly group similar tweets into clusters and then label the clusters,

(iii) Reliable ham tweets detection to label tweets that are non-spam,

One major contribution of this work is the creation of HSpam14 dataset, which can be used for hashtag oriented spam research in tweets. Another contribution is the observations made from the preliminary analysis of the HSpam14 dataset.

A. Designing of model

In designing of topic modeling follow the step 3.2 and system architecture shown in Figure 3.1 with the tweet corpus then examine the topics shared by the spam and non-spam information;

we applied Latent Dirichlet allocation (LDA) technique. First, we built a corpus and a dictionary from the collected tweets, where corpus represents the occurrences of words for each document and dictionary contains ids for each word and each document. Next, we train the LDA model by using the created corpus and dictionary. LDA converts the document-terms matrix (corpus) into two matrices: document-topics matrix and topics - terms matrix. Initially, terms are randomly distributed to topics by using the gamma distribution. The probability density for the distribution is:

$$P(x)=x^{k-1}e^{-x/\theta}/\ \theta^{k}\rho(k) \qquad (1)$$

where k is number of topics to be extracted, is scale (1/number of topics), and is the gamma function. Then, this distribution is updated until the convergence point of LDA. The mean change between prior distribution and updated distribution is less than the given threshold than it is called as convergence point for LDA. Topics corresponding to the spam and non-spam tweets are returned by this technique, where each topic contains 30 most probable words, which we will analysis in the next chapter it includes the Implementation part.

Algorithm on tweet filtering and topic modeling

Input- List of extracted tweet.
Step 1 - Categorized them spam/ham
Step 2 -Tweet contain #free, #sail, #retweet, based on Hspam , labeled spam. Step 4 – input the spam tweet list
Step 5 – Chunk the given input into list of max.100 elements.
Step 6 – Pass each list to status_ lookup module.
Step 7 - Output of the lookup module.
Step 8 – output tweet object.

Topic modeling

Step 9 – apply tweet pooling techniques
Step 10 – apply LDA
1. The topic distribution in the given corpus is β_k.
2. The topic collection from each document in the corpus.
(a) Each topic has θ_d amount of distribution in each document.

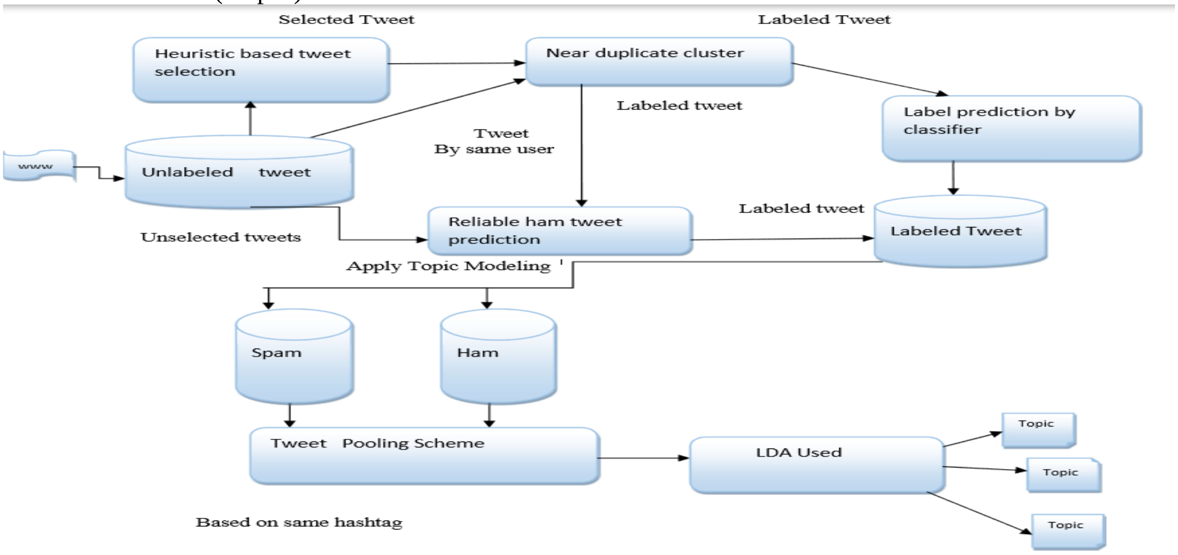


Fig 3.1 Architecture of spam detection and topic modeling in Twitter

(b) Then consider the each word in the corpus.

i. Then select the topics from the topic distribution, that is zn.

ii. Then Choose a particular word from that topic occur is wn and the topic distribution is $\beta_{z^{\wedge}n}$.

IV. RESULT ANALYSIS

Twitter offers a functionality of "retweeting" which empowers people to spread the information of their choice beyond the followers of original tweet's author. It is a key mechanism by which information gets diffused on Twitter. We used retweet count as a measure of diffusibility and divide the collection of spam and non-spam tweets into 9 categories. Figure 4.1 shows the percentage of tweets under each category. In the category "<=2", tweets received upto two retweets. It is observed in figure 4.1 that 69 percent of spam tweets fall under this category whereas non-spam tweets are only 41 percent. It shows that a major portion of spammy information do not get a high diffusion. But, it is clear from Figure that spammy information is capable of getting retweet even on the scale of thousands. So, there is a need of further analysis to get the best parameters contributing to the diffusibility of spammy information.

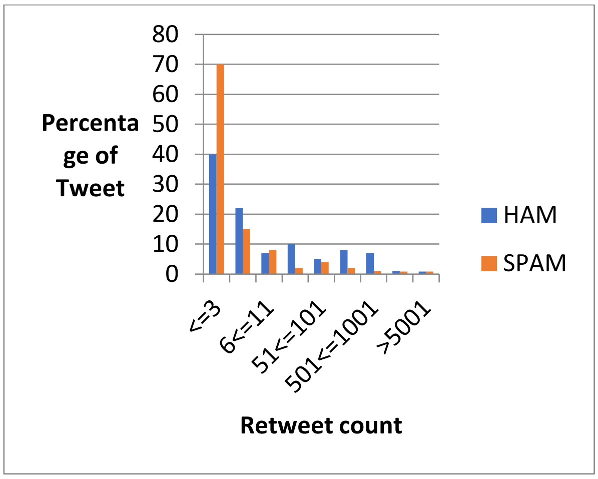


Figure 4.1: Diffusion of spam & ham tweet

Table 4.1 Dataset statistics

	Tweets	Hashtags	Mentions	URls
Ham	**7264**	**3557**	**2871**	**1658**
Spam	**3375**	**5319**	**1567**	**2486**

Result shown in the table 4.1 dataset statistics, we observed that spam tweets tend to have more number of URLs and hashtags than non-spam tweets. To examine the diffusion dynamics in a better way, in this section, we analyze spam/non-spam information through topic modelling.

V. CONCLUSION

Social spam has led a serious threat to the utilities of online social media. To combat this situation, construct a dataset which serves as the ground-truth for analyzing the variation spam/non spam information diffusion in the previous section. In modern age the Online Social Networks have changed the way of communication and information access. The popularity of these sites is the main cause of spam diffusion, because of the popularity the spammer also attract toward the social media. The motive of spammer is to spread the malicious contents and earn money or defame the reputation of others. Spammer mainly spread unwanted posts. Most of the receiver even doesn't know about these links, and they click on those link and become the victim of spammer. Such spam not only pollutes the platforms but also exploit users' critical information. To solve this issue this work is focused on the spam diffusion on the twitter and to analysis the pattern of the spammy contents. This work contains spam detection through twitter data analysis and also analyzes the spam through topic modeling and extract the latent topic from the corpus.

REFERENCES

1. Ajay Rastogi and Monica Malhotra "Article in Journal of Information & Knowledge Management "Sept 2017.
2. Mohammad Ahsan and Madhu Kumari,."Spam Diffusion in Twitter using Sentiment and Latent Dirichlet Allocation "
3. Surendra Sedhaiand AixinSu.," HSpam: A Collection of 14 Million Tweets for Hashtag-Oriented Spam Research".
4. Chao Chen," a Survey of Spam Detection Methods on Twitter [digrame]",International Journal of Advanced Computer Science and Applications.
5. Grier,C. et all,"Spam: the underground on140 characters or less". In Proceedings of the 17th ACM conference on Computer and communications security, CCS '10, pages 27{37, New York, NY, USA, 2010.ACM.
6. Thomas K.," Suspended accounts in retrospect:an analysis of twitter spam", Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference, IMC '11, pages 243- 258, New York, NY, USA, 2011. ACM.
7.]. Wei. Et all, "Exploring characteristics of suspended users and network stability on twitter", Social Network Analysis and Mining, 6(1), 1-18, 2016.
8. H. Gao, et all, "Detecting and characterizing social spam campaigns"., Proceedings of the 10th ACM SIGCOMM conference on Internet measurement, IMC- '10, pages 35- 47, New York, NY,
9. USA, 2010.
10. Priyanka rai and poonamtanwar," A propose system for opinion mining using Machine Learning Approach using NLP and classifiers ", Proceedings of the 12th INDIACom; INDIACom-2018;
11. H. Achrekar et all. "*Online social networks u trend tracker": A* novel sensory approach to predict u trends, editors, Biomedical Engineering Systems and Technologies, volume357 .
12. Castillo, C. et all, "Information credibility on twitter". In Proceedings of the 20th international conference on World wide web , pp. 675-684, ACM.
13. Benevenuto, F. et all,"Detecting spammers on twitter", electronic messaging, anti-abuse and spam conference (CEAS) (Vol. 6, No. 2010, p. 12).
14. Elenberg, E. R.et all, "Distributed estimation of graph 4-profiles". Proceedings of the 25th International Conference on World Wide Web, pp. 483-493, 2016, April.
15. Yang, Z. et all," Uncovering social network sybils in the wild", *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *8*(1), 2, 2014.

16. 0. Carley, K. M., “ORA: A toolkit for dynamic network analysis and visualization. In *Encyclopedia of social network analysis and mining* Springer”, New York, NY. , pp. 1219-1228, 2014.
17. P. Landwehr et all, “Using tweets to support disaster planning, warning and response. *Safety science”*, volume *90*,pp 33-47.
18. S . Cresci,., Di Pietro, R., M Petrocchi,.,A.Spognardi,., &M. Tesconi, “Fame for sale: “*efficient detection of fake Twitter followersDecision Support Systems”* , volume *80*, pp 56-71.
19. Yang, C., et all, “Analyzing spammers' social networks for fun and profit : a case study of cybercriminalecosystem system on twitter”, *Proceedings of the 21st international conference on World Wide Web* , pp. 71-80, ACM , 2012.
20. Mukherjee, A. et all ,” What yelp fake review filter might be doing?. In *ICWSM*, pp. 409-418, July 2013.
21. L. Backstrom and J. Leskovec.,” Supervised random walks”, “*Predicting and recommending links in social networks.”*CoRR, abs/1011.4071, 2010.