Spam Detection and Filtration using Data Mining for Social Networking Sites

Ritesh Kumar, Mayur Girnar, Archana Darwatkar, Shital Ghadage, Prof. G.S Navale

Abstract - We live in the era of social media where everybody is socially connected. But recently Online Social Network [3] is vulnerable to mass spam incidence as well as users data theft such as financial credential, user system data and user trends for exploitation purposes. Recent trend of spam incidents is causing a very serious threat to Social Networking World which has in turn become an important means of interaction and communication between socially online users. It is not only dangerous to the socially online users, but it also occupies large portion of the traffic on networks. Majority of current spam filters in use are based on the metadata, subject and content of emails and other social networking sites such as Facebook, twitter etc. Social Networking Services also provide great possibilities to take advantage of user identification and other social graph-dependent features to improve classification. In this paper, our system uses machine learning [1] approach for spam detection based on features extracted from social networks constructed from social networking site message metadata and logs. Email subject headers are used to verify spam email, spam on Social networking Sites is often accompanied by a wealth of data on the sender, user messages, posts, content can be used to build more relevant detection mechanisms. System uses these terminologies to choose features that best differentiate spammers from the legitimate users. In this technique system flag user system or message as spam and non-spam messages. Legitimacy credential are assigned to senders based on their possibility of being a legitimate user or spammer. Also, proposed System also explores various spam filtering techniques and possibilities.

Keywords: Machine Learning [1], SNS (Social Networking Sites) [2], OSN (Online Social Network) [3], tf-idf (Term Frequency-Inverse Document Frequency) [4].

I. INTRODUCTION

With advent of more online user on social media in public has made Social Networking Sites [2] most appropriate targets for spam and fraud, vulnerable to mass attacks. Earlier checking on email subject was done to check spam messages. Many of these techniques previously used to combat conventional email and web spam. Social Networking Sites [2] give facility to take advantage of user identification and other social graph-dependent features to improve classification. Majority of research has been carried out on public research available data from Social Network Sites, making it difficult up until now to measure the effect of private user data on algorithms for detecting site misuse. The System proposed to reconstruct spam messages into categories for classification rather than checking them individually.

Although categorization and identification has been done for offline spam analysis, we apply this technique to be used in the online spam detection problem with sufficiently high accuracy. Our proposed system undergoes a set of parsing algorithm that effectively distinguishes spam messages. It pins messages category as “spam” before they reach the targeted recipients, thus providing shield from various kinds of hacks and vulnerabilities.

II. BACKGROUND

Vast amount of online data and traffics requires new strategies for combating its usage and vulnerabilities. Earlier checking for spam message is done manually from data set or generic algorithms were used in case of text based spam checking. Earlier algorithm like Bayes theorem, regression analysis and keyword matching were used for uncovering patterns or spam, this algorithms were meant to handle small data set and are incompetent to handle large dataset.

Large data with properties like unstructured, size and complexity, direct analysis is next to impossible or is tiresome process. Using Machine learning [1] data processing technique, such as big data analysis, cluster analysis, decision trees and decision making, text mining and support document classification. We can easily disclose the uncovered pattern and spam messages. Data mining can be easily applied using these methods which help in discovering hidden patterns in large data sets and moreover analyzing association and trends.

III. RELATED WORK

Recently, Persistent Systems Pvt. Ltd, Pune has done work based on data mining (Text Classification) for the reality TV show “Satyamev Jayate”, for which they have taken the

data from social networking site (twitter). For this project they have analyzed the feedback on twitter for that show. Significant research has been done in studying spam detection using data mining approach.

The proposed system uses social networking sites as the example application. System can start monitoring of new posting activities in social media network. For each new instance, System first make prediction based on the algorithm, if it is non-deterministic, send the message for manual labeling using human interference.

Note. This demo uses Facebook as the example since it is currently the most popular social networking sites. However, the designed system can be easily generalized and applied to other social networking sites, such as Facebook, Quora and Twitter.

IV. PROPOSED METHODOLOGY

The algorithm which we are using for our system is based on the spam detection using data mining approach. In the first stage, we collect social media content (including text with time stamp) performs machine learning to build classified documents and identify spams. In the second stage, we monitor the activity and trend of user, make prediction on basis of algorithm and send flag or spam alert to client about detected spams and update the model.

The complete process is divided into six tasks explained as follows:

Task 1 First of all we have to extract and fetch the Facebook comments to our local machine.

Task 2 After storing comments we need to parse each comment.

Task 3 Next to parsing of data is the tokenizing of comments for proper structuring.

Task 4 After tokenization of the comments we need to check for its spam classification.

If it contains URL check for suspicious URLs (URL Features: Number of (.), special URLs etc.)

If it not contains URL check for suspicious word by comparing with spam keyword stored in database.

Task 5 Determine whether a comment on aspect is spam, no spam or neutral.

Task 6 Produce all messages expressed in document based on results of the above tasks.

Algorithm Steps:

1. Let IP = {S1, S2, S3… Sn}
2. Splitting:
   S = {V1, V2, V3… Vn, Ad1, Ad2, Ad3… Adn, Adj1, Adj2, Adj3, Adjn, Sw1, Sw2, Sw3, Swn }
   Where,
   V1, V2, V3… Vn = Set of verbs (walk, play, etc.)
   Ad1, Ad2, Ad3… Adn = Set of adverbs (very, extremely, etc.)
   Adj1, Adj2, Adj3… Adjn = Set of adjectives (good, bad, excellent, etc.)
   Sw1, Sw2, Sw3 …Swn = Set of Stock word (from, his, to etc.)
3. Let Sw = { Sw1, Sw2, Sw3 …Swm }
4. SF = {S}-{SW}
   Where SF = Set of verbs + Set of adverbs + Set of adjectives
5. Let each element of SF be W1, W2 and so on.
6. Let T = { T1, T2, T3… Tn }
   Where,
   T1 = {W1, Wd1}
   W1 = Dictionary word
   Wd1 = Weight (-1 and 1)
7. For each x of SF then assign weight form set T if and only if word found in set T else set 0.
8. Adding weight of each word in SF as
   $Spam = \sum_{i=0}^{m} weight(W(i))$
9. If Spam > 0, then spam of that sentence is positive.
10. If Spam < 0, then spam of that sentence is negative.
11. Else spam is neutral.
12. End.

V. GRAPHICAL & STATISTICAL ANALYSIS

For Statistical analysis we used precision and recall method to calculate the accuracy percentage of our project.

Precision (also called positive predictive value) is the fraction of retrieved instances that are relevant, while Recall (also known as sensitivity) is the fraction of relevant instances that are retrieved

Normal Dataset: Actual Objects = 20, Retrieved Objects = 18, Relevant Retrieved Objects = 17.

Spam Dataset: Actual Objects = 25, Retrieved Objects = 28, Relevant Retrieved Objects = 23.

Precision Class 1 = (Relevant Intersect Retrieved) / (Retrieved Objects) = 0.944444444

Precision Class 2 = (Relevant Intersect Retrieved) / (Retrieved Objects) = 0.821428571

Recall Class 1 = (Relevant Intersect Retrieved) / (Actual Objects) = 0.85

Recall Class 2 = (Relevant Intersect Retrieved) / (Actual Objects) = 0.92

Accuracy Percentage = 0.885

We are showing the output on a dashboard. For our project the dashboard images are as follows:

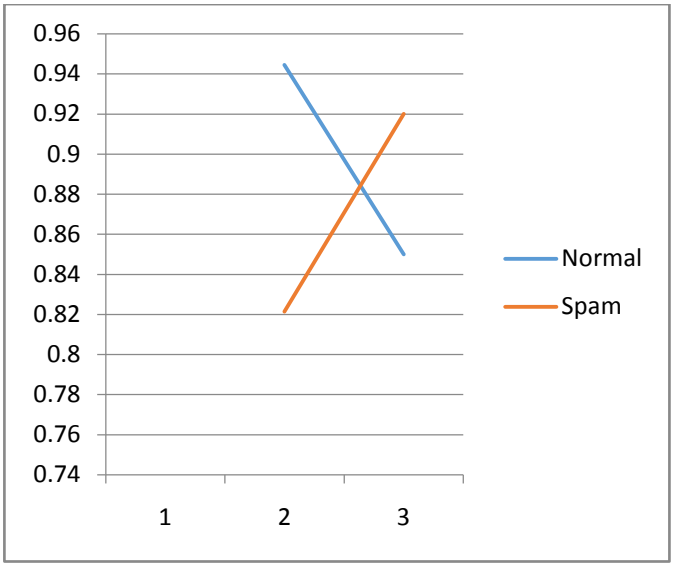

Figure. The figure shows the Percentage Accuracy for Normal and Spam dataset in our project.

VI. Conclusion

This paper Demonstrates the data mining approach on OSNs [2] (Online Social Networks) to detect the spam on data generated from feedbacks, comments to produce a categorized summary of messages. For the spam detection, the spam word dataset, URL blocking, keyword blocking is applied on rendered message. Then, the data mining or text classification algorithm is used to detect the overall spam. The designed system can be applied over several social networking sites such as Facebook, Quora, Twitter, and Instagram etc.

The future work will include implementation of Spam filtering in real-time environment in multi-threading environment. Also, the ability to handle more languages other than English as the proposed System can process only English as a language while filtering spam due to natural language processor restrictions are also left as a part of future work.

Acknowledgement

This work was supported by Sinhgad Institute of Technology and Science, Narhe, Pune, India in part of Final Year Project Paper Submission. The Views and conclusions contained in this document are those of authors and should not be interpreted as representing official policies, either expressed or implied, of the sponsors.

References

[1] Xin Jin, Cindy Xide Lin, Jiebo Luo, Jiawei Han A Data Mining based Spam Detection System for Social Media Networks. PVLDB , 2011

[2] Ritesh Kumar, Mayur Girnar, Prof. G.S Navale Spam detection using approach of data mining.In IJCA, 2014

[3] Kyumin Lee, James Caverlee, Steve webb *Uncovering social spammers: social honeypots + machine learning*, Published by ACM, 2010

[4] Prashant Bhosale, Shivani Sharma Sentiment analysis of big data using clickstream in hadoop, May 2014

[5] F. Benevenuto, T. Rodrigues, F. Benevenuto, T. Rodrigues, V. Almeida, J. M.Almeida, C. Zhang, and K.W. Ross. Identifying video spammers in online social networks. In AIR, 2008.

[6] Hongyu Gao, Yan Chen, Kathy Lee, Diana Palsetia, Alok Choudhary Towards Online Spam Filtering in Social Networks. In Proceedings of the 19th Annual *Network* & Distributed System Security Symposium, February 2012.