Filtering Unwanted Messages from OSN User walls using MLT

Prof.Sarika.N.Zaware[1], Anjiri Ambadkar[2], Nishigandha Bhor[3], Shiva Mamidi[4], Chetan Patil[5]
Department of Computer Engineering
AISSMS Institute of Information Technology
Savitribai Phule Pune University, India

***Abstract* – One of the basic issues these days on the social networking site is that our walls get spammed easily by spammer. Spam being unwanted advertisements, Vulgar messages, etc.Keeping these issues in mind we propose a system which gives OSN users direct control over the messages being posted on their walls. We attain this by obtaining the dataset of users details and putting this dataset through TF-IDF and NLP parser, based on machine learning concepts which automatically labels the posts as spam, Vulgar or legit.**

Keywords: Machine Learning Techniques (MLT), Online Social Network (OSN), Term Frequency -Inverse Document Frequency, Natural Language Processing, Spam, Filtering Wall, Data Set.

I. INTRODUCTION

We are well aware that online social networks are used immensely for staying in touch with people; Exchange of data, leisure, sharing of information, in short any kind of interaction is easily done by online social networks. Data could be images, Videos, Audios, Texts. This type of data is shared and exchanged all over the world in large amounts, a statistics on facebook reveals that a regular user roughly shares around 85-90 pieces of content(i.e. links, gossip stories, news, advertisements, pictures, audios, videos,etc)in each month. This Huge amount of information creates a scope for the web content mining strategies to automatically locate useful information within the whole data that is OSN management is highly supported by them. Information filtering gives the user ability to have an automatic control over messages written on their walls by filtering out the unwanted posts. OSN these days provide little support to prevent spam messages on user walls. For e.g. Sites like Google+, Facebook ,Twitter allows the user to choose the circle that is allowed to post on their walls(i.e. Friends, acquaintances, Friends of Friends, and any predefined group)But no content based filtering or preferences are given support in existing systems, which doesn’t prevent undesired messages like political statements, advertisements, Vulgar messages.

The main objective is to develop a filtering wall for the OSN user, and to filter out the unwanted messages, malicious spams, so that undesired events or consequences are avoided and the OSN user has no worry of unwanted things being posted on his wall. This task is to be completed by use of TF-IDF algorithm and NLP parser technique, which are robust techniques used for classification.

II. SYSTEM MODEL

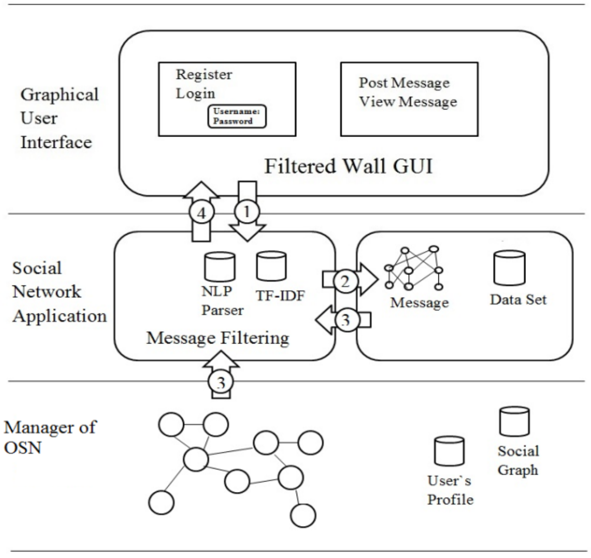

Fig 1.Conceptual System Architecture of Filtered Wall

The proposed conceptual architecture is a three –tier architecture. The First tier is the Manager of OSN, which provide basic functionality i.e. management of relationships and our overall profile. The Second tier nothing but the supporting layer for the external Social Network Applications. And the third tier provides a support to the second layer i.e. the Graphical user interface is required for interaction with second layer. Our proposed system is supposed to be located in the second and third tier of the System Architecture. And the GUI is used by the user for interaction with System and to setup and manage TF-IDF and NLP parser for obtaining the filtered messages. In our system GUI provides the user a Filtering Wall where the unwanted malicious, and spam messages are filtered and only legitimate messages are published according to the TF-IDF and NLP parsers.

The main elements of our proposed system are the Term Frequency-Inverse Document Frequency (TF-IDF) and the Natural Language Processing (NLP) [Explained in IV.]

The Route followed by messages (comments, posts) in our fig 1. Is stated as below:

a) As the user logs in into his private wall he tries to post or comment. This post/comment is intercepted by Filtering Wall.

b) The Data set is extracted for the message content.

c) Now, the filtering wall combines the Dataset from previous step along with the user's social graph and profile and it is sent through TF-IDF and NLP to Filter out the remaining content of Data.

d) Now from the result of the above step the message will be filtered or marked as spam and posted/removed respectively.

III. PREVIOUS WORK

The main aim of this paper is to design a system that provides content based filtering, that can be customized by user according to his needs. This concept is based on machine learning techniques. As per our best knowledge we've pointed out the main concerns of previously present systems in the introduction. Our work is related to content based filtering as well as policy based filtering so we provide the literature survey for both as follows Email related filtration:-

An anti-spam system proposed in the early years compared with traditional system, is based on an uncertain learning approach. This approach is integrated through commission collaboration mechanism. This Newly developed system was capable of handling dual-way spam filtering, i.e. both out-going as well as in-coming spam. After a real-time performance test of 6months on an email server it was proved that the new system has very low filtering output.

Online social network filtering using GAD Clustering:-

Then in the era of Social media networks represented by Fb, Twitter, YouTube and Flickr. It is observed that the users spend more time on social networks than search engines or any other sites. Also we know that famous entities or public figures set up social networking fan pages in order to have a direct interaction with their fans. Social media systems completely depend on users for content sharing and its contribution. Social networks are a medium to share and spread Information quickly and effectively. But, at the same time social media networks is susceptible to various kinds of unwanted and malicious spammer, or hackers. In today's society security solution in social media is of core importance to. This proposed system had a scalable online spam detection system for social network security. using GAD clustering algorithm. This was for large scale clustering and along with that integrating it with the designed active learning algorithm to deal with the scalability and real-time detection challenges.

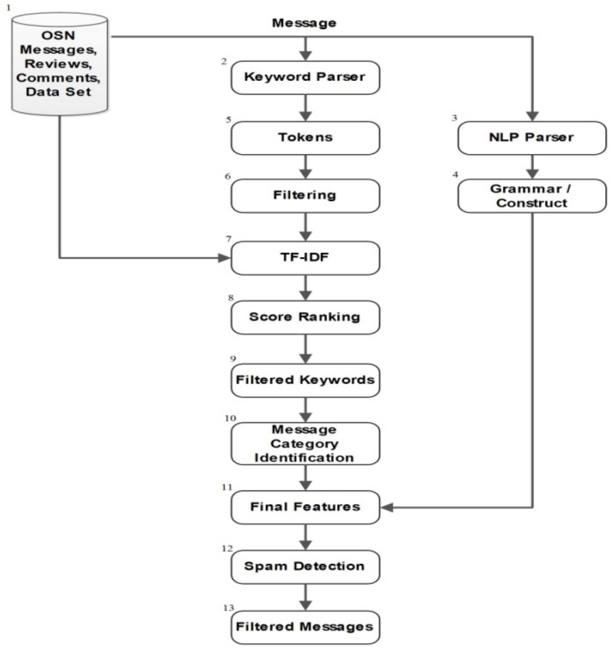

Fig 2. Flow Graph for Proposed System

The SMTP protocol falls short of a mechanism for authenticating the origin of a message, and protocol extensions are still far from standard. As a result of which Content-based automatic spam filters are used very often, and simple filtering techniques like blacklists and whitelists are very prevalent. Analysis is made to the retrieve the behaviour of email sources which might lead to delivery errors. The proposed novel spam filtering approach, founded the analysis of the above results.

Also sites like Facebook, Twitter, Google+ doesn't support content based filtering, We can just give access to post on our wall for selected users in our list or completely block them. It is possible that if a person posts a spam one time, next time he might post a useful message, so blocking a person completely is not the solution. Hence we, propose our system.

IV. PROPOSED METHODOLOGY

On the concepts based on paper [1], and taking the theme of that paper we propose a system model illustrating the complete flow of our system as shown in Fig 2. The stepwise explanation of the flow of the proposed system is as follows-

Step 1: It is the Dataset which contains all the comment, reviews posted on the OSN user walls.

Step 2& 3: The comments and messages from the Dataset go to the Keyword parser and the NLP parser. The job of the Keyword parser is to accept the input string and the NLP parser is based on machine learning so according to the large set of corpus it predicts the parts of speech(noun, verb, adjective)also the relation between them.(object, Subject)

Step 4: The output from NLP parser is input in this state and it constructs the grammar in this stage.

Step 5: The Comments parsed from the step 2 are tokenized in this stage. The tokens are input to the next stage.

Step 6: Here common stop words (like –is, and, the) are compared sequentially and filtered out, so that our final data set has the words that occur less frequently.

Step 7: The TF-IDF algorithm is applied here, the input is from the previous step 6 as well as the Data set. The TF-IDF value is calculated as below:

Example: Suppose you want to see which pages on your blog are the most related, you could check their category, or their tags. But that's an outdated technique. There are better way, Indeed! Such as tf-idf to determine the most important terms in each post, and then compare those results with every other post on your blog to find any overlaps.

The td-idf weighs words by their inverse frequency. So the more common a word is (think words like AND, BUT, IT, etc), the less weight it will have for a potential document.

Here's how it works:

Step a – calculate term frequency of a word in a document, and then divide it by number of total words in the document. Let's use the term "System" for this example calculation.

Suppose in a previous post, the word System showed up 2 times, in 725 total words. The TF value = 0.002759

Step b – To calculate (IDF) inverse document frequency divide the total number of documents, by number of documents containing the actual keyword which we intend to search. Then take logarithm of the obtained result.

For simplicity, let's assume that none of the other 7 blog posts, at the time of posting this article, contained the word System. The IDF value is therefore, log (7/1), or 0.8451.

Step c – multiply the TF by the IDF, to get the result 0.002759 * 0.8451 = 0.0023312

To put that into perspective, let's calculate TF-IDF for the word like. Since it's more generic, commonly used word in the English language, we expect that the term weight for like should be much lower.

I counted 4 instances of the word like in the link building blog post. Also, 4 of the 7 blog posts contain the word like. That gives us the following calculations:

$$4/725 = 0.005517$$

$$\text{Log } (7/4) = 0.2430$$

$$0.005517 * 0.2430 = 0.001341$$

The result, that the weight for the term system (0.0023312), is higher than the weight for the term like (0.001341), is what we expected, given that the word system comes up less often in the English language, than the term like. This tells us that the word System is more significant to this posting about link building, than the word like is – even though like was used twice as often (4 times vs 2).

Step 8 & 9: The values of TF and IDF are combined and the total TF-IDF score is given. Then according to our previously given threshold value we compare it with the TF-IDF score calculated and again the less valued words are further filtered out.

Step 10: The category of message is identified. That is whether the message is legitimate, Vulgar or spam.

Step 11: In Final feature the output of TF-IDF and NLP parser is combined together and a final average score is calculated forwarded to the spam detection module.

Step 12 & 13: Finally the final message score is compared with our spam threshold value and it is declared if a message is spam or normal message.

This intense filtering enables to give a reliable output to the user.

V. CONCLUSION

This paper is a study paper for our proposed system that filters unwanted and malicious messages from OSN walls; this helps the user to keep out of any crimes or issues regarding the contents posted on the walls. The System based on machine learning algorithms enables filtering of data. System also provides user to manage his/her posts content value during the posts on other wall. This proposed system is the stepping stone to a wider scope for advanced projects in this subject.

VI. FUTURE SCOPES

As we develop filtered system for OSN's similarly we can develop this system for other online information sites like Wikipedia, GOOGLE etc. We can give user to keep other users in blacklist for some time and also warn him regarding the content posted on the wall. This helps to block unwanted information from different users. Currently our System works on changing only filtering unwanted messages in future we can extend our project that works on unwanted Images, Audio, Video format filtering.

REFERENCES

[1] A System to Filter Unwanted Messages from OSN User Walls M. Vanetti, E. Binaghi, B. Carminati, M. Carullo, and E. Ferrari, "Content-based filtering in on-line social networks," in Proceedings of ECML/PKDD Workshop on Privacy and Security issues in Data, Mining and Machine Learning (PSDML 2010), 2010.Name of Authors, "Title of the research", Citation Details, year.

[2] Understanding Inverse Document Frequency: On theoretical arguments for IDF By - Stephen Robertson Microsoft Research7 JJ Thomson Avenue Cambridge CB3 0FB UK (and City University, London, UK)

[3] F. Sebastiani, "Machine learning in automated text categorization," ACM Computing Surveys, vol. 34, no. 1, pp. 1–47, 2002.

[4] http://www.tfidf.com/

[5] http://www.site.uottawa.ca/~diana/csi4107/cosine_tf_idf_example.pdf

[6] Natural language processing - Wikipedia, the free encyclopedia

[7] Machine Learning-Mark K Cowan

[8] http://nlp.stanford.edu/nlp/javadoc/javanlp/edu/stanford/nlp/parser/lexparser/package-summary.html

[9] http://ccl.pku.edu.cn/douBTfire/NLP/Parsing/Introduction/Grammars%20and%20Parsing.htm

[10] Towards Online Spam Filtering in Social Networks, Hongyu Gao, Yan Chen, Kathy Lee†, Diana Palsetia†, Alok Choudhary†

[11] Content-based Filtering in On-line Social Networks, M. Vanetti, E. Binaghi, B. Carminati, M. Carullo and E. Ferrari, Department of Computer Science and communication University of Insubria 21100 Varese, Italy.