Automatic Removal of Malicious Post at Zero Hour from Facebook

Rugveda Mane[1], Tazkiya Momin[2], S Jeni[3], Paurnima Shewale[4]
[1, 2, 3, 4] Department of Information Technology
All India Shri Shivaji Memorial Society's Institute of Information Technology, Pune
Savitribai Phule Pune University, Pune, Maharashtra India

Abstract-*Now a days the adoption rate of Facebook like social networking website is increasing with significant ratio. These sites are widely used for posting posts, sharing images and group communications. Possibilities of illegal activities are unavoidable as the post and activities are public to everyone directly or indirectly. These illegal activities include creating fake accounts, posting malicious posts, adult images etc. Anything posted on OSN gets viral within a short span of time. If the post is malicious in nature it may cause a riot which would disturb the normal working of society. Our proposed system is addressing this issue by automatically removing malicious posts in zero hour by creating a portal which would classify the user's post into different categories and to further analyze and recognize the malicious post using NLP (Natural Language Processing). Sentiment analysis is done on user posts and comments to detect user sentiments. Adult images are blocked using adult detection based on image processing.*

Keywords-Malicious post, sentiment analysis, adult image

I. INTRODUCTION

Web applications, especially social networks (such as Facebook, Twitter etc.) are enjoying ever growing popularity. One of the most famous and popular social networks is Facebook with 1.17 billion monthly active users in 2016 and has recently surpassed Google as the most visited site on the Internet.[1] A multitude of examples exist which demonstrate how Facebook influences our daily life. Even in areas of life which have always considered being private and/or intimate is shared publically now, the usage of Facebook has become more popular .So eventually the rise in Facebook activities is rapidly increasing day by day. For example during 32 days of the FIFA world cup 2014 Facebook noted 350 million users posting over 3 billion posts, comments and likes [2]. Every 60 seconds on Facebook 510 comments are posted, 293000 statuses are updated and 136000 photos are uploaded. This enormous magnitude of activities makes Facebook a lucrative venue for malicious entities to seek monetary gains and compromise system reputation. Today Facebook, being the most preferred OSN for users to interact with each other, group communications, to post their opinions and get news, is potentially the most attractive platform for malicious entities to launch cyber-attacks. These cyber-attacks include misinformation on Facebook, luring victims into scams, phishing attacks, malware infections, malicious post etc. It has been claimed that Facebook spammers make $200 million just by posting links. Such activity not only degrades user experience but also violates Facebook's terms of service. Lately one post can create a havoc or cause riots if group of people find it offending. Thus the environment of the society is disturbed as well as properties are damaged because of a single malicious post. There have been numerous real time examples of this over the world where facebook posts caused riots. For e.g. in Mumbai on 21 June 2014, riots took place in Dhule district because objectionable content about minority community had been posted on Facebook. Another famous riot took place in Pune in June 2014 where a controversial facebook post that contained defamatory pictures with allegedly derogatory references to warrior King Shivaji Maharaja were posted on Facebook which caused a violent protest and affected the place for two days.

We propose to address the problem of automatic detection of malicious content posted by user at zeroth hour. We intend to develop a portal that classifies user posts with the help of NLP into different categories such as Politics, Education, Entertainment and Sports etc. Classifying the post gives probable effect of the original post, so it's easier to understand social effect of any post on society or any such social media. So we focus on user statuses, which can be viewed as opinions of users or their reaction on concern we want to analyze. Texts are extracted from posts, images to know how people feel about different posts thus sentiment analysis is done. Sentiment analysis is applied on classified post to identify good and bad words; the post containing maximum bad words are further automatically removed by implementing NLP. Therefore, any user posting any malicious post which would cause disturbance in society is automatically removed from this Social Networking Portal. Adult images are blocked using adult image detection based on Image Processing.

II. RELATED WORK

Detection of malicious content on Facebook: Gao et al.[3] used facebook accounts of different users to do an initial study to quantify and characterize spam campaigns with the help of a set of automated techniques to detect and characterize the coordinated spam campaigns .The authors observed a huge anonymized dataset of 187 million asynchronous wall messages between various Facebook users. In return authors detected approximately 200,000 malicious wall posts with embedded URLs, which were originating from more than 57,000 user accounts. Following this, Gao et al.[4] then proposed an online spam filtering system to inspect messages generated by users in real time as a component of the OSN platform. Rather than analyzing each post individually, this approach mainly focused on redeveloping spam messages into campaigns for classification. Resultant was that, using 187 million facebook wall posts as their dataset they got true positive rate of roughly over 80%. Also authors achieved 1,580 messages/sec as average throughput and 21.5m as an average processing latency rate. However, this approach was not successful in detecting any new malicious post if the system has not noted it previously as the clustering approach used always marked a new cluster as legitimate.

To protect Facebook users from real time malicious posts, Rahman et al.[5] took advantage of the social context of posts to deploy a social malware detection method. Using a SVM based classifier trained on 6 features; a maximum true positive rate of 97% was achieved by the authors. The classifier took 46ms to classify a post. MyPageKeeper, a facebook app was developed using this model to protect its users from malicious posts. This model also targeted at detecting spam campaigns, and depended on message similarity features. Such techniques are efficient in detecting content which they have seen in the past, for example, campaigns. However, if the system is not familiar with the post in past, then these techniques are incapable of detecting malicious posts in real time. But in our propose system we overcome this flaw by using NLP to detect malicious post at zeroth hour.

Facebook's Current Techniques:

A. Detect Malicious URLs in Real Time

For detecting malicious URLs in real time and preventing them from entering the social graph, Facebook's immune system uses multiple URL blacklists [6]. The limitation of the blacklist is that it is incapable in detecting URLs at zero-hour which limits the effectiveness of this technique [7].After taking an analysis using Graph API to check if Facebook removed any of the 11,217 malicious posts identified by blacklists after being posted. The result was disappointing as only 3,921 out of the 11,217 (34.95%) malicious posts had been deleted the remaining got past Facebook's real time filters i.e. almost two thirds of all malicious posts (65.05%) and it remained undetected even after 4 months (July - November, 2014) from the date of post.

Fig. 1 shows an example of a malicious post from Facebook, this URL in the post ask users to like a post on Facebook to earn money as indirectly its pointing to a scam website.

B.WOT (Web of Trust) warning pages:

To protect its users from malicious URLs, in 2011 Facebook partnered with Web of Trust [8]. This partnership states that whenever a user clicks on a link which has been reported on WOT as malware, phishing, spam or any other kind of abuse , then Facebook shows a warning page to the user (Figure 2).

To verify this claim and to cross check the existence and effectiveness of the warning pages , we visited some random 1000 posts on Facebook containing a URL marked as malicious by WOT, and clicked on the URL. Surprisingly, the warning page did not appear even once.

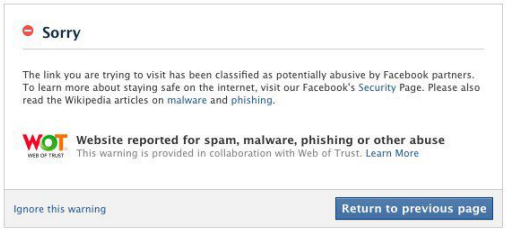

Fig. 2. Example of WOT warning page which Facebook claims to show whenever a user clicks on a link which is noted as abusive on Web of Trust.

TABLE I. RIOTS CAUSED BY FACEBOOK STATUS ACROSS WORLD

Sr. No.	Description	Place	Consequences
1	Communal violence erupted over 'objectionable video' posted on Facebook over Hindu God and Goddess	Chhapra, Bihar Aug 6, 2016	-Mosques were damaged by petrol -shops of Muslim were looted and shops were set to fire
2	Communal violence erupted from an alleged objectionable Facebook post against Prophet Muhammad	Birbhum, West Bengal March 3, 2016	-1 killed in police firing -1police station was ransacked
3	Defamatory post morphing photos of Chhatrapati Shivaji, Bal Thackeray and others on Facebook sparks violence across the city	Pune , Maharashtra Jun 2, 2014	-24 out of 33 police stations were affected stones were pelted at vehicles and damaged 130 PMPML buses and 21 private vehicles, as also set fire to one bus, tempo
4	Woman and two children killed by mob in riots over 'blasphemous' Facebook post	Pakistan July 2014	Houses of religious minority group Ahmadiyyas, were torched by a mob
5	A boy posted a morphed image of a Hindu goddess on Facebook	Gujarat's Vadodara	The place was disturbed for a week

III.METHODOLOGY

1. NLP (Natural Language Processing):

NLP is used for detecting and removing malicious post on zeroth hour. Following steps are performed in NLP:

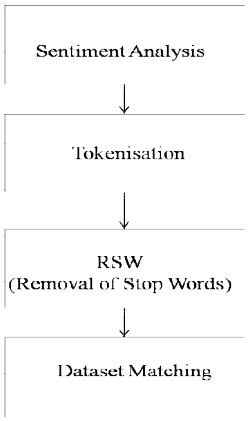

Fig.3. Steps performed in NLP

- Step 1. Sentiment Analysis: Sentiment analysis is a technique that determines the attitude of text. Sentiment analysis is a type of classification. It is a concerned with determining what text is trying to convey to a reader, usually in the form of a positive and negative attitude. Over here, we use 'sentiment analysis' to refer to the task of automatically determining feelings whether text, ismalicious in nature or not.
- Step 2. Tokenization: Tokenization is the process of breaking a stream of text up into words, phrases, symbols, or other meaningful elements called tokens. The list of tokens becomes input for further processing such as parsing or text mining. We use tokenization for breaking the user's post or comment for further processing.Then each token is inserted into stack.
- Step 3. RSW(Removal of stop words):
  Sometimes, some extremely common words which would appear to be of little value in helping select documents matching a user need are excluded from the vocabulary entirely. These words are called stop words. Stop words are filtered out before or after processing of data. Stop words are usually referred to the most common words in a language; there is no single universal list of stop words used by all natural language processing tools. After tokenization, we remove stop words from the stack which are not useful in detecting the malicious word e.g. When, the, an etc.
- Step 4. Dataset Matching:The words which are left after the removal of stop words are compared with the dataset which contains a list of abusive/ malicious words. If the word founds a match in the dataset then it is malicious in nature and is blocked from further processing and the words which are not malicious in nature are categorised into various category in the dataset.

2. Adult Image Detection(AID):

This method detects whether image is adult or not using Skin Tone Pixels detection and threshold value for number of Skin tone pixels allowed.Following steps are performed in AID:

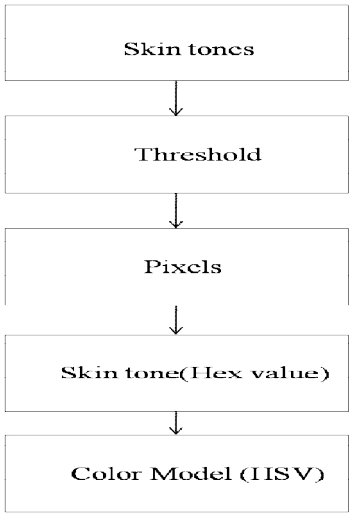

Fig.4. Steps performed in AID

- Step 1. Skin tone: The goal of skin tone detection is to build a decision rule that will differentiate between skin and non-skin pixels.

- Step 2.Threashold: The next step for skin detection in an image is by assigning a threshold value(skin tone) i.e. if the probability of skin tone in image is more or equal to the assigned threshold value then that image is considered as adult and will be blocked from uploading .

- Step 3.Pixels and HEX value: To do the comparing of the skin tone with the assigned threshold value, it is converted into hex value.

- Step 4. Color Model (HSV): HSV is named for three values - Hue Saturation Value. Hue-saturation based color spaces describes hue or tint , saturation or amount of gray in term of colors and shade and brightness value .Hue defines color (such as red, green, yellow and purple) of an area, saturation measures the colorfulness of an area in proportion to its brightness .The ”intensity”, ”lightness” or ”value” is related to the color luminance. To do all these processing e we are using HSV color model

IV. CONCLUSION

Lately OSNs like Facebook is a medium used by people to express their views and opinions on everything. People post their views on Facebook through post/status or comments. In a twinkling of an eye a large amount of data is uploaded, shared and liked by many users. A single post can give a positive or a negative outcome. But if the post is controversial in nature it can cause uproar in society. A riot has always caused a lot of damage to the physical materials as well to the people’s life. We noted the riots caused by Facebook post across the world and their consequences. And as the saying goes precaution is always better than cure, we intend to deploy a real world solution using NLP to detect and automatically remove malicious post from Facebook at zeroth hour. And also to categorize the post into different categories using sentiment analysis. And using adult image detection technique to block adult images.

REFERENCES

[1] Zephoria Digital Marketing. [Online] https://zephoria.com/top-15-valuable- facebook-statistics/.

[2] CNN.[Online] http://edition.cnn.com/2014/07/14/tech/social-media/world-cup-socialmedia/.

[3] H. Gao, J. Hu, C. Wilson, Z. Li, Y. Chen, and B. Y. Zhao. Detecting and characterizing social spam campaigns in Internet MeasurementConference, ACM, 2010.

[4] H. Gao, Y. Chen, K. Lee, D. Palsetia, and A. N. Choudhary. Towards online spam filtering in social networks in NDSS, 2012.

[5] M. S. Rahman, T.-K. Huang, H. V. Madhyastha, and M. Faloutsos. Efficient and scalable socware detection in online social networks in USENIX Security Symposium, pages 663–678, 2012.

[6] T. Stein , K. Mangla and E. Chen,. Facebook immune system in workshop on Social Network Systems, page 8. ACM, 2011.

[7] S. Sheng, G. Warner, B. Wardman , L. Cranor, J. Hong, and C. Zhang. An empirical analysis of phishing blacklists in 6th Conference on Email and Anti-Spam (CEAS), 2009.

[8] Facebook Developers: Keeping you safe from scams and spam. https://www.facebook.com/notes/facebook-security/keepingyou-safe-from-scams-and-pam/10150174826745766, 2011.

[9] https://apps.facebook.com/mypagekeeper/