ISSN: 2582-8118 ***Volume 2, Issue 1; Jan 2022***

INTERNATIONAL JOURNAL OF RESEARCH AND ANALYSIS IN SCIENCE AND ENGINEERING

Web: https://www.iarj.in/index.php/ijrase/index

4. Malicious Posts Detection Using Emotions and Reputation on Facebook Data

Dr. R. Senthil Kumar
Professor,
Department of CS with Cognitive Systems.

Mrs. S. Revathy
Assistant Professor (SG),
Department of Computer Science
Dr. N.G.P. Arts and Science College,
Coimbatore, Tamilnadu.

ABSTRACT

Online Social Networks (OSNs) witness a rise in user activity whenever an event takes place. Malicious entities exploit this spur in user-engagement levels to spread malicious content that compromises system reputation and degrade user experience and has recently been reported to face much abuse through scams and other type of malicious content, especially during news making events. It has been observed that there is a greater participation in Facebook pages regarding malicious content generation.

These contents will be in greater amount as compared to legitimate content. These issues are addressed in this research work, whose main goal is to detect the malicious activities involved in the social web pages accurately. In the proposed research method, Reputation and EMOtional (REMO) score based malicious post detection framework is introduced. This method combines the power of reputation score observed through WOT and the emotional score obtained from the annotation profile of the post. Annotation in the Facebook data refers to the user's emotion about the post such as like, dislike, angry and so on. In this case, posts that receive more dislikes are probably malicious. Similarly, posts with no annotation also likely to be verified as malicious. The overall implementation evaluation of the proposed research method is done in the python simulation environment from which it can be proved that the proposed work can provide optimal outcome than the existing research methods.

KEYWORDS

Social web pages, Facebook, Twitter, Malicious post, Emotions, post, annotations, reputation score.

1. Introduction:

Social network activity rises considerably during events that make the news, like sports, natural calamities, etc [1]. For example, the 2014 FIFA World Cup final inspired more than 618,000 tweets per minute, a new record for Twitter. Facebook also saw 350 million users generating over 3 billion posts, comments and likes during the 32 days of the world cup [2]. This enormous magnitude of activity during sports and other news-making events makes OSNs a lucrative venue for malicious entities to seek monetary gains and compromise system reputation [3]. Facebook, being the most preferred OSN for users to get news, is potentially the most attractive platform for malicious entities to launch cyber-attacks.

Recently, cyber criminals exploited the context of various news events to spread hoaxes and misinformation on Facebook, luring victims into scams, phishing attacks, malware infections, etc [4]. It has been claimed that Facebook spammers make $200 million just by posting links. Such activity not only degrades user experience but also violates Face book's terms of service. Facebook has acknowledged spam and hoaxes as serious issues, and taken steps to reduce malicious content in users' newsfeed [5].

Researchers have used various supervised learning models to detect spam and other types of malicious content on OSNs and achieved good results [6]. However, existing approaches to detect malicious content on other OSNs like Twitter, YouTube etc. cannot be directly ported to Facebook because these approaches heavily rely on features that aren't publicly available from Facebook [7]. These features include pro- file and network information, age of the account, total number of messages posted, social connections, etc [8].

In this work, we address the problem of automatic real time detection of malicious content generated during news making events that is currently evading Face book's detection techniques. To this end, we collect 4.4 million public posts generated by 3.3 million unique entities during 17 news making events that took place between April 2013 and July 2014. We first study the effectiveness of existing techniques used by Facebook to counter malicious content. Then, we identify some key characteristics of malicious content spread on Facebook, which distinguish it from legitimate content.

We propose an extensive feature set consisting of 42 features to automatically distinguish malicious content from legitimate content in real time. This feature set is used to train multiple machine learning models to identify malicious posts on Facebook, and attains a maximum accuracy of 86.9% using the Random Forest classifier. Our experiments show that prior clustering based spam detection techniques are able to detect less than half the number of malicious posts as compared to our model. We use our model to deploy a publicly available REST API 2 and a browser plug-in that can be used to identify malicious content on Facebook in real time. Our broad contributions are as follows:

- Characterization of malicious content generated on Facebook during news-making events. Our dataset of 4.4 million public posts is one of the biggest datasets of Facebook posts in literature.
- Extensive feature set for identifying malicious content in real time, excluding features like likes, comments, shares, etc. which are absent at post creation time.

- Publicly available end-user solution (API and browser plug-in) to identify malicious posts in real time.

In the proposed research method, Reputation and EMOtional (REMO) score based malicious post detection framework is introduced. This method combines the power of reputation score observed through WOT and the emotional score obtained from the annotation profile of the post. Annotation in the Facebook data refers to the user's emotion about the post such as like, dislike, angry and so on. In this case, posts that receive more dislikes are probably malicious. Similarly, posts with no annotation also likely to be verified as malicious.

The overall organization of the research work is given as follows: In this section, detailed introduction about the malicious activities involved in the social media websites has been discussed in detailed. In section 2, discussion about the various related research methodologies that are conducted with the goal of accurate malicious post identification is provided. In section 3, detailed discussion about the proposed research methodology along with required examples and explanation has been given. In section 4, simulation evaluation of the proposed research methodology is given with suitable examples and explanation. Finally in section 5, overall conclusion of the research method based on simulation outcome is given.

2. Related Works:

Gao, Hu, et al. (2010) examined a study on detecting and characterizing social spam campaigns. The author states that their system detected roughly 200,000 malicious wall posts with embedded URLs, originating from more than 57,000 user accounts. It is also find that more than 70% of all malicious wall posts advertise phishing sites. Further, author examined the characteristics of malicious accounts, and see that more than 97% are compromised accounts, rather than "fake" accounts created solely for spamming. It is observed that spamming dominates actual wall post activity in the early morning hours, when normal users are asleep.

Rahman, Huang et al. (2012) proposed a framework known FRAppE (Facebook's Rigorous Application Evaluator) which is arguably the first tool focused on detecting malicious apps on Facebook. The author states that the proposed toole can detect malicious apps with 99.5% accuracy with no false positives and a low false negative rate (4.1%). Rahman, Huang et al. (2012), studied about efficient and scalable socware detection in OSNs. In this article, the author proposed a method called MyPageKeeper, which is a Facebook application helps to protect Facebook users from socware. The result reveals that their classifier can 97% accurately predict the socware. Finally, it is also identify a new type of parasitic behaviour, which we refer to as "Like-as-a-Service", whose goal is to artificially boost the number of "Likes" of a Facebook page.

Faraz Ahmed, Muhammad Abulaish (2013) studied a generic statistical approach for detecting spam in Online Social Networks (OSNs). This study uses dataset, which was extracted from Facebook and Twitter networks. There are about 14 generic statistical features have extracted to identify the spam. Three type classifications were applied such as

Naïve Bayes, Jrip and J48. The result states that detection rate on Facebook dataset is 95.7%, whereas the detection rate on Twitter dataset is 97.6%.

Chen, Guan et al. (2014) examined the feature set identification and detection of suspicious URL using Bayesian classification. The feature set considered in this study combines the features of traditional heuristics and social networking. Furthermore, a suspicious URL identification system for use in social network environments is proposed based on Bayesian classification. The experimental results indicate that the proposed approach achieves a high detection rate.

Dewan and Kumaraguru (2015) conducted a study on detecting malicious content on Facebook. The author proposed an extensive feature set based on entity profile, textual content, metadata, and URL features to identify malicious content on Facebook in real time and at zero-hour. This feature set was used to train multiple machine learning models and achieved an accuracy of 86.9%. The intent is to catch malicious content that is currently evading Facebook's detection techniques. Further, the author states that the proposed technique is capable of predicting malicious content more than double compared with existing techniques.

Dhawan, Singh et al. (2017) examined a study on identification of malicious posts in Facebook. In this article, the author presents comparison between existing techniques with their pros and cons. The author examined another study on recognizing the malicious posts and user behavior. In this paper, author proposed a framework to distinguish genuine posts or malicious posts shared or posted by users on a Facebook page in Facebook and similarly it may be extended to other social networking sites for example (Twitter, Instagram, Whatsapp etc). To analyze proposed framework real dataset is collected from netvizz a facebook application and process it using Gephi tool.

3. Web of Trust (WOT) Based Reputation Score:

Web of Trust (WOT) is a browser add-on and web site. WOT is an online reputation and Internet safety service, providing crowdsourced reviews and other data about whether websites respect user privacy, are secure, and other indicators of trust. According to the company information the WOT software computes the measure of trust the rating users have in websites, combined with data from, among others, Google Safe Browsing. The WOT browser add-on is available for all major operating systems and browsers. To view or submit ratings, no subscription is required. To be able to write comments on score cards and in the forum, one needs to be registered.

The add-on sends user ratings to the WOT site, and it determines how the computed results are displayed, depending on user's settings. For instance, when visiting a poorly rated site, a warning screen may pop up, or only a red icon in the user's browser tool-bar is shown. Color-coded icons are also shown next to external links on the pages of leading search engines, on email services, on social network sites, and on Wikipedia. Ratings are cast by secret ballot. They can be given in the categories "trustworthiness" and "child safety". To specify at least one reason for a rating is mandatory, via multiple choices in the rating interface. The user rating system is meritocratic; the weight of a rating is algorithmically calculated for each user individually.

Reputation management refers to influencing and controlling an individual's or business's reputation. Originally a public relations term, the expansion of the internet and social media, along with reputation management companies, has made it primarily an issue of search results. Online reputation management, sometimes abbreviated as ORM, is primarily concerned with managing the results on websites that evaluate products and services and make recommendations and referrals. Ethical grey areas include mug shot removal sites, astroturfing review sites, censoring negative complaints or using search engine optimization tactics to influence results.

WOT Reputation Score – Algorithm

The subsequent part explains about the algorithm helps to estimate the reputation score based on WOT.

For all posts do

For all URL domains do

Components = GetComponentFromWOT_API

For all components do

If reputation < 60 and confidence ≥ 10 then

Post = malicious

End if

End for

End for

End for

The above algorithm describes the WOT reputation score. We marked a URL as malicious if one or more of these services categorized the URL domain as spam, malicious, or phishing. The WOT API returns the reputation score for a given domain. Reputation scores are measured for domains in several components, for example, trustworthiness. For each {domain, component} pair, the system computes two values: a reputation estimate and the confidence in the reputation. Together, these indicate the amount of trust in the domain in the given component. A reputation estimate of below 60 indicates unsatisfactory. The WOT browser add-on requires a confidence value of more than 10 before it presents a warning about the website. In addition, the WOT rating also computes categories for websites based on votes from users and third parties.

The reason for including WOT reputation scores in our labeled dataset of malicious posts was two-fold. Firstly, to study Facebook’s current techniques to counter malicious content.

Facebook partnered with WOT to protect its users from malicious URLs. Secondly, during news-making events, malicious entities tend to engage in spreading fake, untrustworthy and adult content to degrade user experience. This kind of information, despite being malicious, is not captured by blacklists like Google Safe-browsing and SURBL, since they do not fall under the more obvious kinds of threats like malware and phishing. WOT scores helped us to identify and tag such content. In all, we found 4,622 unique malicious URLs across 11,217 unique Facebook posts.

4. Reputation and Emotional Score:

This section describes the proposed approach known as REMO (Reputation and EMOtional), which combines the power of reputation score observed through WOT and the emotional score obtained from the annotation profile of the post. Annotation in the Facebook data refers to the user's emotion about the post such as like, dislike, angry and so on. In this case, posts that receive more dislikes are probably malicious. Similarly, posts with no annotation also likely to be verified as malicious. Therefore, these two scenarios help to filter the actual list of data and those extracted list of data are seeking for the reputation score using WOT.

REMO Algorithm

```
i = 0

For all posts do

    If post[i].total_emotions = 0 then

      post[i].malverify = True

    Else If post[i].total_emotions > 0 then

      risk = (post[i].total_emotions – (post[i].like+post[i].haha))/post[i].total_emotions

      If risk > 0.5 then

          post[i].malverify = True

      End If

    End If

    i = i + 1

End for

i = 0
```

```
For all malicious verify posts do
    For all URL domains do
        Components = GetComponentFromWOT_API
        For all components do
            If reputation < 60 and confidence ≥ 10 then
                post[i].malicious = True
            End if
        End for
    End for
    i = i + 1
End for
```

As mentioned above, the objective of this algorithm is to estimate the given Facebook post as malicious based on reputation score and emotion score. The dataset contains 19850 posts and fourteen attributes.

Further, forty five features are extracted from the dataset for further evaluation, which forms a new dataset. Every Facebook post contains certain annotations, which can help us to find the specified post as malicious. In general, if any post receives more dislikes, angry kind of feedback which is probably a malicious post.

Specifically when the computed risk factor exceeds more than 50% referred as probably malicious. Similarly, all posts to be considered for validating as malicious when no feedback was received. Finally, the algorithm validating the posts which are identified as risk and no feedback through reputation score obtained from WOT_API. This approach is straight forward and it helps to reduce the computation time and enhance the quality of prediction.

5. Performance Evaluation:

The Figure-1 shows the classification accuracy of WOT and REMO algorithms, while predicting the malware from the Facebook posts. The result states that the REMO algorithm finds 83.7% of malicious posts whereas the WOT algorithm finds 82.1% of malicious posts.

It is also observed that very little overlap between the URLs flagged by blacklists and those flagged by the classifier. Therefore, the chances of misclassification error are low in the REMO algorithm.

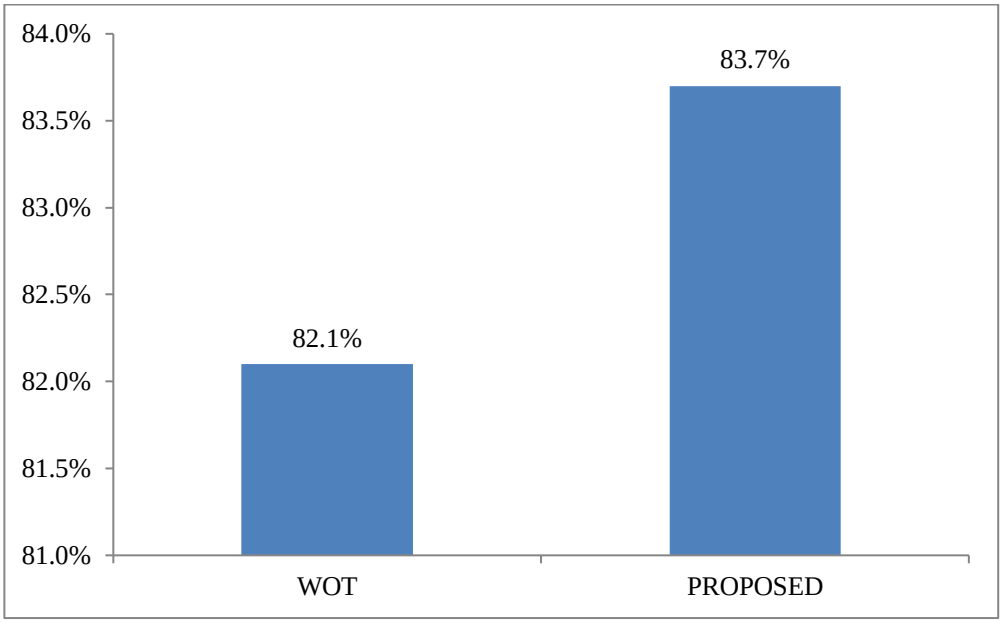

Figure 1: Accuracy Prediction using WOT and REMO algorithms

As like the accuracy of predicting malicious posts, the execution time also validated to measure the efficiency. Subsequent Figure-2 exhibits the performance of WOT and REMO algorithms in the context of execution time. It is observed from the result that the REMO algorithm utilize less time (31 seconds) than WOT algorithm (39 seconds). Hence, it is decided that the proposed algorithm is superior in terms of prediction accuracy and less utilization of execution time.

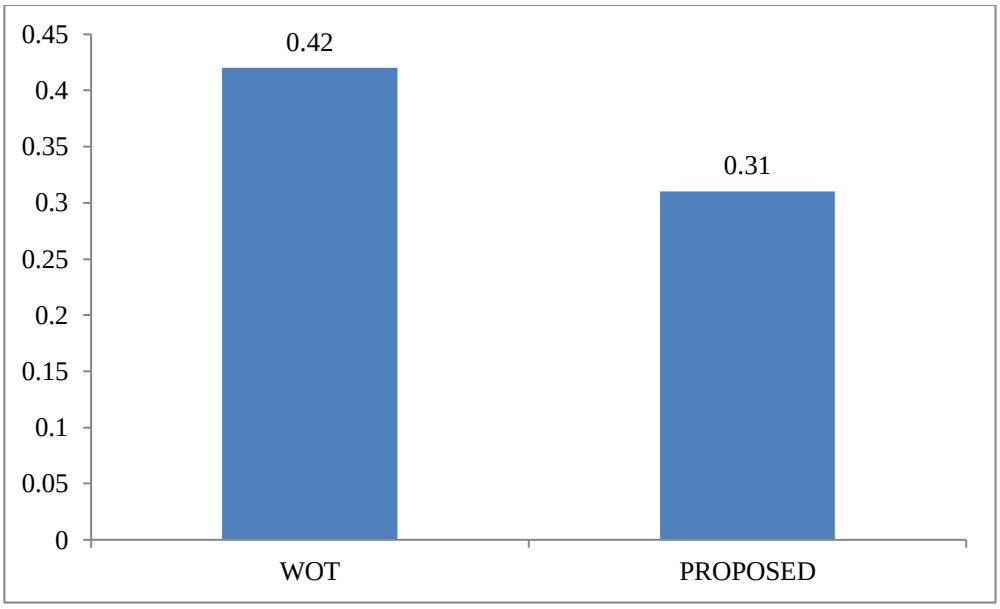

Figure 2: Execution Time of WOT and PROPOSED models

6. Conclusion:

The Online Social Networks (OSNs) has opened up new possibilities for the dissemination of malware. In the proposed research method, Reputation and EMOtional (REMO) score based malicious post detection framework is introduced. This method combines the power of reputation score observed through WOT and the emotional score obtained from the annotation profile of the post. Annotation in the Facebook data refers to the user's emotion about the post such as like, dislike, angry and so on. In this case, posts that receive more dislikes are probably malicious. Similarly, posts with no annotation also likely to be verified as malicious. The overall implementation evaluation of the proposed research method is done in the python simulation environment from which it can be proved that the proposed work can provide optimal outcome than the existing research methods.

References:

1. Imran, M., Castillo, C., Diaz, F., & View eg, S. (2015). Processing social media messages in mass emergency: A survey. ACM Computing Surveys (CSUR), 47(4), 67.
2. Foote, J. G., Butterworth, M. L., & Sanderson, J. (2017). Adrian Peterson and the "Fuzzification of America": Football and Myths of Masculinity. Communication Quarterly, 65(3), 268-284.
3. Dewan, P., Suri, A., Bharadhwaj, V., Mithal, A., & Kumara guru, P. (2017, July). Towards Understanding Crisis Events On Online Social Networks Through Pictures. In Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017 (pp. 439-446). ACM.
4. Wani, M. A., Jabin, S., & Ahmad, N. (2017). A sneak into the Devil's Colony-Fake Profiles in Online Social Networks. arXiv preprint arXiv:1705.09929.
5. Dewan, P., & Kumaraguru, P. (2015). Detecting Malicious Content on Facebook. arXiv preprint arXiv:1501.00802.
6. Crawford, M., Khoshgoftaar, T. M., Prusa, J. D., Richter, A. N., & Al Najada, H. (2015). Survey of review spam detection using machine learning techniques. Journal of Big Data, 2(1), 23.
7. Kumar, S., Gao, X., Welch, I., & Mansoori, M. (2016, March). A machine learning based web spam filtering approach. In Advanced Information Networking and Applications (AINA), 2016 IEEE 30th International Conference on (pp. 973-980). IEEE.
8. Egele, M., Stringhini, G., Kruegel, C., & Vigna, G. (2017). Towards detecting compromised accounts on social networks. IEEE Transactions on Dependable and Secure Computing, 14(4), 447-460.
9. Ahmed, F., & Abulaish, M. (2013). A generic statistical approach for spam detection in online social networks. Computer Communications, 36(10-11), 1120-1129.
10. Chen, C. M., Guan, D. J., & Su, Q. K. (2014). Feature set identification for detecting suspicious URLs using Bayesian classification in social networks. Information Sciences, 289, 133-147.
11. Dewan, P., & Kumaraguru, P. (2015). Detecting Malicious Content on Facebook. arXiv preprint arXiv:1501.00802.
12. Dewan, P., & Kumaraguru, P. (2015, July). Towards automatic real time identification of malicious posts on Facebook. In Privacy, Security and Trust (PST), 2015 13th Annual Conference on (pp. 85-92). IEEE.

13. Dhawan, S., Singh, K., & Sagwal, S. (2017). Identification of Malicious Posts in Facebook Social Networks.
14. Dhawan, S., Singh, K., & Sagwal, S. (2017). Recognition of Malicious Posts among Facebook Social Network Groups to Investigate User's Behavior. Journal of Network Communications and Emerging Technologies (JNCET) www. jncet. org, 7(9).
15. Gao, H., Hu, J., Wilson, C., Li, Z., Chen, Y., & Zhao, B. Y. (2010, November). Detecting and characterizing social spam campaigns. In Proceedings of the 10th ACM SIGCOMM conference on Internet measurement (pp. 35-47). ACM.
16. Rahman, M. S., Huang, T. K., Madhyastha, H. V., & Faloutsos, M. (2012, August). Efficient and Scalable Socware Detection in Online Social Networks. In USENIX security symposium (pp. 663-678).
17. Rahman, M. S., Huang, T. K., Madhyastha, H. V., & Faloutsos, M. (2012, December). Frappe: detecting malicious facebook applications. In Proceedings of the 8th international conference on Emerging networking experiments and technologies (pp. 313-324). ACM.