International Journal of Innovative Research in Technology and Science

ISSN: 2321-1156 Volume 12, Issue 2, March 2024

Unveiling Malicious Intent: A Theoretical Framework for Detecting Anomalies in Social Media Behaviours

Rajesh Nigam [1], Pradeep Pandey [2],Gaurav Mishra [3],
[1,2] Computer Science and Engineering Department
SAM Global University, Bhopal, India

Selection and peer review of this article are under the responsibility of the scientific committee of the International Conference on Current Trends in Engineering, Science, and Management (ICCSTEM-2024) at SAM Global University, Bhopal.

Abstract

The proliferation of social media platforms in business and politics has, unfortunately, facilitated the emergence of undesirable behaviours. From personal vendettas to corporate and political propaganda, engagement on social media has incentivised actions that are detrimental to individuals and society at large. The anonymity of hacked profiles has encouraged individuals to spread criticism and misinformation with impunity. This study delves into social media data to uncover the intricate relationships that underpin these problematic behaviours. Employing a sophisticated system, it identifies aberrant behaviours by detecting malicious or fraudulent accounts within social networks. Drawing on theories such as Influence, Homophily, and Balance Theory, a theoretical social framework is employed to enhance the accuracy of classifying potentially harmful users. Utilising metrics like the Jaccard coefficient, the system evaluates the similarity between user behaviours, incorporating graphical and linguistic cues to categorise end-users. The framework's efficacy is rigorously assessed using standard parameters, including confusion matrices, to gauge performance. Additionally, the study recommends employing a friend connection identification framework for enhanced social atom anomaly detection, further fortifying efforts to combat undesirable behaviour on social media platforms.

Keywords- Social Media, Social Media Mining, Fake user, Influence, Homophily, Balance Theory

1 Introduction

Detecting anomalies in social media behaviour is a significant and intellectually stimulating pursuit. Traditionally, the literature has advocated using user profiles and user-generated data to classify end-users as hostile [1]. However, this approach has limitations as it neglects the linguistic and graphical features, compromising diversity and accuracy. The conventional methodology overlooks linguistic nuances and opinion expressions, failing to capture end-users ideologies and perspectives on various subjects such as products, politics, and national and local issues. An efficient anomaly classification model for social media platforms is proposed to address these shortcomings, integrating user graphical and linguistic elements within a social theory-based relationship identification framework. This framework aims to create a nuanced understanding of user space within these platforms, thereby influencing the propagation of malicious behaviour [2]. Consequently, scalability within large networks becomes a feasible solution. Social networks serve as pivotal arenas for discussing a myriad of crucial topics. This study aims to identify social workers who can fill the void left by employers. It is observed that power users play a significant role in fostering the growth of workers within a community, particularly in contexts where public access is limited, such as within vast networks.

2 Related Work

With the proliferation of Internet technology, social media platforms like Facebook, Twitter, Google, Sina Weibo, and others have become indispensable aspects of people's daily lives, owing to their convenience, versatility, and abundant content [3]. The user base of social networks is expanding rapidly, making them attractive targets for criminals seeking to evade the law due to the vast amounts of private user information and their significant financial value. One prevalent method employed by criminals is disseminating unsolicited text messages, which poses a serious threat to the security of social networks. These spam messages encompass various types, including promotional messages for products, fraudulent reviews, and spreading rumours about trending events, all aimed at disseminating false information and compromising network security. According to a 2013 social network spam statistics report, there was a 35.5% increase in social spam messages between January and June of that year, with approximately 1 out of every 200 social texts classified as spam [4]. These

spam messages have had a discernible impact on approximately 5% of social applications, deteriorating the overall social network environment, disrupting user experience, and jeopardising user information security. Concurrently, with the rapid global expansion of social media, there has been a corresponding increase in malicious activities such as spamming, creation of bogus accounts, phishing, and dissemination of viruses. Various forms of spam, including profile spam, insider spam, eclipse spam, and outside spam, have emerged as significant challenges in social media. Profile spammers create spam, hacked, and phishing accounts using stolen user profile credentials, which can be utilised for internal and external spam activities [**?**]. Additionally, user-generated data spam manifests in various forms, including retweet spam, opinion spam, fake trend spam, and follower spam [**?**]. External engagement in spam activities encompasses downloading, financing, and product spam [5].

3 IDENTIFYING MALICIOUS PROFILE OVER SOCIAL MEDIA

This study introduces an analytical and methodological framework for identifying harmful users, which integrates implicit and explicit link connections from the perspective of end-users' social graphs. Additionally, a malicious user identification using proposed framework [6]. The framework involves classifying end-users as malevolent or legitimate users, extracting their communal information, and creating a sockpuppet node based on this information. In order to tackle the issue of noise within text content at the filter layer, the DCNN model employs a keyword-based detection method. Specifically, it utilizes the concept of Naive Bayes for analyzing word sequences $[y_1, y_2, \ldots, y_n]$ found in specific social network information. The attention mechanism employed incorporates the Yes weighting technique. The attention mechanism based on naive Bayes weight technology calculates the Naive Bayes weight of each word using formula (1) below, and then selects a certain number of keywords based on specified conditions to filter out noise.

$$s_i = \frac{(q_t^{y_i} + \alpha)/|q_t|_1}{(q_{\sim y}^{y_i} + \alpha)/|q_{s\sim}|_1} \tag{1}$$

In the filter layer, $q_t^{y_1}$ represents the number of texts that contain the word y_i in the text of the spam category y, $q_{\sim y}^{y_i}$ represents the number of texts that contain the phrase y in the text of the non-spam category s, and $|q_t|_1$ and $|q_{s\sim}|_1$ represent the number of texts in the spam and non-spam categories, respectively. The parameter α denotes the smoothing parameter. Words with significance greater than 1 are selected to enter the embedding layer.

Embedded Layer- The extraction of keywords through the attention mechanism of the filtering layer, suppose a piece of social network information is filtered from n words to k words and input into the embedding layer, denoted as $[y_1, y_2, \ldots, y_n]$. The embedding matrix is then constructed using various representation methods, and each word y_t is mapped into a real-number domain eigenvector h_t given by $h_t \in S^{f\times 1}$.

The embedding matrix $Z^{wrd} \in S^{f\times|K|}$, where K is a fixed-size dictionary set, f is the dimension of the embedding word, and Z^{wrd} needs to be learned using different representations methods such as random embedding, Skip-Gram, CBOW, and Glove. For each word y_t, it is transformed into an embedded representation h_t by matrix multiplication, as shown in formula (2) below.

$$h_t = Z^{wrd} k^t \tag{2}$$

Here, $k^t \in S^{|k|\times 1}$, where the index position of h_t is 1, and the remaining parts are 0. Finally, the feature vector of a piece of social network information formed by the embedding layer is expressed as $[h_1, h_2, h_3, \ldots, h_t]$.

Pooling Layer- Addressing the limitation of the CNN pooling strategy's inability to be dynamically updated, the DCNN model introduces a pooling strategy based on the attention mechanism. This strategy dynamically updates the weight of the attention mechanism pooling strategy, as shown in equation (4) below.

$$\begin{aligned} Z_i &= \tanh(X(Z)D_iR) \\ b_i &= \text{softmax}(Z_i v_e) \\ k_i &= D_i R b_i \end{aligned} \tag{3}$$

Here, $X^{(Z)}$ represents the attention mechanism pooling strategy matrix, Z_i represents the updated output of the i convolution kernel, v_e represents the environment vector, b_i denotes the weight of the attention mechanism pooling strategy of the i convolution kernel structure, and k_i represents the output of i convolution kernel structure of the pooling layer. Since MA-CNN adopts m types of convolution kernel structure, the characteristic of specific social network information is expressed as $K = k_1 \oplus k_2 \oplus \ldots \oplus k_m$, where $K \in S^{m\times M}$.

4 Result Analysis

The evaluation of the Proposed DNN-RIF classifier's performance involves calculating Accuracy (Acc) based on the confusion matrix parameters. Comparative analysis has been conducted to assess the performance of various classifiers, including Random Forest (RF), Bagging, J48, Random Tree (RT), Logistic Regression (LR), and the proposed DNN-RIF framework with feature fusion vector. According to Table 1, which illustrates accuracy over

the Social Media Data Set, the proposed DNN-RIF classifier demonstrates notable improvements, achieving approximately 95.89% and 98.54% Accuracy over the Crude and CCSD datasets, respectively. Figure 2 showcases these accuracies graphically, highlighting the superiority of the proposed method over existing classifiers. Specifically, in comparison to Random Forest, Bagging, J48, Random Tree, and Logistic Regression, the DNN-RIF model achieves accuracy gains ranging from 2.54% to 12.7%. Additionally, Figure 3 illustrates the accuracy of the Proposed DCNN model, further emphasizing its effectiveness. Despite these achievements, Figure 4 indicates a training loss of 1.01 percent, suggesting areas for further optimization. Nonetheless, the proposed work exhibits significant enhancements in accuracy over both Crude and CCSD datasets, with improvements ranging from 2.61% to 14.61%, underscoring its potential for practical application and performance superiority in sentiment analysis tasks.

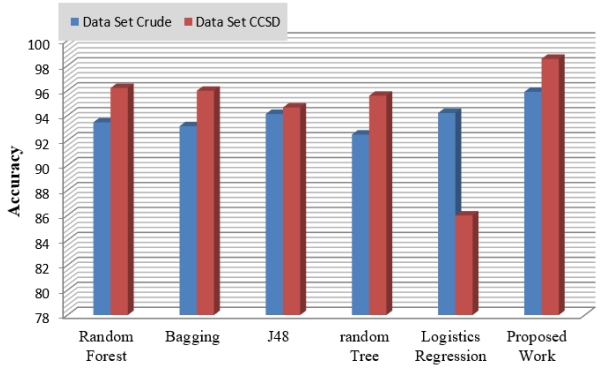

Figure 1: Accuracy of Social Media Dataset

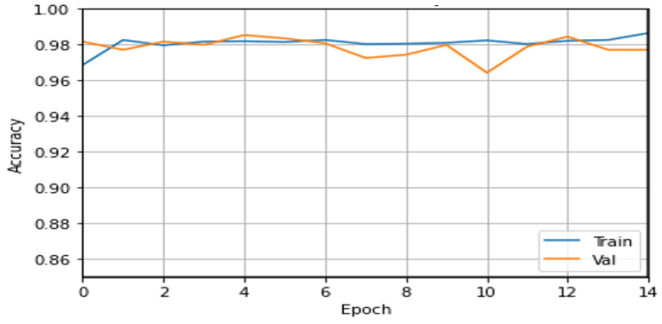

Figure 2: Accuracy of Proposed DCNN model

5 Conclusion and Future Work

The contemporary social media landscape has witnessed a significant rise in end-user participation across various domains such as business marketing, political propaganda, educational endeavors, and entertainment. While social media platforms serve as pivotal channels for marketing strategies, brand establishment, and content dissemination, they also present challenges associated with the proliferation of fraudulent activities and deceptive practices. This includes the proliferation of fake profiles, dissemination of misleading information, and the manipulation of social media users towards unlawful behavior. By leveraging communal data, this study offers insights into user behaviors and societal dynamics within these digital spaces. The findings of this research hold potential in identifying and addressing issues such as spam reviews, rumors, and fake news circulating on social networking platforms. Moreover, the implications extend to sectors such as military strategy, election campaigning, and criminology, wherein the understanding of political ideologies, prediction of criminal behavior, and analysis of societal influences can inform decision-making processes and policy formulations.

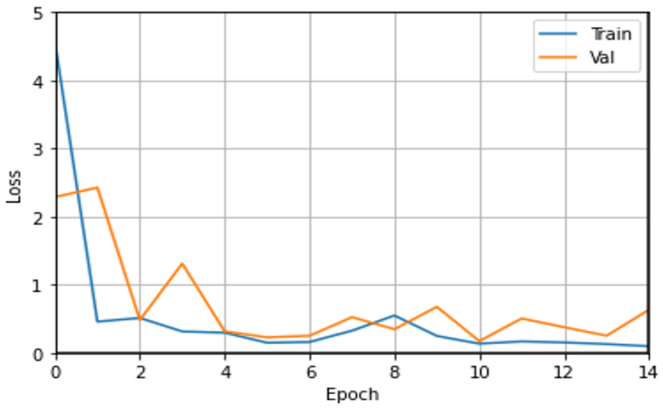

Figure 3: Training Loss of Proposed DCNN model

References

[1] A. Benamira, B. Devillers, E. Lesot, A. K. Ray, M. Saadi, and F. D. Malliaros, "Semi-supervised learning and graph neural networks for fake news detection," in *2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2019, pp. 568–569.

[2] H. Matsumoto, S. Yoshida, and M. Muneyasu, "Propagation-based fake news detection using graph neural networks with transformer," in *2021 IEEE 10th Global Conference on Consumer Electronics (GCCE)*, 2021, pp. 19–20.

[3] Z. Wang, W. Wei, X.-L. Mao, G. Guo, P. Zhou, and S. Jiang, "User-based network embedding for opinion spammer detection," *Pattern Recognition*, vol. 125, p. 108512, 2022.

[4] S. Rao, A. K. Verma, and T. Bhatia, "A review on social spam detection: Challenges, open issues, and future directions," *Expert Systems with Applications*, vol. 186, p. 115742, 2021.

ISSN: 2321-1156 Volume 12, Issue 2, March 2024

[5] J. D. R. P and W. Stalin Jacob, "Multi-objective genetic algorithm and cnn-based deep learning architectural scheme for effective spam detection," *International Journal of Intelligent Networks*, vol. 3, pp. 9–15, 2022.

[6] J. Zhang, B. Dong, and P. S. Yu, "Deep diffusive neural network based fake news detection from heterogeneous social networks," in *2019 IEEE International Conference on Big Data (Big Data)*, 2019, pp. 1259–1266.