Malicious Social Network Messages

Subjects: Computer Science, Artificial Intelligence

Contributor: Aušra Čepulionytė , Jevgenijus Toldinas , Borisas Lozinskis

The primary methods of communication in the modern world are social networks, which are rife with harmful messages that can injure both psychologically and financially. Most websites do not offer services that automatically delete or send malicious communications back to the sender for correction, or notify the sender of inaccuracies in the content of the messages. The deployment of such systems could make use of techniques for identifying and categorizing harmful messages.

social network messages malicious messages Internet

1. Introduction

The Internet is increasingly essential for interacting with others and exchanging information as a single platform in which users leave digital records of their own behavior and usage patterns, which, if properly analyzed, might provide crucial information about user behavior [1][2]. Tweets, blogs, chat messaging, and other forms of social media are the primary means of community communication today, and Internet-based crime and can be used offline to investigate crimes or, in real time, to prevent them [3]. According to Stacy Jo Dixon, in the first quarter of 2023, Facebook had approximately three billion active monthly users [4]. The amount of financial harm caused by cybercrime that was reported to the Internet Crime Complaint Center (IC3) rose considerably between 2001 and 2022. The annual financial loss as a result of complaints referred to the IC3 increased from USD 6.9 billion to USD 10.3 billion in the most recent reported period [5].

Social media platforms must understand the fundamentals of human social interaction and find simple, effective ways to maintain the necessary standards of confidentiality, security, and reliability. To use and manipulate the vast amount of information on the social web, governments, intelligence agencies, and technical specialists must step forward and try to adopt new technologies and paradigms [6]. Sentiment analysis is a way of determining a text's sentiment polarity, which is used to identify whether the text is conveying a positive or negative message [7]. To automatically determine the sentiment polarity of a comment is the aim of the sentiment classification of online social networks (OSNs). It requires investigation into handling emotional ranges to achieve a better interpretation of OSN messages, because messages can have a range of sentiments in addition to positive or negative ones, including neutral and neutral with gradations [8].

Machine learning approaches make it easier to develop models from sample data, speeding up decision-making processes based on real-world inputs. These techniques allow learning from input data via descriptive statistics as well as production values within a predetermined range [9]. Input data from a batch or the real-time collection of

data instances are needed for machine learning algorithms to train their models. The terms “data point,” “vector,” “event,” “sample,” “case,” “object,” “record,” “observation,” and “entity” can all be used to describe a single datum instance [10]. Unlabeled data are utilized in unsupervised learning since it lacks additional information while labeled data have useful tags and are used in supervised learning. Benchmark datasets are used in machine learning for model accuracy comparisons and performance measures.

2. Malicious Social Network Messages

Based on data that can be found on social networks, the information is separated into four categories: hyperlinks, images, audio, and text (a subset of spoken language primarily produced with a text or string to examine the content) [11]. OSNs are receiving attention from users who are malicious or abnormal and engage in malicious activities such as harassing others, plotting attacks (in which terrorists may be involved), and disseminating false information [12]. Spam is the term for unsolicited messages that are sent in large quantities by fostering a sense of community trust. Spammers engage in illegal acts including phishing, advertising, surveillance, assault against women, and cyberbullying, among others [13]. Instead of using legitimate accounts, spammers typically distribute spam using fraudulent, compromised, or cloned accounts, crowd-sourcing strategies, and automated bots [14]. The taxonomy of various social spam detection techniques and approaches are observed as follows: URL list-based spam filtering (Blacklist, Whitelist, Greylist), honeypot and honeynet-based spam detection, and machine learning (ML) and deep learning (DL)-based social spam detection. ML and DL are used for social spam content detection including malicious URL detection [15] and text-based spam detection [16][17].

Social media bots (SMBs) are tools that people and organizations employ to spread information, expand their reach, and boost their impact. Malicious bots can annoy or burden users by participating in unethical actions, including stealing the identities of real users, persuading voters to favor politicians, spreading hate speech, and other divisive material [18]. SMBs are classified into three main groups: benign bots, neutral bots, and malicious bots. For SMB detection, the most used ML methods are random forest (RF), SVM, and AdaBoost, while LSTM and CNN are the two most widely used DL algorithms; unfortunately, there is a lack of large datasets to train models [18].

In [19], bidirectional encoder representations from transformers (BERT) are proposed. It has been shown that the pre-training of linguistic models is effective in improving many tasks related to the processing of natural languages, including the intuition and paraphrase of natural languages, the recognition of named entities and, the answer to questions. The development of pre-trained language models based on transformer architectures has stimulated the evolution of modern techniques for many tasks in the field of natural language processing (NLP) [20][21][22]. The study of [23] proposed text classification using BERT for natural language processing and the results of the experiment showed that combinations of BERT with CNN, RNN, and BiLSTM performed well with precision, recall rate, and F1 score, compared to Word2vec. The new BSTC (BERT, SKEP, and TextCNN) fake review detection model is proposed [24] based on a pre-trained language model and a convolutional neural network. The highest accuracy was achieved with all three gold standard datasets (Hotel, Restaurant, and Doctor), with 93.44%, 91.25%, and 92.86%, respectively. The process of choosing, modifying, and transforming raw data into features

that can be utilized to enhance the performance of machine learning models is known as feature engineering. In some tasks, effective feature engineering combined with conventional machine learning methods could produce outcomes comparable to BERT [25]. Although there has been a rise in interest in learning general-purpose sentence representations, the majority of the research in that field has been conducted in English and has mostly been monolingual [26].

Spam is typically defined as undesired text that is sent or received over social media platforms like Facebook, Twitter, YouTube, e-mail, etc. [27]. The authors of [28] proposed a novel four-layered, state-of-the-art detection strategy, with graph-based, neighbor-based, automation-based, and time-based features to find spammers on social networking sites. The majority of SMS spam classifiers use supervised algorithms like Naïve Bayes (NB), support vector machine (SVM), neural networks, and regression, because the availability of the output column (labeled data) of the SMS dataset makes it possible to train classification models [29]. Using a total of 20 samples from the dataset (SMS Spam Corpora and Twitter Corpora), the suggested solution in [30] employs reinforcement learning to identify the malicious social bots. It also makes use of k-nearest neighbor (KNN) and a recurrent neural network (RNN). A social bot is a computer program that uses an application programming interface (API) to operate a social media account. It can be used for malicious activities, such internet trolling and fraud. Bots are classified as malicious or benign in the study cited [31].

Information phishing began as a marketing tactic, but it has since evolved into destructive internet interactions that expose users to significant security risks using tools including emails, comments, blogs, and messaging. Given their adaptability and ability to make the most of current hardware and computational limitations, deep learning architectures like convolutional neural networks (CNNs), multi-layer perceptrons (MLPs), and the long short-term memory (LSTM) have been successfully used for email spam classification [32]. The identification of fake news [33] [34] is a difficult challenge for social media platforms like Facebook, Twitter, etc., because of the volume of data that people publish on these sites. To determine whether a news article is authentic or fake, a deep CNN for fake news detection was presented in [35] and models were tested using binary class datasets. For NLP researchers, sarcasm presents a formidable challenge and can entirely alter the meaning of a statement, making it challenging for modern models and systems to recognize it. In order to create models that can accurately identify the settings in which sarcasm may occur or is suitable, an approach for the automatic detection of sarcasm context has been developed [36].

Cyber social media security examines the dynamics of online social networks, the data's vulnerabilities, and the potential effects of their abuse by social media attackers. Due to their nature, the volume of content they include, and the sensitive information they use, social media are the most attack-prone section of the internet [37][38][39]. To classify a social media message as a part of a particular crisis event, it is important to take into account a number of factors, such as the message's nature, the information it contains, the source of that information, its credibility, the timing, and its location [40]. Some of the features can be automatically extracted, whereas some need to be manually labeled. The best performance is achieved with an ensemble approach for the identification and classification of crime-related tweets that uses logistic regression (LR), SVMs, KNN, a decision tree (DT), and an RF classifier assigned the weights of 1, 2, 1, and 1, respectively, ensemble together via a soft weighted voting

classifier along with a term frequency–inverse document frequency (TF-IDF) vectorizer with an accuracy of 96.2% on the testing dataset [41]. When compared to the ground truth labeled by network experts, an RNN-LSTM model that was trained to identify five different social engineering attacks (SEA) that may show signs of information gathering achieves classification precision and recall scores of 0.84 and 0.81, respectively [42].

References

1. Luna, S.; Pennock, M.J. Social media applications and emergency management: A literature review and research agenda. Int. J. Disaster Risk Reduct. 2018, 28, 565–577.
2. Bhattacharjee, S.D.; Tolone, W.J.; Paranjape, V.S. Identifying malicious social media contents using multi-view Context-Aware active learning. Future Gener. Comput. Syst. 2019, 100, 365–379.
3. Soomro, T.R.; Hussain, M. Social Media-Related Cybercrimes and Techniques for Their Prevention. Appl. Comput. Syst. 2019, 24, 9–17.
4. Dixon, S. Social Media-Statistics & Facts. Available online: https://www.statista.com/topics/1164/social-networks/#topicOverview (accessed on 20 July 2023).
5. Statista. Cyber Crime: Reported Damage to the IC3 2022. Available online: https://www.statista.com/statistics/267132/total-damage-caused-by-by-cyber-crime-in-the-us (accessed on 20 July 2023).
6. Thakur, K.; Hayajneh, T.; Tseng, J. Cyber Security in Social Media: Challenges and the Way Forward. IT Prof. 2019, 21, 41–49.
7. Wanda, P.; Huang, J. Model of Sentiment Analysis with Deep Learning in Social Network Environment. In Proceedings of the 2nd International Conference on Electronic Information and Communication Technology (ICEICT), Harbin, China, 20–22 January 2019.
8. Wanda, P.; Jie, H.J. DeepSentiment: Finding Malicious Sentiment in Online Social Network based on Dynamic Deep Learning. IAENG Int. J. Comput. Sci. 2019, 46, 616–627.
9. Mishra, S.; Shukla, P.; Agarwal, R. Analyzing Machine Learning Enabled Fake News Detection Techniques for Diversified Datasets. Wirel. Commun. Mob. Comput. 2022, 2022, 1575365.
10. Toshniwal, A.; Mahesh, K.; Jayashree, R. Overview of Anomaly Detection techniques in Machine Learning. In Proceedings of the Fourth International Conference on I-SMAC, Palladam, India, 7–9 October 2022.
11. Kondamudi, M.R.; Sahoo, S.R.; Chouhan, L.; Yadav, N. A comprehensive survey of fake news in social networks: Attributes, features, and detection approaches. J. King Saud Univ.-Comput. Inf. Sci. 2023, 35, 101571.

12. Sharma, K.; Singh, A. A Systematic Review: Detection of Anomalies in Social Networks. In Proceedings of the International Conference on Sustainable Computing and Data Communication Systems (ICSCDS), Erode, India, 23–25 March 2023.
13. Koggalahewa, D.; Xu, Y.; Foo, E. An unsupervised method for social network spammer detection based on user information interests. J. Big Data 2022, 9, 7.
14. Rao, S.; Verma, A.K.; Bhatia, T. A review on social spam detection: Challenges, open issues, and future directions. Expert Syst. Appl. 2021, 186, 115742.
15. Al-Haija, Q.A.; Al-Fayoumi, M. An intelligent identification and classification system for malicious uniform resource locators (URLs). Neural Comput. Appl. 2023, 35, 16995–17011.
16. Martinez-Romo, J.; Araujo, L. Detecting malicious tweets in trending topics using a statistical analysis of language. Expert Syst. Appl. 2013, 40, 2992–3000.
17. Almutlaq, R.; Hafez, A. Detection Mechanism for Malicious Messages on KSU Student Social Network. Int. J. Data Sci. Technol. 2020, 6, 23–36.
18. Ellaky, Z.; Benabbou, F.; Ouahabi, S. Systematic Literature Review of Social Media Bots Detection Systems. J. King Saud Univ. Comput. Inf. Sci. 2023, 35, 101551.
19. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv 2019, arXiv:1810.04805.
20. Pattanaik, B.; Mandal, S.; Tripathy, R.M. A survey on rumor detection and prevention in social media using deep learning. Knowl. Inf. Syst. 2023, 65, 3839–3880.
21. Zhang, X.; Malkov, Y.; Florez, O.; Serim Park, S.; McWilliams, B.; Han, J.; El-Kishky, A. TwHIN-BERT: A Socially-Enriched Pre-trained Language Model for Multilingual Tweet Representations. arXiv 2022, arXiv:2209.07562.
22. Dai, Z.; Yang, Z.; Yang, Y.; Carbonell, J.; Le, Q.V.; Salakhutdinov, R. Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context. arXiv 2019, arXiv:1901.02860.
23. Bello, A.; Ng, S.-C.; Leung, M.-F. A BERT Framework to Sentiment Analysis of Tweets. Sensors 2023, 23, 506.
24. Lu, J.; Zhan, X.; Liu, G.; Zhan, X.; Deng, X. BSTC: A Fake Review Detection Model Based on a Pre-Trained Language Model and Convolutional Neural Network. Electronics 2023, 12, 2165.
25. Gani, R.; Chalaguine, L. Feature Engineering vs BERT on Twitter Data. arXiv 2022, arXiv:2210.16168.
26. Lample, G.; Conneau, A. Cross-lingual Language Model Pretraining. arXiv 2019, arXiv:1901.07291.

27. Kaddoura, S.; Chandrasekaran, G.; Popescu, D.E.; Duraisamy, J.H. A systematic literature review on spam content detection and classification. PeerJ Comput. Sci. 2022, 8, e830.
28. Bankar, S.H.; Shinde, S.A. Spammer Detection of Social Networking Sites Using 4 Novel Techniques. Available online: https://www.academia.edu/download/34105340/Sachin_Bankar.pdf (accessed on 20 July 2023).
29. Odera, D.; Odiaga, G. A comparative analysis of recurrent neural network and support vector machine for binary classification of spam short message service. World J. Adv. Eng. Technol. Sci. 2023, 9, 127–152.
30. Kumar, R.M.; Bharathi, P.S. Detection of Malicious Social Bots with reinforcement learning technique with URL Features in Twitter Network with KNN in comparison with RNN. In Proceedings of the Eighth International Conference on Science Technology Engineering and Mathematics (ICONSTEM), Chennai, India, 6–7 April 2023.
31. Mbona, I.; Eloff, J.H.P. Classifying social media bots as malicious or benign using semi-supervised machine learning. J. Cybersecur. 2023, 9, tyac015.
32. Baccouche, A.; Ahmed, S.; Sierra-Sosa, D.; Elmaghraby, A. Malicious Text Identification: Deep Learning from Public Comments and Emails. Information 2020, 11, 312.
33. Alkhodair, S.A.; Ding, S.H.H.; Fung, B.C.M.; Liu, J. Detecting breaking news rumors of emerging topics in social media. Inf. Process. Manag. 2020, 57, 102018.
34. Meel, P.; Vishwakarma, D.K. Fake news, rumor, information pollution in social media and web: A contemporary survey of state-of-the-arts, challenges and opportunities. Expert Syst. Appl. 2020, 153, 112986.
35. Kaliyar, R.H.; Goswami, A.; Narang, P.; Sinha, S. FNDNet—A deep convolutional neural network for fake news detection. Cogn. Syst. Res. 2020, 61, 32–44.
36. Băroiu, A.-C.; Trăușan-Matu, Ș. Comparison of Deep Learning Models for Automatic Detection of Sarcasm Context on the MUStARD Dataset. Electronics 2023, 12, 666.
37. Sharma, S.; Jain, A. Role of sentiment analysis in social media security and analytics. WIREs Data Min. Knowl. Discov. 2020, 10, 5.
38. Lippmann, R.P.; Campbell, W.M.; Weller-Fahy, D.J.; Mensch, A.C.; Zeno, G.M.; Campbell, J.P. Finding malicious cyber discussions in social media. Linc. Lab. J. 2016, 22, 46–59. Available online: https://apps.dtic.mil/sti/citations/AD1034416 (accessed on 3 August 2023).
39. Rahman, M.S.; Halder, S.; Uddin, M.A.; Acharjee, U.K. An efficient hybrid system for anomaly detection in social networks. Cybersecurity 2021, 4, 10.
40. Krishna, Y.V.; Jahnavi, G.; Tharun, M.; Yegineti, S.G.; Raja, G.; Suneetha, B. Survey: Analysis of Security Issues on Social Media using Data Science techniques. In Proceedings of the

International Conference on Inventive Computation Technologies (ICICT), Lalitpur, Nepal, 26–28 April 2023.

41. Siddiqui, T.; Hina, S.; Asif, R.; Ahmed, S.; Ahmed, M. An ensemble approach for the identification and classification of crime tweets in the English language. Comput. Sci. Inf. Technol. 2023, 4, 149–159.
42. Aun, Y.; Gan, M.; Wahab, N.H.B.A.; Guan, G.H. Social engineering attack classifications on social media using deep learning. Comput. Mater. Contin. 2023, 74, 4917–4931.