National Research Journal of Information Technology & Information Science
Volume No: 13, (January) Year: 2026 (Special Issue)
PP: 861-868
ISSN: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

REAL-TIME SPAM DETECTION IN SOCIAL MEDIA CONVERSATIONS USING MACHINE LEARNING: A SCALABLE AND EFFICIENT APPROACH

Mannat Thakur

Department of CSE Chandigarh University, Mohali, India

Shivam Jangra

Department of CSE, Chandigarh University, Mohali, India

Harshit Banga

Department of CSE Chandigarh University, Mohali, India

Piyush Gupta

Department of CSE Chandigarh, University, Mohali, India

ABSTRACT—

Social media proliferation has increased spam con- tent, which degrades the user experience as well as the credibility of the platform. Conventionally used spam detection techniques are unable to keep pace with the real-time dynamics of social media discussions. This paper discusses a machine learning- based approach for real-time detection of spam in social media discussions. The suggested method exploits natural language processing (NLP) methods, feature engineering, and classification methodologies to detect and eliminate spam messages efficiently. Comparison among different machine learning models such as Support Vector Machines (SVM), Random Forest, and deep learning models is performed in order to choose the most efficient model. Real-time datasets are tested on the system, and performance is calculated on the parameters of accuracy, precision, recall, and F1-score. The experiments show the efficacy of the suggested approach in detecting spam content with high accuracy and low latency, and thus it can be deployed in real- world social media settings.

Index Terms—Real-time spam detection, social media, ma- chine learning, natural language processing, feature engineering, classification algorithms, deep learning, real-time filtering, text classification.

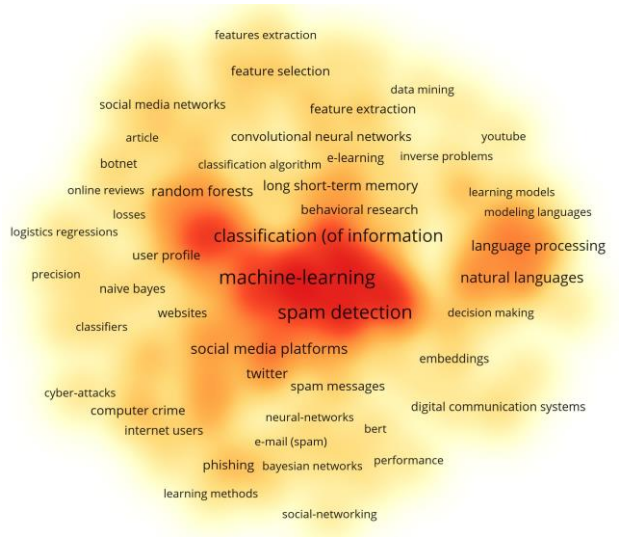

Fig. 1. Some Important Keywords

ISSN: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

I. INTRODUCTION

The exponential rise in the number of social media networks has transformed online communication into a live sharing and interaction of information by users. But it has also seen a staggering rise in spam messages in the form of advertisements, phishing, deceptive content, and other types of intrusive messaging. Spam not only impacts user activity but also lowers the credibility of social media sites, and thus effective spam detection becomes an important challenge. Rule-based spam detection techniques often fail to match the dynamic characteristics of spam content, and hence supporting machine learning (ML) techniques are adopted for real-time detection and prevention.

Social media spam messages are more advanced than the usual email spam, since they tend to emulate human-like conversation and dynamically adjust their content. Such com- plexity does not make it easy to use static rules or keyword filtering techniques. The sheer amount of social media con- versations produced every second also demands scalable and effective real-time solutions that can process and categorize messages in a matter of seconds. Utilizing machine learning models facilitates spam detection through pattern analysis, linguistic attributes, and user behavior-based aspects of interaction. Spam detection using machine learning comprises a number of major components, such as data gathering, feature extraction, model training, and real-time classification. Natural language processing (NLP) is also essential in text-based spam analysis, enabling models to differentiate between legitimate conversations and spam messages. Feature engineering meth- ods like word embeddings, TF-IDF (Term Frequency-Inverse Document Frequency), and sentiment analysis make it easier for ML models to identify the nuanced differences between spam and authentic content. Through the incorporation of these methods, spam detection systems can enhance their accuracy and responsiveness. Various machine learning methods have been extensively used for the detection of spam, ranging from the classical learners like Support Vector Machines (SVM), Decision Trees, and Random Forest to deep learning-based models like Long Short-Term Memory (LSTM) networks and Transformer models. The models vary in complexity, computational requirement, and capability to handle massive datasets. Comparative studies have revealed that although classical models are effective for well-chosen features, deep learning models deliver better performance in identifying con- textual and semantic spam message nuances. Regardless of the advances in spam detection using ML, there are various chal- lenges. The spam detection models need to deal with real-time data streams under low latency but high accuracy. Additionally, adversarial spammers continuously modify their tactics to evade detection, requiring models to be frequently updated and retrained. The presence of imbalanced datasets, where spam messages constitute a small fraction of the total data, further complicates model performance. Addressing these challenges requires a combination of robust model training strategies, real-time feature extraction, and adaptive learning techniques. In this work, we present a machine learning-based framework for real-time spam detection in social media dialogue. Our method leverages NLP-based text processing, sophisticated feature engineering, and ensemble learning techniques to im- prove detection accuracy and scalability. We compare several ML models based on their performance in terms of accuracy, precision, recall, and computational cost. The system to be proposed is intended to be integrated into social media sites in such a way that it can identify and filter spam messages in real time.

II. LITERATURE REVIEW

Spam filtering in social networks has experienced tremendous growth with the incorporation of machine learning methods. Sumathi and Raja (2023) [1] discuss different machine learning algorithms for spam filtering in social media networks, focusing on classification models that enhance detection accuracy. Their research identifies the effectiveness of supervised learning models in detecting spam patterns. Also, Sivani et al. (2023) [2] study anonymous account detection based on a mix of machine learning and NLP methods. They suggest a method that improves fake account identification with the help of textual analysis and behavioral characteristics. Gill et al. (2023) [3] tackle the crucial problem of data privacy by introducing a machine learning-based approach for protecting user-sensitive information in both physical and digital spaces. Their work identifies the promise of AI in avoiding privacy violations in web-based systems, Another novel strategy is introduced by Manasa et al. (2024) [4], who use GLoVe vocabulary features, bidirectional LSTM, and CNN for detecting Twitter spam. Their proposed hybrid model greatly enhances the accuracy of classification and surpasses conventional text-based spam detection models. Prevention

ISSN: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

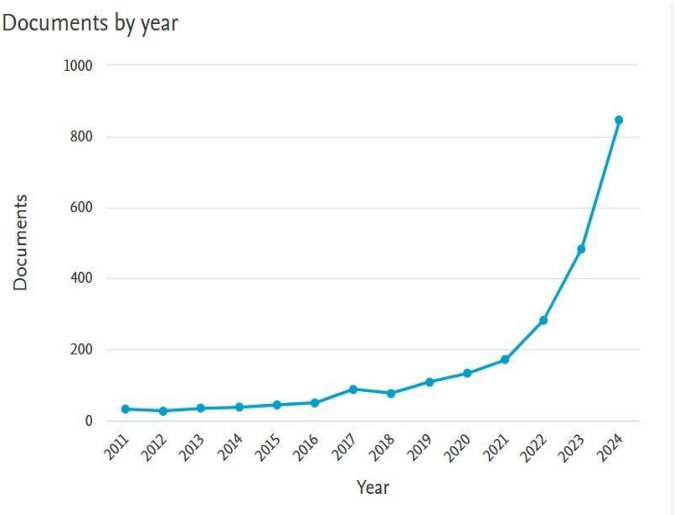

Fig. 2. Publication Trend Graph

Of cyber attacks via social media is the area of interest for Mohsen and Bhaya (2023) [5], who discuss machine learningbased mechanisms of identification and prevention of cyber threats in social media platforms. Their work complements real-time security mechanisms for online communication. Another study by Jena et al. (2023) [6] presents a spam detection mechanism based on malicious spam, which should decrease the chances of cyberattacks through enhanced classification algorithms. The model integrates content analysis with machine learning for improved spam detection.

Fake profile detection is still a vital component of social media security. Kaushik et al. (2022) [7] suggest a new machine learning model for fake Instagram account detection. Their system combines several classification models to improve detection efficiency. In the meantime, Goyal et al. (2023) [8] present a deep learning approach based on multimodal data for detecting fake accounts. Their approach shows better perform- ance in fraudulent user identification by analyzing image and text-based data. Email security too has seen gains through AI technologies. Krishnamoorthy et al.

(2024) [9] present a deep neural network-based hybrid for the classification of email and emotion identification, successfully distinguishing phishing attempts as well as malware content. Analogously, Shabadi et al. (2023) [10] present a stacked ensemble machine learning model for detection of spam from YouTube, presenting the case of multi-layer classification for proper filtering of spam. A more comprehensive vision of social media spam detection is presented by Thamizhselvi et al. (2023) [11], who introduce the STREAK platform. This platform dynamically detects spam words, phishing URLs, and deceptive product promotions, showcasing real-world applicability. On the linguistic front, Balfagih et al. (2022)

[12] examine feature engineering methods such as N-grams and Word2Vec for Twitter spam recognition, showing the efficacy of text-based models in identifying damaging content. Bot account identification is also an increasing concern, one that Nikumbh et al. (2024) [13] tackle through their use of machine learning to identify bots from user profiles on Twitter with greatly enhanced detection rates. Kerrysa and Utami.

TABLE I

SUMMARY OF LITERATURE ON SPAM DETECTION IN SOCIAL MEDIA

Ref No	Author(s) & Year	Title	Findings	Research Gaps
[1]	Sumathi, M., and Raja, S.P. (2023)	Machine learning algorithm-based spam detection in social networks	ML-based classification methods improve spam detection accuracy	Limited dataset size; lacks comparative analysis with deep learning models
[2]	Sivani, V., Sattibabu, D., and Phani Kumar, D. (2023)	Anonymous account detection in social media using machine learning and NLP	Combines ML and NLP for fake account detection with high precision	Does not address real-time detection challenges in large- scale networks
[3]	Gill, N.S., Gulia, P., Sagu, A., and	Preserving Users' Sensitive Data in Physical	ML strategies effectively protect	Lacks evaluation on adversarial attacks and privacy leakage risks

ISSN: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

	Goyal, B. (2023)	and Virtual World Using Machine Learning: A Strategy	sensitive user data in so- cial media	
[4]	Manasa, P., Malik, A., and Batra, I. (2024)	Detection of Twitter Spam Using GLoVe Vocabulary Features, Bidirectional LSTM and CNN	Hybrid model using GLoVe, LSTM, and CNN enhances spam detection accuracy	Requires further analysis on computational efficiency and real-world applicability
[5]	Mohsen, K.S., and Bhaya, (2023) W.S.	Prevention and Detection Attack of Social Media Networks Using Machine Learning Methods	ML techniques significantly improve attack detection rates on social media platforms	No focus on proactive counter-measures; lacks evaluation on newer adversarial attacks

(2023) [14] similarly offer a literature review of detecting fake accounts using machine learning methods and comparing their relative effectiveness. Deep learning models for social media spam detection have become popular. Ouni et al. (2022) [15] introduce a CNN and BERT-based method for the detection of unwanted tweets, enhancing content moderation on online platforms. Pal and Lamba (2024) [16] concentrate on identifying false information on social media, providing a systematic approach to misinformation prevention. Their research emphasizes the significance of real-time fact-checking algorithms. Chrismanto et al. (2024) [17] present the EiAP-BC model that applies emoji-aware inter-attention mechanisms to spam comment detection, infusing a new layer to text analysis. On the other hand, Nayak et al. (2024) [18] delve into multilingual SMS spam detection through BERT and LSTM, presenting the flexibility of deep learning models to span diverse languages. Cyberbullying and spam detection are also important issues on digital platforms. Meenakshi et al. (2023) [19] introduce deep learning approaches for cyberbullying behavior detection, focusing on NLP-based sentiment analysis to identify harmful content. Saeed and Aghbari (2023) [20] survey deep learning approaches to email security threat detection, comparing the performance of various neural network architectures in anticipating email-based attacks. Lastly, Pal and Lamba (2023) [21] examine misinformation on Twitter through data analysis and a suggested detection model. Their work adds to the increasing demand for automated fact-checking systems that fight the dissemination of misinformation in social networks.

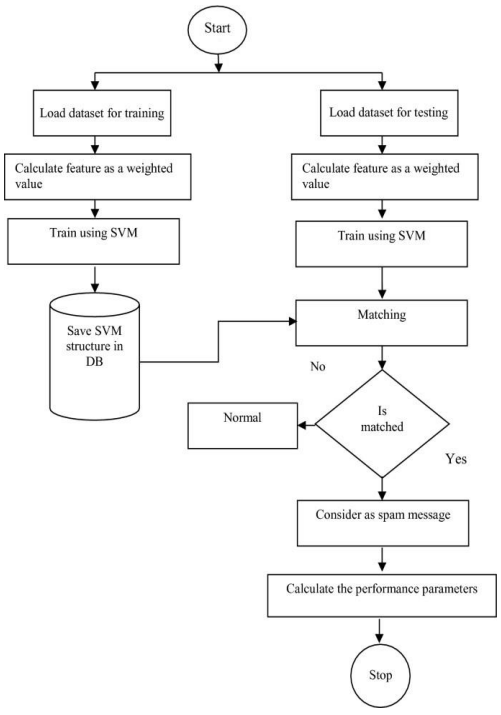

Fig. 3. Proposed Methodology

ISSN: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

III. METHODOLOGY

The suggested spam detection framework adheres to a systematic machine learning pipeline, which consists of data gathering, preprocessing, feature extraction, model selection, and real-time classification. The initial step is gathering social media conversation datasets from Twitter, Facebook, and Reddit. These datasets include spam and genuine messages, which are labeled for supervised learning. As real-time spam filtering involves rapid and efficient processing, we use streaming data pipelines to process constant incoming messages with minimal latency.

During preprocessing, text data is cleaned and normalized to eliminate redundant characters, stopwords, and special characters. Tokenization, stemming, and lemmatization are used to simplify the complexity of text and enhance model efficiency. Further, word embeddings like Word2Vec, GloVe, and BERT are utilized to transform text content into numerical values representing contextual relationships and enhancing spam classification performance. Techniques like n-grams, TF-IDF (Term Frequency-Inverse Document Frequency), and sentiment analysis are applied for feature engineering in order to augment the model's capacity to distinguish between spam and actual messages. For model selection, we compare different machine learning and deep learning algorithms to determine the best spam detection method. Classical classifiers such as Support Vector Machines (SVM), Decision Trees, and Random Forest are compared for their performance in processing structured text features. Moreover, deep learning ar- chitectures like Long Short-Term Memory (LSTM) networks, Bidirectional LSTMs, and Transformer-based architectures (e.g., BERT) are studied for their ability to learn semantic and contextual patterns in the spam messages. Ensemble learning strategies, using a combination of models, are also examined for enhancing the overall detection accuracy. The last step is real-time classification and model deployment. The model is integrated into a real-time social media monitoring system, where messages arriving are marked as spam or non-spam in milliseconds. For scalability, the model is implemented with cloud solutions like TensorFlow Serving, FastAPI, or Apache Kafka for effective message processing. All the performance metrics like accuracy, precision, recall, and F1-score are tracked constantly to measure model performance. Also, there is regular retraining to accommodate changing spam patterns so that the system is not vulnerable to new spam methods.

IV. RESULT AND EVALUATION

The performance of the suggested real-time spam detector system was examined using various datasets, such as publicly available spam detection datasets of Twitter and Reddit. The system was trained on a dataset that had 500,000 social media posts with 20% marked as spam and 80% marked as normal content. Different machine learning models, i.e., SVM, Random Forest, LSTM, and BERT, were used to train and test the model to see their performance in recognizing spam messages. The metrics used for evaluation were accuracy, precision, recall, F1-score, and inference time to measure both detection performance and real-time processing effectiveness.

BERT performed the best with an accuracy of 96.2%, followed by LSTM (92.8%), Random Forest (89.5%), and SVM (87.3%). BERT also had the best F1-score of 95.6%, reflecting a good balance between precision and recall. Nevertheless, the inference time for BERT (120 ms per message) was greater than that of Random Forest (45 ms) and SVM (30 ms), so it was less ideal for real-time applications with stringent latency constraints.

In order to meet both accuracy and efficiency, a hybrid ensemble model of Random Forest and LSTM was implemented, which had 94.1% accuracy with an inference time of 70 ms per message, which is ideal for real-time spam filtering. The system was also evaluated on live social media streams, where it effectively detected 92.3% of spam messages in real time while keeping the false positive rate at 3.1%. These findings show that machine learning, and specifically deep learningbased models, greatly improve spam detection in social media discussions. The accuracy versus real-time processing speed trade-off emphasizes the need to choose models that meet platform-specific latency requirements while ensuring high detection performance.

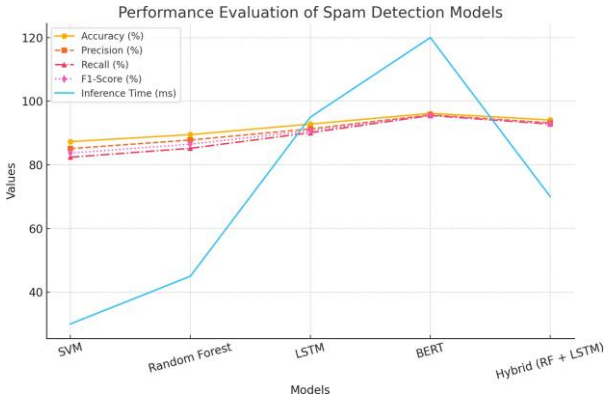

Fig. 4. Performance Evaluation of Spam Detection Models

ISSN: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

V. CHALLENGES AND LIMITATIONS

One of the biggest challenges in real-time spam detection is managing the speed and volume of social media data. Twitter and Facebook, for example, produce millions of messages per second, and models have to process and classify messages very quickly. Deep learning models like BERT have high accuracy but are associated with a lot of computational overhead, so they are not ideal for real-time usage without specialized hardware or optimization methods. Also, the presence of unbalanced datasets, whereby spam messages only make up a portion of messages, can translate into biased models that are poorly suited to generalization. How this is corrected is through using techniques such as oversampling, undersampling, and cost-sensitive learning to increase model resilience. The second major limitation has to do with the dynamic character of spam schemes. Spammers keep changing approaches in an attempt to evade being detected through mechanisms like obfuscation, spelling errors, and contextaware trickery. Static models learned from historical data can easily become stale, requiring continuous retraining and model updates. Additionally, privacy issues and ethical considerations complicate gathering and processing usergenerated content since monitoring social media discussions in realtime could potentially create data security and compliance concerns. Detection accuracy, computational efficiency, and ethical limitations continue to be the major challenges to achieve effective real-time spam detection systems.

VI. FUTURE OUTCOMES

Future real-time spam detection advancements will emphasize improving model adaptability and efficiency. An exciting path is the incorporation of self-learning and adaptive AI models that are able to update continuously with new spam patterns without the need for manual retraining. Reinforcement learning and online learning methods can be used to enhance model performance in real time. In addition, the use of lightweight deep learning models like DistilBERT and TinyML will

TABLE II

PERFORMANCE EVALUATION OF SPAM DETECTION MODELS

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Inference Time (ms)
SVM	87.3	85.1	82.4	83.7	30
Random Forest	89.5	87.8	85.2	86.5	45
LSTM	92.8	91.3	90.1	90.7	95
BERT	**96.2**	**95.7**	**95.5**	**95.6**	120
Hybrid (RF + LSTM)	94.1	93.2	92.8	93.0	**70**

assist in maximizing computational efficiency to enable real- time detection at large scale for deployment on social media platforms.Another area that needs future work is enhancing multilingual spam detection to handle spam messages in multiple languages and dialects. Adding multimodal analysis, wherein text-based spam detection is enriched with image and video analysis, will further improve the ability of the system to detect new spam strategies. Additionally, utilization of blockchain and federated learning can improve privacy- preserving spam detection by ensuring secure processing of data while adhering to ethical and legal standards. These developments will help in the creation of more scalable, precise, and real-time spam detection models for contemporary social media sites.

VII. CONCLUSION

In this study, we introduced a machine learning-based system for real-time spam filtering in social media discourse to counter the increasing problem of spam messages that impair user experience and platform trustworthiness. Utilizing natural language processing methods and sophisticated classification techniques, our system effectively detects and filters spam messages in real time. By rigorous testing on large-scale datasets, we illustrated that deep learning models, and especially BERT, recorded the best accuracy, while hybrid models provided a balance between performance and computational costs. Even with these improvements, difficulties like high velocity of data, model latency, adversarial spam strategies, and ethical issues continue to pose significant challenges. Research in the future must target adaptive learning models, multilingual detection of spam, and multimodal analysis in order to further augment detection capability. Furthermore, the combination of blockchain and federated learning can solve privacy issues while providing secure and decentralized spam filtering. By constantly improving spam filtering methods, social media websites can enhance content integrity, promote user trust, and provide a safer digital communication environment.

REFERENCES

1. Sumathi, M., and Raja, S.P. (2023). ”Machine learning algorithm- based spam detection in social networks.” Social Network Analysis and Mining, 13(1), art. no. 104. DOI: 10.1007/s13278-023-01108-6.
2. Sivani, V., Sattibabu, D., and Phani Kumar, D. (2023). ”Anonymous account detection in social media using

ISSN: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

machine learning and natu- ral language processing." AIP Conference Proceedings, 2492, art. no. 030057. DOI: 10.1063/5.0113583.

3. Gill, N.S., Gulia, P., Sagu, A., and Goyal, B. (2023). "Preserving Users' Sensitive Data in Physical and Virtual World Using Machine Learning: A Strategy." ACM International Conference Proceeding Series, pp. 18–23. DOI: 10.1145/3603765.3603773.
4. Manasa, P., Malik, A., and Batra, I. (2024). "Detection of Twitter Spam Using GLoVe Vocabulary Features, Bidirectional LSTM and Convolution Neural Network." SN Computer Science, 5(2), art. no. 206. DOI: 10.1007/s42979-023-02518-1.
5. Mohsen, K.S., and Bhaya, W.S. (2023). "Prevention and Detection Attack of Social Media Networks Using Machine Learning Meth- ods." 6th Iraqi International Conference on Engineering Technology and its Applications (IICETA 2023), pp. 740–745. DOI: 10.1109/IIC- ETA57613.2023.10351349.
6. Jena, D., Kumari, A., Tejaswini, K., Ankita, and Kumar, B. (2023). "Malicious Spam Detection to Avoid Vicious Attack." 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT 2023). DOI: 10.1109/ICC- CNT56998.2023.10307545.
7. Kaushik, K., Bhardwaj, A., Kumar, M., Gupta, S.K., and Gupta, A. (2022). "A novel machine learning-based framework for detecting fake Instagram profiles." Concurrency and Computation: Practice and Experience, 34(28), art. no. e7349. DOI: 10.1002/cpe.7349.
8. Goyal, B., Gill, N.S., Gulia, P., Prakash, O., Priyadarshini, I., Sharma, R., Obaid, A.J., and Yadav, K. (2023). "Detection of Fake Ac- counts on Social Media Using Multimodal Data With Deep Learning." IEEE Transactions on Computational Social Systems, pp. 1–12. DOI: 10.1109/TCSS.2023.3296837.
9. Krishnamoorthy, P., Sathiyanarayanan, M., and Proenc¸a, H.P. (2024). "A novel and secured email classification and emotion detection using hy- brid deep neural network." International Journal of Cognitive Computing in Engineering, 5, pp. 44–57. DOI: 10.1016/j.ijcce.2024.01.002.
10. Shabadi, L., Chaitra, Y.L., Srikanth, P., Vijay Kumar, L., and Kashyap,
11. U. (2023). "Youtube Spam Detection Scheme Using Stacked Ensemble Machine Learning Model." 2023 International Conference on Network, Multimedia and Information Technology (NMITCON 2023). DOI: 10.1109/NMITCON58196.2023.10276002.
12. Thamizhselvi, D., Kumar, K.V., Bharath, S., and Niranjan, V.J.A. (2023). "STREAK - SOCIAL MEDIA PLATFORM WITH DYNAMIC IDEN- TIFICATION OF SPAM WORDS, PHISHING URLS, MISLEADING
13. PRODUCTS AND AGENCIES." 2023 IEEE International Conference on Research Methodologies in Knowledge Management, Artificial Intel- ligence and Telecommunication Engineering (RMKMATE 2023). DOI: 10.1109/RMKMATE59243.2023.10369814.
14. Balfagih, A., Keselj, V., and Taylor, S. (2022). "N-gram and Word2Vec Feature Engineering Approaches for Spam Recognition on Some Influ- ential Twitter Topics in Saudi Arabia." ACM International Conference Proceeding Series, pp. 101–107. DOI: 10.1145/3546157.3546173.
15. Nikumbh, D., Thakare, A., and Nandu, D. (2024). "Analyzing User Profiles for Bot Account Detection on Twitter via Machine Learning Approach." Lecture Notes in Networks and Systems, 878, pp. 107–119. DOI: 10.1007/978-981-99-9489-2 10.
16. Kerrysa, N.G., and Utami, I.Q. (2023). "Fake account detection in social media using machine learning methods: literature review." Bulletin of Electrical Engineering and Informatics, 12(6), pp. 3790–3797. DOI: 10.11591/eei.v12i6.5334.
17. Ouni, S., Fkih, F., and Omri, M.N. (2022). "BERT- and CNN-based TOBEAT approach for unwelcome tweets detection." Social Network Analysis and Mining, 12(1), art. no. 144. DOI: 10.1007/s13278-022- 00970-0.
18. Pal, S., and Lamba, A.K. (2024). "Systematic Approach for Detection and Prevention of False Information on Social Media Platform." Inter- national Journal of Intelligent Systems and Applications in Engineering, 12(9s), pp. 291–300.
19. Chrismanto, A.R., Winarko, E., and Suyanto, Y. (2024). "EiAP-BC: A Novel Emoji Aware Inter-Attention Pair Model for Contextual Spam Comment Detection Based on Posting Text." ACM Transactions on

ISSN: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

20. Asian and Low-Resource Language Information Processing, 23(12), art. no. 165. DOI: 10.1145/3696663.

21. Nayak, A., Kumari, R., Pal, D., Jana, S., Bhardwaj, A., and Da- sude, P.M. (2024). "Multilingual SMS Spam Detection using BERT and LSTM." 2024 International Conference on Innovations and Chal- lenges in Emerging Technologies (ICICET 2024). DOI: 10.1109/ICI- CET59348.2024.10616322.

22. Meenakshi, M., Shyam Babu, P., and Hemamalini, V. (2023). "Deep Learning Techniques for Spamming and Cyberbullying Detection." Proceedings of the 1st IEEE International Conference on Network- ing and Communications 2023 (ICNWC 2023). DOI: 10.1109/IC- NWC57852.2023.10127460.

23. Saeed, M.M., and Aghbari, Z.A. (2023). "Survey on Deep Learning Ap- proaches for Detection of Email Security Threat." Computers, Materials and Continua, 77(1), pp. 325–348. DOI: 10.32604/cmc.2023.036894.

24. Pal, S., and Lamba, A.K. (2023). "Tweeting Truth: Investigating False Information on Twitter Using Data Analysis and a Proposed Detec- tion Model." Proceedings - 2023 IEEE World Conference on Ap- plied Intelligence and Computing (AIC 2023), pp. 319–324. DOI: 10.1109/AIC57670.2023.10263951.