Hybrid deep learning model for YouTube spam comment detection

Muhammad Sam'an[1], Khrisna Imaddudin[2]
[1]Department of Informatics, Universitas Muhammadiyah Semarang, Semarang, Indonesia
[2]Undergraduted Students in Department of Informatics, Universitas Muhammadiyah Semarang, Semarang, Indonesia

Article Info

Article history:

Received Aug 24, 2023
Revised Jan 9, 2024
Accepted Feb 25, 2024

Keywords:

Convolutional neural network
Gated recurrent unit
Hybrid deep learning classification
Long short-term memory
YouTube spam detection

ABSTRACT

Social media platforms, including YouTube and Facebook, allow users to interact through comments and videos. However, the openness of these platforms also makes them susceptible to spammers engaging in phishing, malware distribution, and advertisement dissemination. In response, our study introduces an innovative technique for detecting features indicative of spam within comments associated with shared videos. The initial phase involves data collection from the University of California, Irvine (UCI) machine learning repository and preprocessing using tokenization and lemmatization. Subsequently, a rigorous feature selection process is executed, and experiments are conducted with various proposed classification models. The performance evaluation demonstrates outstanding accuracy in identifying spam comments on YouTube: convolutional neural network with gated recurrent unit (CNN-GRU) at 95.92%, convolutional neural network with long short-term memory (CNN-LSTM) at 95.41%, convolutional neural network with bidirectional long short-term memory (CNN-biLSTM) at 96.43%, gated recurrent unit (GRU) at 95.41%, long short-term memory (LSTM) at 94.13%, and bidirectional long short-term memory (biLSTM) at 96.94% and convolutional neural network (CNN) at 94.64%. These results highlight the substantial contribution of our approach to spam detection and the fortification of online security.

Corresponding Author:

Muhammad Sam'an
Department of Informatics, Universitas Muhammadiyah Semarang
Semarang, Indonesia
Email: muhammad92sam@unimus.ac.id

1. INTRODUCTION

YouTube is a well-known social platform that serves as a medium for users to share and upload relevant videos. Internet users from various parts of the world can watch these videos online. Through the videos foundon YouTube, users can share their creations and provide comments on those videos. Comments originating from users are not only limited to praising good videos or criticizing disliked videos. However, they can also take the form of unwanted or irrelevant electronic messages, which are then sent massively to several recipientsin a form known as spam [1].

Spam is not just a problem; it also leads to negative impacts such as wasting time, excessive memory usage, and inefficient network bandwidth utilization. The threats within spam can result in financial losses for organizations and users [2]. Some use YouTube comments for advertising, while others spread computer viruses; there are also intentionally designed spam messages to steal financial identities [3]. The most serious threat arises when spam is linked to malicious actions that direct users to phishing sites and distribute malware [4]. The spam ratio on YouTube is 100 to 1, indicating the severity of the spam threat as depicted in Table 1. It can be concluded that spam can pose a dangerous security threat to users. Spammers exploit this opportunity to spread malicious software through comments, exploiting user device vulnerabilities. This can also involve financial information theft, damaging web page content, and disturbing visitors by reducing content quality [5]. Researchers have extensively studied the detection of spam on YouTube. Classifying YouTube comments as spam and ham using machine learning [5]–[13], cascaded ensemble machine learning model [14], Markov decision process [15], artificial neural network [16], Microsoft structured query language server data mining tools [17], contextual feature based one-class classifier approach [18], hybrid ensemble machine learning models [19], multi-stage spam account [20]. Brain-inspired hyperdimensional computing [21], genetic algorithmic multi evaluation [22], n-gram assisted [23]. This comprehensive exploration demonstrates dedication to combat spam through various sophisticated techniques encompassing machine learning paradigms, algorithmic advancements, and even bio-inspired computing, collectively working to uphold the integrity of platforms like YouTube. Implementing deep learning models provides a robust solution for tackling the complexity of textual data [24]. Combinations of convolutional neural networks with long short-term memory (CNN-LSTM), convolutional neural networks with bidirectional long short-term memory (CNN-biLSTM), and convolutional neural networks with gated recurrent unit (CNN-GRU) are commonly used approaches in spam detection in YouTube comments or other domains involving text analysis. This approach is used because convolutional neural network (CNN) effectively recognizes visual patterns in text, while long short-term memory (LSTM) and bidirectional long short-term memory (biLSTM) effectively understand sequence and context in text. Combining the advantages of these two worlds, a hybrid model like this can help detect spam that utilizes visual elements and text contextto trick spam detection algorithms. This model can effectively merge an understanding of temporal context and essential feature extraction from the text, enabling more precise detection of increasingly diverse spam patterns. Training the model on preprocessed comment datasets can automatically distinguish between legitimate and potential spam comments. Implementing this model in real-time moderation efficiently handles suspiciouscomments, ensuring the safety and cleanliness of the online environment for YouTube users.

Table 1. Analysis of widespread spam on a well-known social platform [4]

Description	Data
Social media applications characterized by spam-like behavior	5%
Social media applications owned by brands exhibiting spam-like behavior	20% (that is 1% overall)
The mean count of social profiles engaged by a spam account	23
Count of newly generated spam accounts	5 out of every new account
The social platform most preferred by spammers	Facebook and YouTube
Percentage of spam messages containing URLs	15%
Total count of messages across social media	1 out of every 200

2. METHOD

Figure 1 illustrates the workflow of the proposed method for YouTube spam detection. There are 4 phases in this workflow: data collection, data pre-processing, feature selection and extraction, and classification. These phases are explained in the following subsections.

2.1. Data collection

The experimental dataset is obtained from the University of California, Irvine (UCI) machine learning repository. The dataset consists of five specifically chosen videos sourced from YouTube using the application programming interface (API) [25]. These videos belong to Park Jae-sang (PSY), KatyPerry, laughing my freaking ass off (LMFAO), Eminem, and Shakira. The aggregate number of spam and legitimate comments in the PSY video amounts to 350, while Katy Perry’s video contains 350, LMFAO’s has 438, Eminem’s has 448, and Shakira’s has 370. A dataset comprising 1,956 review data points was utilized to streamline the computational process, consisting of 1,005 spam comments and 935 legitimate (ham) comments.

2.2. Data pre-processing

Data preprocessing is a vital step in machine learning methods. This process aims to prepare the dataset by cleaning it, allowing relevant features to be extracted according to this specific detection framework. Two processes are carried out in the preprocessing steps for this research. These processes involve tokenization and stemming. Tokenization involves the separation of comments based on spaces (-) and punctuation marks [26]. The stemming process involves transforming words into their root forms. For instance, the word ”Subscribe” becomes ”Subscribe” [25].

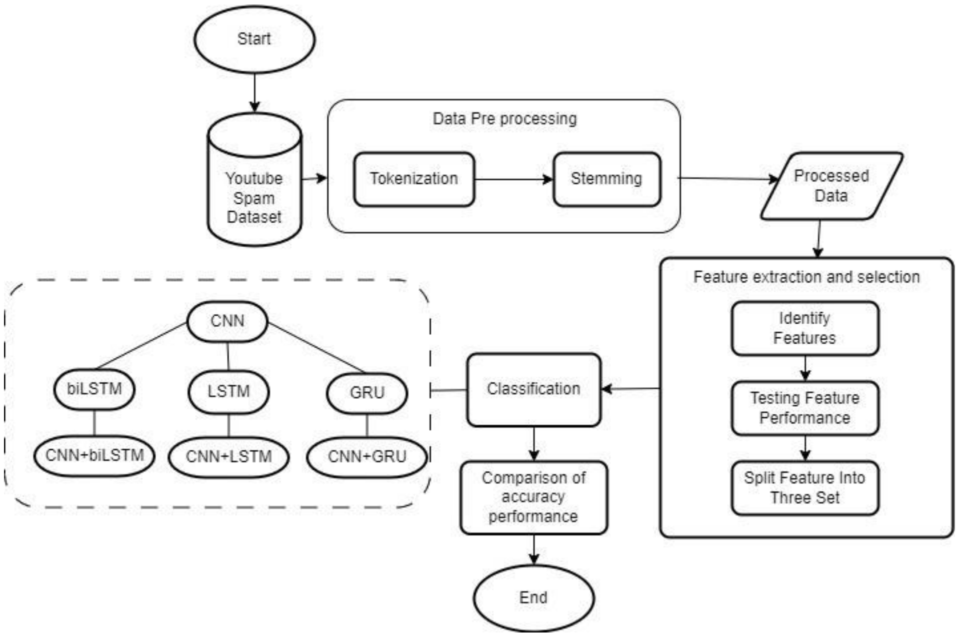

Figure 1. The workflow of proposed method

2.3. Features selection and extraction

This phase involves cleaning data, identifying relevant features based on YouTube comments, dividing the features into three sets to identify the best feature set, and testing these features using classification techniques. Links or uniform resource locators (URLs) are frequently recognized as indicative of spam messages or comments [27]. We record this attribute through a Boolean expression, where 1 signifies existence, and 0 indicates nonexistence. Comment length is calculated post-pre-processing in this study, and this attribute holds a numerical nature [28]. It is also depicted as a Boolean expression, assigning a value of 1 if spam-related keywords are within the comment and 0 when such keywords are absent [29].

2.4. Classification

In this phase, a training and testing process is involved. 80% will be allocated for training and 20% for testing. We also set the global hyperparameters: $number\ of\ epochs = 20$, $batch\ size = 10$, and $optimizer = Adam$. After completing this phase, there should be features considered as spam. Therefore, the dataset needs to be trained based on proposed models (CNN, LSTM, biLSTM, GRU, CNN-LSTM, CNN-biLSTM, and CNN-GRU). The result performance will be used accuracy in (1).

$$Accuracy = \frac{TP+TN}{TN+FP\ TP+\ FN} \tag{1}$$

where TP: true positive, TN: true negative, FN: false negative, and FP: false positive

3. RESULTS AND DISCUSSION

The learning process generates a model or classifier adapted to the input data. The model generated through this learning process requires accuracy testing to determine whether it is appropriate. The proposed model is tested by splitting the data into 10 segments. In the first iteration, the model is tested using the first segment of the data divided into 10 parts. In the second iteration, the testing data originates from the second segment of the same data. The third segment of the data is used as the testing data for the third iteration, and so on. The accuracy of the testing data from each iteration is averaged to evaluate the model’s performance.

The training and testing data are independent, enabling the assessment of the model’s capability to process unfamiliar data. Before being tested with the testing data, the proposed model or classifier produced from the training data is initially tested on the same training data. This step aims to evaluate the performance during the learning process. Figure 2 compares model accuracy during the training and testing throughout 20 epochs. The model generated from the training data in this study demonstrates an extremely high level of accuracy. Figure 2 shows that the accuracy of the training data reaches 100% after just 20 epochs, except for CNN, which displays minor variations. This phenomenon highlights the exceptional ability of the proposed model to comprehend and capture patterns present within the training data. However, the CNN model exhibits slightly different results, possibly due to the complexity of the structure and features within the data. Nonetheless, the achieved results remain impressive and underscore the substantial potential of the model in addressing the issue of spam detection on the YouTube platform.

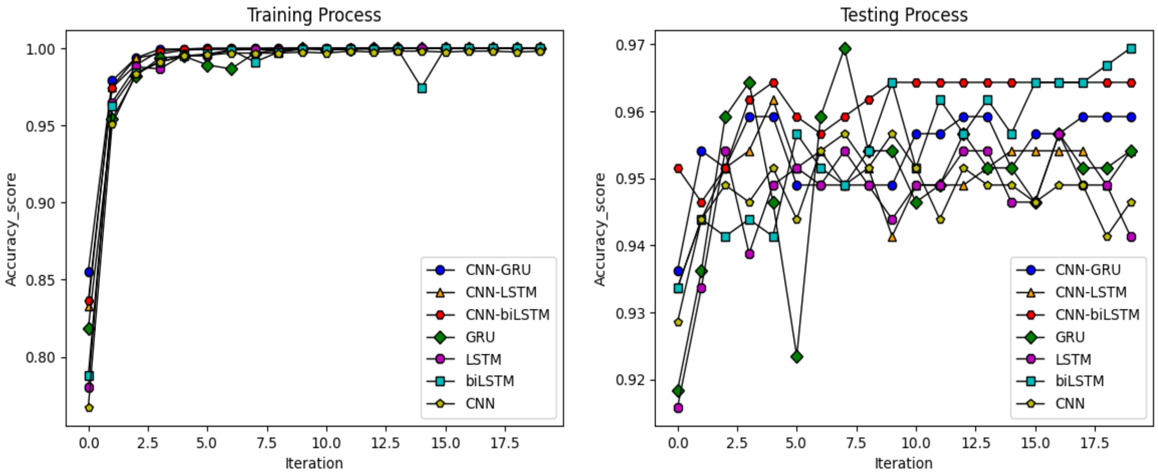

Figure 2. The comparison of the accuracy trends in the training and testing processes

All generated models achieved a final accuracy rate of 95.55% on average. This is evident when analyzing the accuracy testing graph, which displays fluctuations across all models, with the GRU model standing out as the primary focus. The CNN model exhibited the highest accuracy at 96.94%, the lowest at 91.84%, and concluded at 95.41%. The graph depicting the accuracy trends of the models against the testing data indicates a relatively stable pattern. The difference in the accuracy graph between training and testing data arises from monitoring classification accuracy using the training data during each testing phase, resulting in an increase or stability of training data accuracy in each iteration.

On the other hand, testing data accuracy is not monitored during the model testing process using the testing data. The comparison of movement in the training and testing result graphs indicates significant consistency in the performance of the models on both datasets. The relatively small gap between these two graphs suggests that the models did not experience overfitting, a condition where a model becomes excessively tailored to the training data, leading to decreased performance when tested on unseen data. Successfully maintaining a controlled difference between the training and testing result graphs is a positive indicator of the quality of the models produced in this study.

Figure 3 compares accuracy results obtained using (1). The highest accuracy rate is achieved by CNN-biLSTM, reaching 96.94%, followed by biLSTM (96.42%) and CNN-GRU (95.92%). CNN-LSTM and GRU have similar accuracy rates, at 95.41%, while CNN achieves 94.61% and LSTM at 94.13%. This visual representation reinforces these findings, providing a solid understanding of the effectiveness and quality of each model in tackling the challenge of spam detection on the YouTube platform. Consistency and high accuracy levels in identifying spam signify significant potential for creating a safer and spam-free online environment. Given the threat of increasingly sophisticated spam tactics, a deep understanding of the performance and characteristics of these models becomes ever more crucial. The superiority of CNN-biLSTM in detecting YouTube comment spam can be explained from the theoretical aspect of their functionalities. CNN effectively extracts spatial features, and biLSTM understands contextual relationships between words. By combining both, CNN-biLSTM becomes a robust solution for handling the task of spam detection on the YouTube platform.

Several studies have proposed various models for spam detection. The comparison of previous study results with the proposed model is presented in Table 2. This comparison provides insightful perspectives on the performance of diverse classification methods, demonstrating that the models introduced in this study exhibit remarkable quality and effectiveness. Mainly, models such as CNN-GRU, CNN-LSTM, and CNN-biLSTM show superior performance, surpassing various other classification methods in this analysis.

The accuracy performance results discussed earlier are based on the theoretical principles of each classification method. Neural network methods, as used in the study [6], excel in extracting intricate features from text data, enabling the recognition of patterns indicative of spam comments. Random forest, as outlined in [11], relies on ensemble methods, combining decisions from individual decision trees to handle data variations effectively. Naive Bayes and logistic regression, as described in [10], utilize statistical probabilities for data classification. In contrast, the support vector machine from the research [13] separates data using an optimal hyperplane for class separation.

In contrast, the K-nearest neighbor is limited in dealing with complex data variations. The Markov decision process in [15] reflects context understanding and the influence of the past in decision-making. In line with these principles, the proposed models in this study integrate innovative architectures, such as CNN-GRU, CNN-LSTM, and CNN-biLSTM, to combine feature processing, pattern recognition, and temporal context comprehension. These hybrid models effectively address the complexities and temporal patterns commonly found in text data, including spam comments, resulting in impressive accuracy levels, as demonstrated in the evaluation.

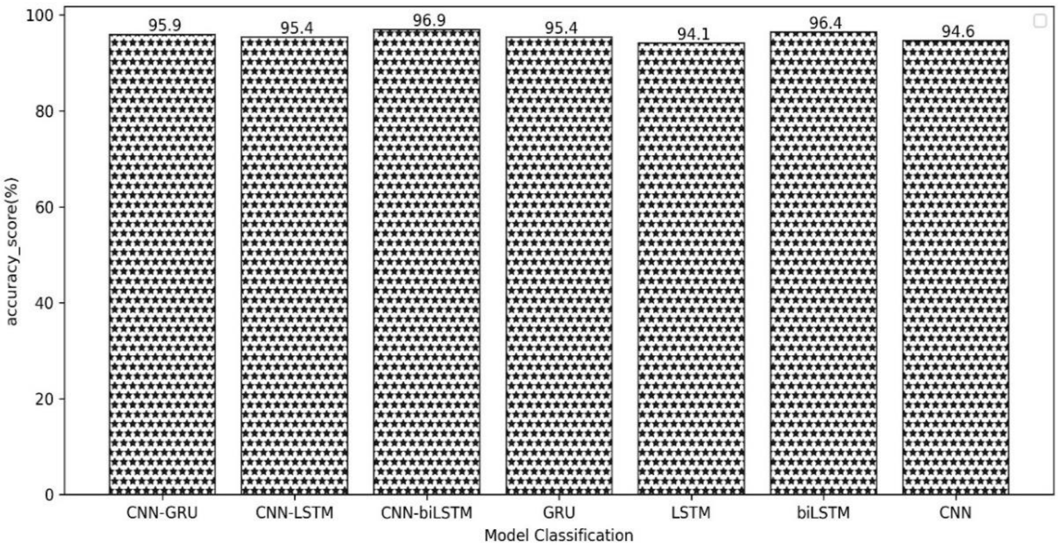

Figure 3. The comparison of testing accuracy performance

Table 2. The comparison accuracy of proposed model and previous study

Ref.	Classifier	Accuracy (%)
[6]	Neural network	91.65
[11]	Random forest	90.57
[10]	Naive Bayes	87.21
	Logistic regression	85.29
[13]	Support vector machine	74.40
	K–nearest neighbor	56.70
[15]	Markov decision process	78.82
Proposed model	CNN-GRU	95.92
	CNN-LSTM	95.41
	CNN-biLSTM	96.94
	GRU	95.41
	LSTM	94.13
	biLSTM	96.43
	CNN	94.64

4. CONCLUSION

Based on previous studies, several studies applied and evaluated various classification models for spam detection on YouTube, including neural network, random forest, naive Bayes, logistic regression, support vector machine, K-nearest neighbor, and Markov decision process. Additionally, we proposed hybrid models that combine CNN with LSTM, biLSTM, and GRU. The performance evaluation results indicate that our hybrid models, such as CNN-GRU (95.41%), CNN-LSTM (95.41%), CNN-biLSTM (96.94%), GRU (95.41%), LSTM (94.13%), biLSTM (96.43%) and CNN (94.64%), provide excellent accuracy in identifying spam comments on YouTube. Overall, this research demonstrates a significant contribution to understanding and addressing the challenges of spam detection in the online environment.

REFERENCES

[1] Y. Yusof and O. H. Sadoon, “Detecting video spammers in YouTube social media,” *ICOCI Kuala Lumpur. Universiti Utara Malaysia*, no. April 2017, pp. 228–234, 2017, [Online]. Available: http://www.uum.edu.my

[2] U. K. Sah and N. Parmar, “An approach for malicious spam detection in email with comparison of different classifiers,” *International Research Journal of Engineering and Technology (IRJET)*, vol. 4, no. 8, pp. 2238–2242, 2017.

[3] I. Dagher and R. Antoun, “Ham-spam filtering using different PCA scenarios,” in *Proceedings - 19th IEEE International Conference on Computational Science and Engineering, 14th IEEE International Conference on Embedded and Ubiquitous Computing and 15th International Symposium on Distributed Computing and Applications to Business, Engi*, Aug. 2017, pp. 542–545, doi: 10.1109/CSE-EUC-DCABES.2016.238.

[4] A. Gupta and R. Kaushal, “Improving spam detection in online social networks,” in *Proceedings - 2015 International Conference on Cognitive Computing and Information Processing, CCIP 2015*, 2015, doi: 10.1109/CCIP.2015.7100738.

[5] M. Alsaleh, A. Alarifi, F. Al-Quayed, and A. Al-Salman, “Combating comment spam with machine learning approaches,” in *Proceedings - 2015 IEEE 14th International Conference on Machine Learning and Applications, ICMLA 2015*, Dec. 2016, pp. 295–300, doi: 10.1109/ICMLA.2015.192.

[6] A. Antony, A. Rajendran, and G. Deepa, “YouTube spam comment detection,” in *Lecture Notes in Electrical Engineering*, vol. 1026 LNEE, Springer Nature Singapore, 2023, pp. 387–394, doi: 10.1007/978-981-99-1410-4_32.

[7] H. Valpadasu, P. Chakri, P. Harshitha, and P. Tarun, “Machine learning based spam comments detection on YouTube,” in *Proceedings of the 7th International Conference on Intelligent Computing and Control Systems, ICICCS 2023*, May 2023, pp. 1234–1239, doi: 10.1109/ICICCS56967.2023.10142608.

[8] Y. Tashtoush, A. Magableh, O. Darwish, L. Smadi, O. Alomari, and A. Alghazoo, “Detecting Arabic YouTube spam using data mining techniques,” *2022 10th International Symposium on Digital Forensics and Security (ISDFS)*, Istanbul, Turkey, 2022, pp. 1-5, doi: 10.1109/ISDFS55398.2022.9800840.

[9] R. K. Das, S. S. Dash, K. Das, and M. Panda, “Detection of spam in YouTube comments using different classifiers,” in *Advances in Intelligent Systems and Computing*, vol. 1082, Springer Singapore, 2020, pp. 201–214, doi: 10.1007/978-981-15-1081-6_17.

[10] N. M. Samsudin, C. F. B. Mohd Foozy, N. Alias, P. Shamala, N. F. Othman, and W. I. S. Wan Din, “YouTube spam detection framework using naïve bayes and logistic regression,” *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 14, no. 3, pp. 1508–1517, Jun. 2019, doi: 10.11591/ijeecs.v14.i3.pp1508-1517.

[11] N. Alias, C. F. M. Foozy, and S. N. Ramli, “Video spam comment features selection using machine learning techniques,” *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 15, no. 2, pp. 1046–1053, Aug. 2019, doi: 10.11591/ijeecs.v15.i2.pp1046-1053.

[12] M. B. Puneeth and V. Ramakrishnan, “The mechanism of spam comment detection using count vectorizer and naive Bayes machine learning algorithms in Python,” *ECS Transactions*, vol. 107, no. 1, pp. 13417–13428, Apr. 2022, doi: 10.1149/10701.13417ecst.

[13] A. Aziz, C. F. M. Foozy, P. Shamala, and Z. Suradi, “YouTube spam comment detection using support vector machine and K–nearest neighbor,” *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 12, no. 2, pp. 612–619, Nov. 2018, doi: 10.11591/ijeecs.v12.i2.pp612-619.

[14] H. Oh, “A YouTube spam comments detection scheme using cascaded ensemble machine learning model,” *IEEE Access*, vol. 9, pp. 144121–144128, 2021, doi: 10.1109/ACCESS.2021.3121508.

[15] S. Kanodia, R. Sasheendran, and V. Pathari, “A novel approach for YouTube video spam detection using Markov decision process,” in *2018 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2018*, Sep. 2018, pp. 60–66, doi: 10.1109/ICACCI.2018.8554405.

[16] T. Abd, H. Altabrawee, and S. Q. Ajmi, “YouTube spam comments detection using artificial neural network,” *Journal of Engineering and Applied Sciences*, vol. 13, no. 22, pp. 9638–9642, Jan. 2018, doi: 10.3923/jeasci.2018.9638.9642.

[17] R. Chowdury, M. N. Monsur Adnan, G. A. N. Mahmud, and R. M. Rahman, “A data mining based spam detection system for YouTube,” in *8th International Conference on Digital Information Management, ICDIM 2013*, Sep. 2013, pp. 373–378, doi: 10.1109/ICDIM.2013.6694038.

[18] V. Chaudhary and A. Sureka, “Contextual feature based one-class classifier approach for detecting video response spam on YouTube,” in *2013 11th Annual Conference on Privacy, Security and Trust, PST 2013*, Jul. 2013, pp. 195–204, doi: 10.1109/PST.2013.6596054.

[19] A. Sinhal and M. Maheshwari, “YouTube: Spam comments filtration using hybrid ensemble machine learning models,” *International Journal of Emerging Technology and Advanced Engineering*, vol. 12, no. 10, pp. 169–182, Oct. 2022, doi: 10.46338/ijetae1022_18.

[20] F. Concone, G. Lo Re, M. Morana, and S. K. Das, “SpADe: Multi-stage spam account detection for online social networks,” *IEEE Transactions on Dependable and Secure Computing*, vol. 20, no. 4, pp. 3128–3143, Jul. 2023, doi: 10.1109/TDSC.2022.3198830.

[21] R. Thapa, B. Lamichhane, D. Ma, and X. Jiao, “SpamHD: Memory-efficient text spam detection using brain-inspired hyperdimensional computing,” in *Proceedings of IEEE Computer Society Annual Symposium on VLSI, ISVLSI*, Jul. 2021, vol. 2021-July, pp. 84–89, doi: 10.1109/ISVLSI51109.2021.00026.

[22] E. Elakkiya and S. Selvakumar, “GAMEFEST: Genetic algorithmic multi evaluation measure based feature selection technique for social network spam detection,” *Multimedia Tools and Applications*, vol. 79, no. 11–12, pp. 7193–7225, Dec. 2020, doi: 10.1007/s11042-019-08334-1.

[23] S. Aiyar and N. P. Shetty, “N-Gram assisted YouTube spam comment detection,” *Procedia Computer Science*, vol. 132, pp. 174–182, 2018, doi: 10.1016/j.procs.2018.05.181.

[24] G. Jain, M. Sharma, and B. Agarwal, “Optimizing semantic LSTM for spam detection,” *International Journal of Information Technology (Singapore)*, vol. 11, no. 2, pp. 239–250, Apr. 2019, doi: 10.1007/s41870-018-0157-5.

[25] T. C. Alberto, J. V Lochter, and T. A. Almeida, “TubeSpam: Comment spam filtering on YouTube,” in *Proceedings - 2015 IEEE 14th International Conference on Machine Learning and Applications, ICMLA 2015*, Dec. 2016, pp. 138–143, doi: 10.1109/ICMLA.2015.37.

[26] T. Verma, R. Renu, and D. Gaur, “Tokenization and filtering process in RapidMiner,” *International Journal of Applied Information Systems*, vol. 7, no. 2, pp. 16–18, Apr. 2014, doi: 10.5120/ijais14-451139.

[27] A. K. Uysal, S. Gunal, S. Ergin, and E. S. Gunal, “The impact of feature extraction and selection on SMS spam filtering,” *Elektronika ir Elektrotechnika*, vol. 19, no. 5, pp. 67–72, May 2013, doi: 10.5755/j01.eee.19.5.1829.

[28] C. Radulescu, M. Dinsoreanu, and R. Potolea, "Identification of spam comments using natural language processing techniques," in *Proceedings - 2014 IEEE 10th International Conference on Intelligent Computer Communication and Processing, ICCP 2014*, Sep. 2014, pp. 29–35, doi: 10.1109/ICCP.2014.6936976.

[29] R. Shams and R. E. Mercer, "Classifying spam emails using text and readability features," in *Proceedings - IEEE International Conference on Data Mining, ICDM*, Dec. 2013, pp. 657–666, doi: 10.1109/ICDM.2013.131.

BIOGRAPHIES OF AUTHORS

Muhammad Sam'an received bachelor's degree from Universitas Negeri Semarang and master degree from Universitas Diponegoro in mathematics 2010 and 2016 respectively. His research interests are in optimization, fuzzy mathematics and computational mathematics. He can be contacted at email: muhammad92sam@unimus.ac.id.

Khrisna Imaddudin is bachelor students from Universitas Muhammadiyah Semarang. His research interests are in natural language processing (NLP). He can be contacted at email: krishnaimad@unimus.ac.id.