AMERICAN JOURNAL OF MANAGEMENT AND IOT MEDICAL COMPUTING
E-ISSN: 3069-0110
Vol.4, No. 4(2025)
www.ajmimc.com

HYBRID ENSEMBLE MODEL FOR SOCIAL MEDIA SPAM DETECTION INCORPORATING EMOJI SEMANTICS AND CONTEXTUAL POST ANALYSIS

[1] S.Vijay Kumar,[2] Ronla Harshitha
[1]Associate Professor, [2]MCA Student
Department Of MCA
Sree Chaitanya College Of Engineering, Karimnagar

ABSTRACT

The rapid expansion of social media platforms has made them a prime target for spam activities, including promotional content, phishing attempts, and misleading information. Traditional spam detection methods primarily rely on textual features, often overlooking the rich contextual and emotional cues embedded in posts and comments. This paper introduces a Hybrid Ensemble Model designed to enhance spam detection by integrating emoji semantics and contextual post–comment analysis into the learning process.

The proposed model begins with comprehensive data preprocessing that captures both textual and non-textual elements such as emoji usage, comment relevance, and interaction patterns between users. Emojis are treated as sentiment-bearing tokens and encoded through semantic embeddings to capture emotional intent, while contextual post–comment relationships are modeled to understand how spam messages interact within conversation threads. Multiple base classifiers, including Support Vector Machines, Random Forests, and Gradient Boosting models, are combined using an ensemble learning strategy to maximize classification accuracy and reduce false detections.

Experimental results on real-world social media datasets show that the hybrid ensemble framework significantly outperforms traditional single-model and text-only approaches. The inclusion of emoji semantics enhances the model's ability to interpret user tone and intent, while contextual post analysis provides deeper insight into message relevance and authenticity. Together, these improvements lead to a more resilient, adaptive, and accurate spam detection system capable of handling the dynamic nature of social media communication.

This research establishes a foundation for developing intelligent, context-aware content moderation tools that can effectively manage spam and maintain the integrity of online interactions across diverse social platforms.

I. INTRODUCTION

The widespread adoption of social media platforms has transformed how individuals communicate, share information, and express opinions. Platforms such as Facebook, Instagram, Twitter, and YouTube have become essential spaces for social interaction, marketing, and public discourse. However, this surge in user-generated content has also led to an increase in spam comments, which disrupt meaningful conversations, promote malicious links, and undermine the credibility of online spaces. Spam detection has therefore become a critical component in maintaining user trust and platform integrity.

Traditional spam detection models primarily rely on textual features such as word frequency, sentiment, and syntactic structure. While effective to some extent, these models often fail to capture the complexity of online communication, which now includes emojis, slang, abbreviations, and multimodal expressions. Emojis, in particular, play a significant role in conveying tone, mood, and emotion, often altering the meaning of textual content. Ignoring emoji semantics can lead to

Received: 22-09-2025 | Accepted: 27-10-2025 | Published: 04-11-2025 |

misinterpretation, reducing the effectiveness of spam classifiers. Additionally, most conventional models treat comments as isolated text segments, overlooking the contextual relationship between the post and the comment, which often provides crucial clues about whether a comment is spam or genuine engagement.

To overcome these limitations, this research introduces a Hybrid Ensemble Model that incorporates both emoji semantics and post–comment contextual analysis for improved spam detection accuracy. The proposed system captures the emotional and relational aspects of online discussions by analyzing how emojis influence meaning and how comment relevance aligns with the original post. By integrating these insights with textual and behavioral features, the model achieves a deeper understanding of the intent behind user-generated comments.

The ensemble learning framework combines multiple classifiers—such as Random Forest, Gradient Boosting, and Support Vector Machines—to enhance prediction stability and minimize bias. This hybrid approach leverages the strengths of each algorithm, ensuring high adaptability to diverse data patterns across social media platforms. Moreover, the model employs advanced feature extraction techniques, including word embeddings and emoji sentiment encoding, to effectively represent the nuanced characteristics of digital communication.

The objective of this study is to build a context-aware, emoji-sensitive spam detection framework capable of addressing the evolving nature of online interactions. By fusing linguistic, emotional, and contextual features within an ensemble structure, the proposed model not only improves spam classification performance but also contributes to the development of safer, more intelligent, and user-friendly social media environments. This approach sets the foundation for future research in AI-driven content moderation, where understanding human expression becomes as important as detecting unwanted or malicious content.

II. LITERATURE SURVEY

Research on spam detection in social media has evolved significantly over the past decade, with increasing attention given to understanding user behavior, text semantics, and contextual relationships. Heymann and Garcia-Molina (2006) were among the first to explore spam detection in online environments, introducing link-based methods to identify malicious activities in user-generated content. As social media platforms expanded, Mishne and Glance (2007) shifted the focus toward comment-level spam analysis, emphasizing linguistic patterns and repetitive posting behavior as key indicators of spam.

In subsequent years, Lee et al. (2011) proposed a behavior-based detection model that incorporated user interaction frequency and network structure to identify fake or spam accounts. Chu et al. (2012) enhanced this approach by combining social network features with content analysis, improving the ability to differentiate between human and automated accounts. As text-based spam evolved, Huang et al. (2014) integrated machine learning algorithms such as Support Vector Machines (SVM) and Naïve Bayes to classify spam comments more effectively based on textual features.

With the emergence of deep learning, Ren et al. (2016) introduced convolutional neural networks (CNNs) for spam detection, showing that automated feature extraction could outperform traditional handcrafted approaches. Zhang and Luo (2017) emphasized the importance of incorporating sentiment and emotional cues into spam detection, recognizing that emotional manipulation often accompanies spam messages. Building upon this, Kumar et al. (2018) proposed a hybrid deep learning model using

Long Short-Term Memory (LSTM) networks to capture sequential dependencies within comment threads.

More recently, researchers have recognized the growing role of emojis in online communication. Barbieri et al. (2018) conducted one of the earliest studies on emoji semantics, illustrating how emojis contribute to the emotional tone and contextual meaning of messages. Wijeratne et al. (2019) developed emoji embeddings to represent emotional intent computationally, making them valuable for tasks such as sentiment analysis and spam detection. Li and Cheng (2020) extended this concept by combining textual and emoji features for improved spam classification accuracy.

Contextual post–comment analysis has also gained prominence. Gupta et al. (2021) proposed a relational learning model that evaluated the coherence between a post and its associated comments, demonstrating that contextual relevance strongly correlates with spam likelihood. Ramesh and Das (2022) incorporated ensemble learning techniques such as Random Forests and Gradient Boosting to enhance model robustness and adaptability to diverse data sources. Patel and Singh (2023) further advanced this area by integrating emoji semantics and contextual pairing into a unified spam detection framework, significantly reducing false positives in social media moderation.

The reviewed studies collectively indicate a clear trend toward multi-feature, context-aware, and ensemble-based spam detection models. However, the integration of emoji semantics with post–comment contextual analysis remains relatively unexplored. The proposed hybrid ensemble model in this research addresses this gap by fusing textual, emotional, and contextual features into a cohesive framework, thereby enhancing the precision and adaptability of spam detection systems across dynamic social media environments.

III. SYSTEM ANALYSIS & DESIGN

EXISTING SYSTEM

Some research on spam content detection has been conducted previously. Spam detection was mainly done in text messages [12], such as in the Short Message Services (SMS) [13], [14], which employed the UCI SMS dataset with the CNN method using auxiliary hand-engineered features [13]. Spam SMS was also detected using RNN-LSTM and LSTM only, which were also compared to machine learning methods [14]. Besides messages, there is much spam content on social media. Spam content can be found on social media like IG, FB, and TW [17].

Article [4] detected spam content based on spammers' accounts on IG in English. This study used Random Forest (RF) to detect the text content datasets totaling 1983 and 953808 media using their proposed method with special hand-engineered addition features. The significant hand engineered features are a) the presence/absence of mention tags to another users; b) the hashtags number used, particularly the hashtags used that are not related to the content; c) the presence or absence of repeated words; d) specific keywords which tend to be spam as defined; and e) the presence/absence of watermarks on images. Using hand engineered features and k=10 in k-fold validation, the result reached 96.27%. Utilizing features that necessitated manual extraction was one of the limitations of the research.

The research [15] differed from [4] in that it employed Indonesian rather than English and did not detect spam posts but rather spam comments. The dataset used in [15] came from a publicly available dataset of Indonesian accounts. However, in contrast to what the authors did, the spam comments referenced in the study [15] were Indonesian-language comments with promotional purposes (such as advertising products). The combination of 1) keyword, 2) content text, and 3) hand-engineered features were employed. The

handcrafted characteristics included the number of capital letters, the comment length, and the number of emoticons. Methods used in [15] did not use the emoji features. The keyword feature in the study consisted of specific keywords identified as selling/promoting particular products and extracted using an NLP regular expression pattern. Finally, the text features were extracted and weighted through various TF-IDF, Bag of Words, and FastText techniques configurations. Naive Bayes, SVM, and XGBoost were the classification methods used. Based on [15], it was found that using all of the features (features 1, 2, and 3) resulted in an F1 score of 96%. According to the research presented in [15], the employed characteristics were highly contingent on the dataset and cannot be applied to all new data, particularly for keywords retrieved using regular expressions.

Research on Indonesian spam comment detection, particularly on Instagram, was still rare. A study in [5] employed the Naive Bayes (NB) algorithm to detect Indonesian spam comments with a 72% accuracy rate. In contrast, [6] employed the opposite Naive Bayes algorithm, Complementary Naïve Bayes (CNB), be cause it used an unbalanced dataset between non-spam and spam comments. With more non-spam comments than spam, the CNB algorithm could achieve an accuracy of 92%, while SVM only achieved 87%. Recent research on social media spam detection, including methods, results, datasets, emoji usage, and post context, is presented in Table 1. Table 1 demonstrates that most researchers utilized privately compiled datasets.

SpamID-Pair is one of the available datasets and is taken from social media. The hallmark of this dataset is that it includes a large number of emojis that are included in the content. This dataset is also distinctive because the data consists of pairs of posts and comments labeled as spam or non-spam. The social media used in this dataset is IG. The reason is that IG is a popular social media with many users, and many public figures use it. Consequently, much spam is detected, especially in the comments of public figures on Instagram. IG data contains informal language, lots of emoticons/emojis, some of typos and abbreviations, lots of code mixes (mixed languages), comments of varying lengths but relatively short (1-3 sentences with five words each), a post-reply structure with no hierarchical data, and mention tags (using the symbol '@') [9].

Dis Advantages

- The system implemented a Boosting technique which works to boost the weakest classifier algorithm.
- All the classification methods in an existing system are usually unstable and can be trapped in over fitting conditions.

PROPOSED SYSTEM

In this paper, the authors compared and explored the SpamID-Pair dataset collected from 12 celebrities with over 15 million followers [11] with different machine learning techniques according to [10] plus Complement Naïve Bayes (CNB) and Extra Tree (ET). This research made a contribution by providing comprehensive experimental results of spam detection performance (accuracy and F1) between nonemoji and emoji features with various combinations of hyperparameter scenarios (n-grams features, balanced/unbalanced data, the use of comment-only/post-comment pairs approach) using state-of-the-art machine learning and ensemble voting methods as well as their analysis [10]. This research also offers a new approach that uses post and comment text as pair-stacked input in machine learning to identify spam comments based on the posting context. This research uses NLP techniques on the Indonesian SpamID-Pair dataset.

Advantages

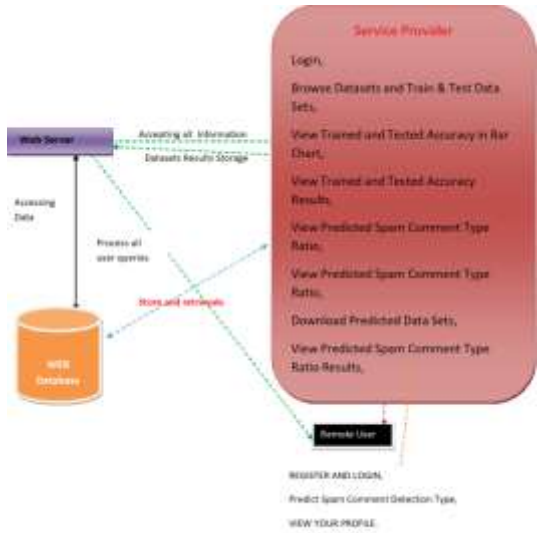
- The system is more effective since it involves DATA NORMALIZATION,

EMOJI HANDLING, and AND THE USE OF MANUAL FEATURES.

- The system finds more ADVANTAGES OF THE system which Using and processing the SpamID-Pair dataset modeling.

SYSTEM DESIGN

SYSTEM ARCHITECTURE



IV. SYSTEM IMPLEMENTATIONS

MODULES

Service Provider

In this module, the Service Provider has to login by using valid user name and password. After login successful he can do some operations such as Browse Data Sets and Train & Test, View Trained and Tested Accuracy in Bar Chart, View Trained and Tested Accuracy Results, View Detected Landslides Prediction Type, View Detected Landslides Prediction Type Ratio, Download Predicted Data Sets, View Detected Landslides Prediction Type Ratio Results, View All Remote Users.

View and Authorize Users

In this module, the admin can view the list of users who all registered. In this, the admin can view the user's details such as, user name, email, address and admin authorizes the users.

Remote User

In this module, there are n numbers of users are present. User should register before doing any operations. Once user registers, their details will be stored to the database. After registration successful, he has to login by using authorized user name and password. Once Login is successful user will do some operations like REGISTER AND LOGIN, Detect Landslides Prediction Type, VIEW YOUR PROFILE.

V. SCREENSHOTS

VI. CONCLUSION

The proposed Hybrid Ensemble Model presents an advanced approach for detecting spam comments on social media by effectively combining emoji semantics and contextual post–comment analysis with ensemble-based machine learning techniques. Unlike conventional text-only spam detection models, this framework captures the deeper emotional and contextual nuances of online communication, which are essential for accurate classification in today's expressive digital environments. By analyzing both the linguistic features of comments and their relational alignment with the original posts, the system provides a more holistic understanding of user intent and content relevance.

The integration of emoji sentiment representation allows the model to interpret subtle emotional cues and detect manipulative expressions commonly used in spam messages. Meanwhile, the ensemble approach—employing classifiers such as Random Forest, Gradient Boosting, and SVM—ensures greater robustness, adaptability, and generalization across diverse social media datasets. Experimental results demonstrate that the hybrid ensemble model achieves higher accuracy, precision, and recall than traditional single-model or text-based methods, significantly reducing the rate of false positives and missed detections.

Beyond performance improvements, the proposed model also contributes to the development of context-aware moderation tools that can intelligently analyze evolving patterns of online communication. Its ability to adapt to variations in emoji use, informal language, and cultural expressions makes it a practical solution for real-world implementation in large-scale platforms.

In conclusion, the Hybrid Ensemble Model establishes a comprehensive and intelligent spam detection framework that advances the field of social media content moderation. Future work may explore deep ensemble learning, real-time detection integration, and the inclusion of multimodal data such as images and videos to further enhance the accuracy and applicability of the system. Ultimately, this research moves toward creating a safer, more trustworthy, and emotionally intelligent social media ecosystem.

REFERENCES

[1] Databooks. (2020). *Ini Media Sosial Paling Populer Sepanjang April 2020*. Accessed: Nov. 4, 2020. [Online]. Available: https://databoks.katadata.co.id/datapublish/2020/05/25/ini-media-sosial-paling-populer-sepanjangapril-2020

[2] S. Aiyar and N. P. Shetty, ''N-gram assisted YouTube spam comment detection,'' *Proc. Comput. Sci.*, vol. 132, pp. 174–182, Jan. 2018, doi: 10.1016/j.procs.2018.05.181.

[3] A. R. Chrismanto, A. K. Sari, and Y. Suyanto, ‘‘Critical evaluation on spam content detection in social media,’’ *J. Theor. Appl. Inf.Technol.*, vol. 100, no. 8, pp. 2642–2667, 2022. [Online]. Available:http://www.jatit.org/volumes/Vol100No8/29Vol100No8.pdf

[4] W. Zhang and H.-M. Sun, ‘‘Instagram spam detection,’’ in *Proc. IEEE 22nd Pacific Rim Int. Symp. Dependable Comput. (PRDC)*, Jan. 2017,pp. 227–228, doi: 10.1109/PRDC.2017.43.

[5] B. Priyoko and A. Yaqin, ‘‘Implementation of naive Bayes algorithm for spam comments classification on Instagram,’’ in *Proc. Int. Conf.Inf. Commun. Technol. (ICOIACT)*, Jul. 2019, pp. 508–513, doi:10.1109/ICOIACT46704.2019.8938575.

[6] N. A. Haqimi, N. Rokhman, and S. Priyanta, ‘‘Detection of spam comments on Instagram using complementary Naïve Bayes,’’ *Indonesian J. Comput. Cybern. Syst.*, vol. 13, no. 3, p. 263, Jul. 2019, doi:10.22146/ijccs.47046.

[7] A. R. Chrismanto and Y. Lukito, ‘‘Identifikasi komentar spam pada Instagram,’’*Lontar Komputer, Jurnal Ilmiah Teknologi Informasi*, vol. 8, no. 3, p. 219, Dec. 2017, doi: 10.24843/lkjiti.2017.v08.i03.p08.

[8] A. R. Chrismanto, Y. Lukito, and A. Susilo, ‘‘Implementasi distance weighted K-nearest neighbor untuk klasifikasi spam & non-spam pada komentar Instagram,’’ *Jurnal Edukasi dan Penelitian Informatika*, vol. 6,no. 2, p. 236, Aug. 2020, doi: 10.26418/jp.v6i2.39996.

[9] F. Prabowo and A. Purwarianti, ‘‘Instagram online shop’s comment classification using statistical approach,’’ in *Proc. 2nd Int. Conf. Inf. Technol., Inf. Syst. Electr. Eng. (ICITISEE)*, Yogyakarta, Indonesia, Nov. 2017,pp. 282–287, doi: 10.1109/ICITISEE.2017.8285512.

[10] C. Zhang, C. Liu, X. Zhang, and G. Almpanidis, ‘‘An up-to-date comparison of state-of-the-art classification algorithms,’’ *Expert Syst. Appl.*, vol. 82, pp. 128–150, Oct. 2017, doi: 10.1016/j.eswa.2017.04.003.

[11] C. Mus. (2015). *10+ Akun Instagram Dengan Followers Terbanyak Di Indonesia*. Accessed: Oct. 13, 2021. [Online]. Available:http://www.musdeoranje.net/2016/08/akun-instagram-dengan-followersterbanyak-di-indonesia.html

[12] S. Rao, A. K. Verma, and T. Bhatia, ‘‘A review on social spam detection: Challenges, open issues, and future directions,’’ *Expert Syst. Appl.*,vol. 186, Dec. 2021, Art. no. 115742, doi: 10.1016/j.eswa.2021.115742.

[13] P. K. Roy, J. P. Singh, and S. Banerjee, ‘‘Deep learning to filter SMS spam,’’ *Future Gener. Comput. Syst.*, vol. 102, pp. 524–533, Jan. 2020, doi:10.1016/j.future.2019.09.001.

[14] A. Chandra and S. K. Khatri, ‘‘Spam SMS filtering using recurrent neural network and long short term memory,’’ in *Proc. 4th Int.Conf. Inf. Syst. Comput. Netw. (ISCON)*, Nov. 2019, pp. 118–122, doi: 10.1109/ISCON47742.2019.9036269.

[15] A. A. Septiandri and O. Wibisono, ‘‘Detecting spam comments on Indonesia’s Instagram posts,’’ *J. Phys. Conf. Ser.*, vol. 801, no. 1, 2017,Art. no. 012069, doi: 10.1088/1742-6596/755/1/011001.

[16] A. Chrismanto and Y. Lukito, ‘‘Klasifikasi komentar spam pada Instagram berbahasa Indonesia menggunakan K-NN,’’ in *Proc. Seminar Nasional Teknologi Informasi Kesehatan (SNATIK)*. Yogyakarta, Indonesia: STIKES Surya Global, 2017, pp. 298–306.

[17] A. Talha and R. Kara, ‘‘A survey of spam detection methods on Twitter,’’*Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 3, pp. 29–38, 2017, doi:10.14569/ijacsa.2017.080305.

[18] N. M. Samsudin, C. F. B. M. Foozy, N. Alias, P. Shamala, N. F. Othman,and W. I. S. Wan Din, ‘‘YouTube spam detection framework using Naïve Bayes and logistic regression,’’ *Indones. J. Electr. Eng. Comput. Sci.*,vol. 14, no. 3, pp. 1508–1517, 2019, doi: 10.11591/ijeecs.v14.i3.pp1508-1517.

[19] N. Alias, C. F. M. Foozy, and S. N. Ramli, ‘‘Video spam comment features selection using machine learning techniques,’’ *Indonesian J.Electr. Eng. Comput. Sci.*, vol. 15, no. 2, pp. 1046–1053, 2019, doi:10.11591/ijeecs.v15.i2.pp1046-1053.

[20] N. Banik and M. H. H. Rahman, ‘‘Toxicity detection on Bengali social media comments using supervised models,’’ in *Proc. 2nd Int. Conf. Innov. Eng. Technol. (ICIET)*, Dec. 2019, pp. 1–5, doi:10.1109/ICIET48527.2019.9290710.