IJITCE

ISSN 2347- 3657

International Journal of Information Technology & Computer Engineering

Email : ijitce.editor@gmail.com or editor@ijitce.com

International Journal of Information Technology & Computer Engineering

ISSN 2347–3657
Volume 15, Issue 3, 2024

https://doi.org/10.62646/ijitce.2024.v15.i3.pp837-845

ENHANCING SPAM COMMENT DETECTION ON SOCIAL MEDIA WITH EMOJI FEATURE AND POST-COMMENT PAIRS APPROACH USING ENSEMBLE METHODS OF MACHINE LEARNING

[1] Mrs. Sri Lavanya Sajja,[2]Lokineni Yuktha,[3] N. Esther,[4] P. Charitha
[1]Assistant Professor,[234]Students
Department Of CSE
Malla Reddy Engineering College for Women

ABSTRACT

Whenever a well-known public person shares anything on social media, a lot of people are inspired to leave comments. Regretfully, not every remark is pertinent to the article. A portion of the comments are spam, which might impede the information's general flow. Two approaches were used in this study to solve problems with text spam identification on social media. The first tactic was using emoticons, which had been widely disregarded in previous research. Emojis are very widely used by social media users to express their intents. Unlike many spam detection algorithms that just looked at comment-only data, the second technique made advantage of stacked post-comment pairings. It was necessary for the post-comment pairings to determine if a remark related to the post context (i.e., not spam) or was spam. The SpamID-Pair dataset, which was obtained from social media, was used in this study to identify spam comments in Indonesian. Following a thorough analysis, it was determined that the stacked post-comment pairings, ensemble voting, and the emoji-text feature might improve detection performance (F1 and accuracy). Performance in detecting was further enhanced by adding manual features. According to the experiment, the soft voting ensemble approach for the best average performance and the SVM (RBF kernel) are the best stand-alone methods for spam comment identification.

I. INTRODUCTION

Social media gives individuals the ability to cooperate, do business, market goods, discuss ideas and goals, and become involved in politics. Popular social media platforms include Twitter (TW) for semi-formal and non-formal text and images, Instagram (IG) for semi-formal and non-formal text, images, and videos, Facebook (FB) for more formal or semi-formal text and image media, YouTube (YT) for semi-formal videos, Tik-Tok (TT) for non-formal videos, and Tik-Tok (TT) for non-formal videos [1]. Celebrities utilize these well-functioning, widely-user-base social media platforms to boost their public image.

Celebrities are public people with substantial social media followings. Celebrities use social media to communicate with their fans, promote their activities, and get more notoriety, among other things. Celebrities tend to have larger followings the more well-known they are. Celebrities may communicate with their fans more often if they have a larger following [2]. As is typical of Web 2.0, people may now leave imaginative comments on the feeds of celebrities.

International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 15, Issue 3, 2024
https://doi.org/10.62646/ijitce.2024.v15.i3.pp837-845

Because there are a lot of spam accounts and spam posts on TW, YT, and IG, these social media platforms are widely employed in spam detection research. Spam material, specifically from Indonesia, is often seen in comments made against Indonesian musicians, particularly on Instagram [2]. Figure 1 shows an example of a spam remark and post from the @abutting account on social media in Indonesia. The information flow in the comments on a particular post or status might be disrupted by spam remarks, which are quite unpleasant. Spam filters are already present on certain social networking sites, but they only support English.

The small number of publicly accessible datasets for recognizing spam text on social media is another issue. The majority of social media statistics are available in English, and it might be difficult to get datasets in other languages, such as Indonesian. Similar investigations were carried out by other researchers utilizing private databases that they had amassed.

Medley Data Repository offers a dataset called SpamID-Pair1 for the identification of spam material in Indonesian. Spam ID-Pair offers articles by Indonesian artists together with tagged/untagged comments in pairs. Emojis are often used on social media, and this dataset has a large number of them. Emojis are widely used by users on social media to express their feelings and intentions. However, the majority of emoji attributes are ignored or not employed in many Natural Language Processing (NLP) research studies [3]. Prior research has been done on the identification of spam material [4, 5, 6, 7, 8, 9, 9]. Unfortunately, there are a number of reasons why it is challenging to identify spam content, especially comments: 1) the unusual and highly unstructured nature of the comment text; 2) the quantity of emoticons and symbols used by users; 3) the frequency of typos, intentional abbreviations, non-standard words, and mixed language usage; 4) some content is purposefully hidden to evade detection as spam, like when the system interprets the \/ sign as the letter V; 5) the comments are spam but contain very subtle ads; and 6) the system fails to identify the semantic meaning or semantic relationship between posts and comments. These problems are complex, need research, and call for several interdependent solution components. 1SPAMID-PAIR available at https://data.medley.com/datasets/fj5pbdf95t on the Medley Data Repository.

Comment spam may be detected using a few machine learning NLP approaches. Support Vector Machine (SVM), Random Forest (RF), Logistic Regression (LR), Extreme Gradient Boosting (XG Boost), K-Nearest Neighbor (KNN), Ada Boost (AB), Naïve Byes (NB), Multi-Layer Perception (MLP), and Decision Tree (DT) are the top 14 Machine Learning (ML) classification techniques that have been examined and compared based on [10]. Shallow learning methods, sometimes referred to as machine learning techniques, are gradually evolving towards deep learning, which necessitates distinct learning approaches.

The authors of this study investigated and contrasted several machine learning algorithms, as stated in [10], using the Spam ID-Pair dataset, which was gathered from 12 celebrities with over 15 million

International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 15, Issue 3, 2024
https://doi.org/10.62646/ijitce.2024.v15.i3.pp837-845

followers [11], in addition to Complement Naïve Byes (CNB) and Extra Tree (ET). With various combinations of hyper parameter scenarios (n-grams features, balanced/unbalanced data, the use of comment-only/post-comment pairs approach), this research made a contribution by providing comprehensive experimental results of spam detection performance (accuracy and F1) between non-emoji and emoji features as well as their analysis [10]. Additionally, this study presents a novel method for identifying spam comments based on the posting context by using the text of the posts and comments as pair-stacked input in machine learning. NLP approaches are used in this study on the Indonesian Spam ID-Pair dataset.

II. LITERATURE SURVEY

Prior studies have been done on the identification of spam material. Text messages were the primary medium for spam identification [12], as shown by the Short Message Services (SMS) [13], [14], which used the CNN approach with the UCI SMS dataset and other hand-engineered characteristics [13]. RNN-LSTM and LSTM alone were also used to identify spam SMS, and their results were compared to machine learning techniques [14]. Social media is full with spam stuff in addition to communications. You may find spam material on Facebook, Instagram, and Twitter [17].

Based on the English-language Instagram profiles of spammers, article [4] identified spam material. This research used Random Forest (RF) with specially hand-engineered addition features to identify text content datasets comprising 1983 and 953808 media. The most important hand-engineered features are: a) the existence or non-existence of mention tags for other users; b) the quantity of hashtags used, especially those unrelated to the content; c) the presence or non-existence of repeated words; d) certain keywords that are frequently considered spam; and e) the presence or non-existence of watermarks on images. With k=10 in k-fold validation and hand-engineered features, the outcome was 96.27%. Using attributes that required manual extraction was one of the study's shortcomings.

In contrast to [4], study [15] used Indonesian instead of English and identified spam comments rather than posts. The Indonesian accounts dataset included in [15] was sourced from a publically accessible dataset.

Nevertheless, the research [15] cited spam comments that were written in Indonesian and had promotional intent (i.e., promoting products)—a departure from the authors' actions. Three strategies were used: 1) keyword, 2) content text, and 3) hand-engineered features. The quantity of capital letters, the length of the comments, and the amount of emoticons were among the handmade features. The emoji characteristics were not utilized in the methods in [15]. The study's keyword feature was a collection of individual terms that were retrieved using an NLP regular expression pattern and recognized as marketing or selling certain items. Ultimately, several combinations of the TF-IDF, Bag of Words, and FastText algorithms were used to extract and weight the text characteristics. XGBoost, SVM, and Naive Bayes were the three classification algorithms used.

International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 15, Issue 3, 2024
https://doi.org/10.62646/ijitce.2024.v15.i3.pp837-845

Using all three characteristics (features 1, 2, and 3) produced an F1 score of 96%, according to [15]. The study published in [15] indicates that the features used were heavily dependent on the dataset and are not generalizable to all fresh data, especially when it comes to keywords that were obtained by regular expressions.

There is currently little research on Instagram in particular on the identification of spam comments in Indonesian. The Naive Bayes (NB) algorithm was used in a research in [5] to identify Indonesian spam comments with a 72% accuracy rate. As an alternative, [6] utilized the Complementary Naïve Bayes (CNB) method, which was designed to handle an imbalanced dataset consisting of both spam and non-spam comments. Whereas SVM only managed 87% accuracy, the CNB algorithm was able to get 92% with more non-spam comments than spam. Table 1 presents recent research on social media spam detection, including techniques, findings, datasets, use of emojis, and post context. Table 1 shows that the majority of researchers used datasets that were produced privately.

Among the datasets that are accessible is SpamID-Pair, which is derived from social media. This dataset's distinguishing feature is the abundance of emojis that are used in the text. The fact that this dataset is made up of pairs of posts and comments that have been classified as spam or not makes it unique as well.

This dataset uses Instagram as its social media platform. The cause is that Instagram is a well-liked social networking platform with a large user base, including several celebrities. As a result, a lot of spam is found, particularly in the comments left by well-known people on Instagram.

Informal language, a lot of emoticons and emojis, typos and abbreviations, a lot of code mixes (mixed languages), comments of different lengths but generally brief (three to five words each), a post-reply structure devoid of hierarchical data, and mention tags (using the symbol "@") are all present in IG data. [9].

Pre-processing was almost the same as in many other research using text data. In order to identify spam comments or postings, the majority of pre-processing needed the use of NLP approaches. A number of sources, including [27], [28], and [29], emphasized the significance of text pre-processing prior to further processing. The techniques that were used were tokenization, case-folding, n-gram features, stemming, post-tagging, and stop-word elimination. Stemming approaches had the least impact, according to these pre-processing methods [29]. The text made up the majority of characteristics in several NLP study features, apart from pre-processing. Tokens feature in the form of BoW or weighted tokens in the form of TF-IDF were used in some studies [30].

III. SYSTEM ANALYSIS

EXISTING SYSTEM

Prior studies have been done on the identification of spam material. Text messages were the primary medium for spam identification [12], as shown by the Short Message Services (SMS) [13], [14], which used

International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 15, Issue 3, 2024
https://doi.org/10.62646/ijitce.2024.v15.i3.pp837-845

the CNN approach with the UCI SMS dataset, using additional characteristics that were hand-engineered [13]. Additionally, RNN-LSTM and LSTM alone were used to identify spam SMS and were compared to machine learning techniques [14]. Social media is full with spam stuff in addition to communications. You may find spam material on Facebook, Instagram, and Twitter [17].

Based on the English-language accounts of spammers on Instagram, article [4] found spam material. This research used Random Forest (RF) with specially hand-engineered addition features to recognize text content datasets comprising 1983 and 953808 media. The most important hand-engineered features are: a) whether or not other users are mentioned in the mention tags; b) how many hashtags are used, especially those unrelated to the content; c) whether or not repeated words are used; d) which keywords are often considered spam; and e) whether or not watermarks are present on images. With k=10 in k-fold validation and hand-engineered features, the outcome was 96.27%. Using attributes that required human extraction was one of the study's shortcomings.

In contrast to [4], study [15] used Indonesian instead of English and identified spam comments rather than spam postings. The Indonesian accounts dataset included in [15] was sourced from a publically accessible dataset. Nevertheless, the spam comments included in the research [15] were written in Indonesian and had promotional intent, unlike what the authors did (e.g., promoting items). Three strategies were used: 1) keyword, 2) content text, and 3) hand-engineered features. The quantity of capital letters, the length of the comments, and the amount of emoticons were among the handmade features. The emoji characteristics were not used in the methods in [15]. The study's keyword feature was a set of specified keywords that were retrieved using an NLP regular expression pattern and classified as either selling or promoting a certain product. Ultimately, several combinations of the TF-IDF, Bag of Words, and FastText algorithms were used to extract and weight the text characteristics. XGBoost, SVM, and Naive Bayes were the three classification algorithms used. Using all three characteristics (features 1, 2, and 3) produced an F1 score of 96%, according to [15]. The study published in [15] indicates that the features used were heavily dependent on the dataset and are not applicable to all fresh data, especially when it comes to keywords that are extracted by regular expressions.

There is currently little research on Instagram in particular on the identification of spam comments in Indonesian. The Naive Bayes (NB) algorithm was used in a research in [5] to identify Indonesian spam comments with a 72% accuracy rate. As opposed to this, [6] used the complementary Naive Bayes (CNB) method since it utilized an imbalanced dataset consisting of both spam and non-spam comments. While SVM only managed 87% accuracy, the CNB algorithm was able to get 92% when there were more non-spam comments than spam. Table 1 presents recent research on social media spam detection, including

International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 15, Issue 3, 2024
https://doi.org/10.62646/ijitce.2024.v15.i3.pp837-845

techniques, findings, datasets, use of emojis, and post context. Table 1 shows that the majority of researchers used datasets that were produced privately.

Among the datasets that are accessible is SpamID-Pair, which is derived from social media. This dataset's distinguishing feature is the abundance of emojis that are used in the text. Another unique feature of this dataset is that it is made up of pairs of comments and posts that have been classified as spam or not. This dataset uses Instagram as its social media platform. The reason for this is that Instagram is a widely used social media platform that is frequented by celebrities. As a result, a lot of spam is found, particularly in the comments left by well-known people on Instagram. Informal language, a lot of emoticons and emojis, typos and abbreviations, a lot of code mixes (mixed languages), shorter comments (three to five words long), a post-reply structure devoid of hierarchical data, and mention tags (using the @ symbol) are all present in IG data. [9].

Disadvantages

- In order to improve the performance of the least effective classification algorithm, the system used a boosting strategy.
- In most cases, the classification algorithms used in preexisting systems are not robust and are prone to being stuck in overfitting situations.

PROPOSED SYSTEM

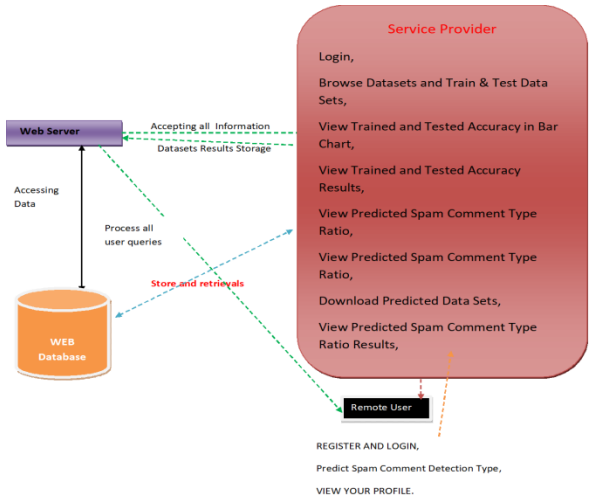
In this work, the authors compared and analyzed the SpamID-Pair dataset, which was collected from 12 celebrities with more than 15 million followers [11], using several machine learning methods, such as Complement Naïve Bayes (CNB) and Extra Tree (ET), in line with [10]. The research contributed by offering thorough experimental results of spam detection performance (accuracy and F1) between nonemoji and emoji features, as well as their analysis, with various combinations of hyperparameter scenarios (n-grams features, balanced/unbalanced data, the use of comment-only/post-comment pairs approach) [10]. Furthermore, this research introduces a new way to detect spam comments contextually by feeding the machine learning model the content of both the post and the comment. Using the Indonesian SpamID-Pair dataset, this research applies natural language processing techniques.

Advantages

- Because it incorporates data normalization, emotion handling, and manual features, the system is more successful.
- By analyzing the SpamID-Pair dataset, the system is able to identify further system benefits.

IV. SYSTEM ARCHITECTURE

International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 15, Issue 3, 2024
https://doi.org/10.62646/ijitce.2024.v15.i3.pp837-845



V. SYSTEM IMPLEMENTATION

Modules

Service Provider

A valid username and password are required for the Service Provider to access this module. After he successfully logs in, he will be able to perform things like browse datasets and run tests and training on them. Check out the Bar Chart for Trained and Tested Accuracy, Check Out the Results for Trained and Tested Accuracy, Access the following: All Remote Users, Forecasted Spam Comment Type Ratio, Forecasted Data Sets for Download, Forecasted Results for Download, and Forecasted Data Sets for Viewing.

View and Authorize Users

The admin can get a complete rundown of all registered users in this section. Here, the administrator may see the user's information (name, email, and address) and grant them access.

Remote User

All all, there are n users in this module. Registration is required prior to performing any operations. Details will be entered into the database after a user registers. He will need to log in using the permitted username and password when registration is completed. After logging in, users will be able to perform things like PREDICT SPAM, REGISTER AND LOGIN, and more. Sort of Remark Detection, Check Out Your Account.

VI. CONCLUSION

The objective of this study was to improve social media spam comment identification by extensive testing and analysis using a range of test cases. This study was different from others in that it solely identified spam based on the content of the comments rather than using the emoji feature in its detection process. In order to identify the best approach, set of circumstances, and characteristics, this research examined the characteristics of emojis and post-comment pair data.

ISSN 2347–3657
Volume 15, Issue 3, 2024
https://doi.org/10.62646/ijitce.2024.v15.i3.pp837-845

The goal of the experiment was to ascertain the value of emoji features—which are often disregarded in many NLP forms of research—using 14 cutting-edge machine learning models with a variety of situations and the Spam ID-Pair dataset. To improve the speed, we also looked at using post-comment pairs of TF-IDF vectors stacked horizontally. The outcomes show how the different situations perform in terms of accuracy and F1 scores, as well as how they compare. Spam comment detection on social media might be improved by the text emoji feature, as shown by the 4% to 6% average increase in performance using machine learning techniques. It was also shown that post-comment pairs data enhanced detection performance by an average of 0.7% to 2.11%. To the best of our knowledge, this is the first instance of spam comment identification based on the post and comment together, particularly when it comes to Indonesian social media users. On average, adding manual characteristics might improve detection performance by 1.35% to 2.18%. Using the C-PCTM and C-PCTMB scenarios, the SVM-RBF, RF, and ET algorithms were the most effective approaches for detecting spam comments. Compared to a single classifier, the ensemble soft voting technique had the best average performance in terms of accuracy and F1 score. In production mode, it may be used. Its large model, as opposed to each/single model without the ensemble approach, is a drawback, nevertheless. In summary, the performance was enhanced by the use of emojis, a post-comment pairs strategy, and balanced-manual features in both comments and pairs of comments.

However, employing machine learning, this study may not yet completely comprehend the context between postings and comments. Further research is still needed to discover the semantic link using an appropriate model and methodology. Understanding the context of posts and comments is essential for determining the relevance of comments to posts. This will improve the accuracy and F1 score of spam comment detection. In order to ascertain their importance, we plan to employ the deep learning model in phrase pairs classification adaption [49] and adjustment between post and comment vector representations. Spam usually appears as a remark that is unrelated to the topic.

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.

ISSN 2347–3657
Volume 15, Issue 3, 2024
https://doi.org/10.62646/ijitce.2024.v15.i3.pp837-845

[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.