SPAMID-PAIR: A Novel Indonesian Post–Comment Pairs Dataset Containing Emoji

Antonius Rachmat Chrismanto[1], Anny Kartika Sari[2]*, Yohanes Suyanto[3]
Department of Computer Science and Electronics, Faculty of Matematics and Natural Science, Universitas Gadjah Mada[1,2,3]
Faculty of Information Technology, Universitas Kristen Duta Wacana, Yogyakarta, Indonesia[1]

***Abstract*—The detection of spam content is an important task especially in social media. It has become a topic to be continuely studied in Natural Language Processing (NLP) area in the last few years. However, limited data sets are available for this research topic because most researchers collect the data by themselves and make it private. Moreover, most available data sets only provide the post content without considering the comment content. This data becomes a limitation because the post-comment pair is needed when determining the context of a comment from a particular post. The context may contribute to the decision of whether a comment is a spam or not. The scarcity of non-English data sets, including Indonesian, is also another issue. To solve these problems, the authors introduce SPAMID-PAIR, a novel post-comment pair data set collected from Instagram (IG) in Indonesian. It is collected from selected 13 Indonesian actress/actor accounts, each of which has more than 15 million followers. It contains 72874 pairs of data. This data set has been annotated with spam/non-spam labels in Unicode (UTF-8) text format. The data also includes a lot of emojis/emoticons from IG. To test the baseline performance, the data is tested with some machine learning methods using several scenarios and achieves good performance. This dataset aims to be used for the replicable experiment in spam content detection on social media and other tasks in the NLP area.**

Keywords—Dataset; natural language processing; spam detection; spamid-pair; post-comment pairs

I. Introduction

Research on text analysis, especially in context-based detection/classification problems, is increasingly important because of the higher need for system automation. A labeled dataset is needed for supervised text classification to be used as machine learning data. Unfortunately, the datasets for text classification are mainly in English. Other languages (Turki [1], Bangla [2], Chinese [3], Arab [4], and Morocco [5]), including Indonesian, are rare enough [6], [7].

Datasets for text classification can be divided into two types: single-text classification and paired-text classification. Some examples of single-text classification datasets are news classification, sentiment classification, hoax classification, spam, topic classification, and emotion text classification. Examples for paired text classification are text entailment classification, duplicate question classification, text pair similarity classification, including spam comment classification based on a particular post on the social media. One of the challenges in the NLP area is how to understand the context to gets the meaning. Context understanding can also be applied to spam comment detection based on its post by detecting the comment's relevance. If the comment is not related/relavance to its post, it is likely to be categorized as spam. To detect spam comments, machine learning methods require training datasets that can be used according to the context, such as in the context of the language, that are still rare.

The motivation of this research is to overcome the datasets scarcity in Indonesian for the text pairs classification to get the context between two texts in pairs correctly. The authors have collected the dataset for spam comment detection based on social media posts. This dataset is taken from Instagram (IG) based on selected 13 public Indonesian artists/actors inspired by [8], [9]. Each of the public Indonesian artists/actors has more than 15 million followers. Each row of this dataset consists of a post and comments text pair called SPAMID-PAIR[1]. SPAMID-PAIR contains 72874 pairs of posts and comments and breaks down into 53837 non-spam data and 19037 spam data.

This article introduces the SPAMID-PAIR dataset, a novel dataset collected, labeled, validated, and used as training data for spam comment detection based on their posts with several machine learning methods. This dataset is intended to contribute as one of the Indonesian datasets in NLP for text pair classification problems based on the context. The SPAMID-PAIR dataset has an advantage because it contains symbols, special characters, and emojis that are widely available in social media posts and comment texts. This dataset is useful for NLP research because most researchers discard emojis in their classification techniques. Some examples are news article classification [10], Twitter without emoji [11], spam comments from the blog [12], Twitter (removed emoji) [13], SMS and Twitter without emoji [14], Twitter without emoji [15], Youtube comment without emoji [16], video spam comment without emoji [17], Youtube comment without emoji [18]. The emoji is essential because most social media users use emojis to express their feelings, such as to support/deny, show sympathy, joy, sadness, and anger. The emojis in the dataset is needed for research in some fields that learn through emoji expression. This dataset uses the UTF-8 format for post and comment data, so both emojis can be used in emoji pairs expression research.

*Corresponding Author.
KemendikbudRistek DIKTI Indonesia Doctoral Research Grant No. 1897/UN1/DITLIT/Dit- Lit/PT.01.03/2022.

[1] This dataset is available at Mendeley Dataset Repository (https://data.mendeley.com/datasets/fj5pbdf95t)

The contribution of this paper is two-fold; first, the novel SPAMID-PAIR dataset, and second, several machine learning algorithms will be used to implement the supervised text-pair classification using this dataset using the F1 score. This paper is written as follows, firstly, the introduction of SPAMID-PAIR and its purpose. Secondly, the related works of the Indonesian NLP dataset, the experiments and results using this dataset, and finally, the conclusion.

II. Related Works

Datasets are the primary data source in machine/computer learning. Various machine learning and deep learning techniques are in dire need of data sources for system learning. But in reality, not all public datasets are available, especially in Natural Language Processing (NLP). Even though learning datasets in NLP are quite widely available in English, such as IMDB Dataset [19]–[21], SMS Spam UCI [22], FLAIR [23], [24], Twitter Spam [25], [26], YouTube Comments [15], PeerRead [27], and Huggingface Community Datasets [28]. Still, there are few public datasets in other languages, especially Indonesian.

IndoNLU [6] is one of two dataset sources in the field of Natural Language Understanding (NLU) for 12 main tasks that have been attempted to be collected in collaboration with universities and industry. IndoLEM [29], as the second source, is a dataset source in NLP for seven main tasks (post tagging, named entity recognition, parsing, sentiment analysis, summarization, and word prediction). IndoLEM, the second, provides datasets, Indonesian Fasttext, and BERT pre-trained that can be used for other tasks. To the best of our knowledge, unfortunately, for the case of the semantic task in detecting spam comments in social media based on the context of the post in pairs, it has not been found. This article introduces SPAMID-PAIR to enrich the Indonesian NLP dataset collection in spam comment detection based on its post context, which has not been done before.

Spam text detection on social media is mostly done on Twitter [11], [13], [14], [15]. Twitter has a structure that is not in the posts and comments pair structure. Otherwise, Youtube, Facebook, and Instagram are examples of social media with posts and comments pair structures. However, the detection of spam comments in previous studies was not based on paying attention to the post. The previous research used some popular machine learning methods. Septiandri and Wibisono use Naïve Bayes, SVM, and XGBoost to detect spam comments from Instagram, and SVM outperformed the others [30]. Zhang uses the Random Forest to detect Instagram spam posts and achieve good [31]. Research [32] investigated 11 state-of-the-art machine learning methods in text classification using 71 datasets and obtained that Stochastic Gradient Boosting, SVM, and Random Forest were the best methods compared to the others. That research can be used as a reference for the best machine learning in classification.

III. Research Method

This research uses the following steps: data acquisition, dataset construction, data profiling, annotation/labeling, pre-processing, feature extraction/generation, ML algorithm implementation, and evaluation. These stages can be seen in Fig. 1, while a more detailed explanation is in the following sub-chapters.

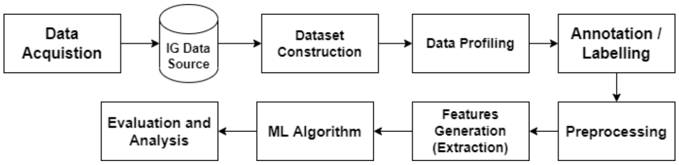


Fig. 1. The Research Method.

A. Data Acquisition

In the data acquisition stage, IG was chosen because 1) IG has a lot of spam comments, especially on Indonesian public figure accounts [33], [34]. 2) Posting and commenting on IG is in pairs suitable for the pair dataset; 3) IG has a lot of non-formal posts and comments, and it also contains a lot of emojis; 4) IG does not have a spam filtering feature in Indonesian yet. For comparison, on Twitter (TW), a tweet is a post, but replies from other users must always use a mention tag, so the form of the reply is not a comment. The reply data is equivalent to the tweet, not as a child node. On Facebook (FB), a user can create a status/post, and others can comment on it. But on FB, the situation tends to be more formal/serious, so it does not contain much spam and emojis. Nowadays, IG is a famous social media with many young IG users; not as serious and formal as FB.

Comparing three leading social media existing today, e.g., IG, TW, dan FB, IG is the best choice for collecting datasets for spam detection. IG is widely used by public figures such as politicians, artists/actors, and well-known people. Very limited datasets are available in languages other than English and Chinese, especially Indonesian [6], making collecting this dataset more critical. The SPAMID-PAIR dataset from IG contains post-comment pairs from 13 Indonesian artists/actors with more than 15 million followers without stating their account names. It is expected that researchers in the NLP field can use this dataset to replicate research and use it as the dataset reference in the topic of spam detection using various algorithms.

The SPAMID-PAIR dataset was retrieved using several tools such as Instaloader and Chrome Selenium Python driver. For the first planning, the data is taken from the 50 most recent posts, and 120 most recent comments are taken from each post. Hence, it was estimated that 78000 data could be collected. However, the data is not as planned in reality because some posts do not have as many comments as expected. The dataset was collected in September 2020, and after data retrieval was completed, 72874 pairs of post and comment data were obtained, which are ready for further processing.

Table I displays all the artists/actor's usernames used in the SPAMID-PAIR dataset. SPAMID-PAIR contains 72874 pairs of posts and comments and breaks down into 53837 non-spam data (73.87%) and 19037 spam data (26.13%). Details of the number of spam and non-spam labels per artist/actor are highlighted in Table II, analyzed using Python Pandas. Table II also shows that the IG ID 24239929 only has 103 data because the user recently had disabled comments, so the data could not be retrieved anymore. Spam comments are detected in all 13

IG users chosen with varying percentages. The SPAMID-PAIR dataset consists of 11 fields and is available in Excel format (.xlsx) and comma-separated value (CSV) with UTF-8 encoding, as described per field in Table III.

TABLE I. THE 13 PUBLIC FIGURES USED IN THE SPAMID-PAIR DATASET WITH MORE THAN 15 MILLION FOLLOWERS (PER DECEMBER 2021)

Account ID	Followers (millions)
1918078581	54.3
522969993	47.4
225064794	42.4
24239929	36.4
2993265	34.1
361869464	33.6
26444210	33.4
1948416	30.7
305384601	27.3
8115577	27.1
5735890	25.8
4934196	25.2
30585021	15.7

TABLE II. DETAILED STATISTICS OF SPAM AND NON-SPAM DATA PER ACCOUNT ID IN THE SPAMID-PAIR DATASET

Account ID	Count of Non-Spam	Count of Spam	%Non-Spam	% Spam	Sub Total
4934196	4565	2251	66,97	33,03	6816
522969993	5712	1108	83,75	16,25	6820
5735890	3397	691	83,10	16,90	4088
30585021	818	1065	43,44	56,56	1883
2993265	4528	2022	69,13	30,87	6550
1948416	4658	1945	70,54	29,46	6603
361869464	6854	2466	73,54	26,46	9320
225064794	4944	1804	73,27	26,73	6748
24239929	65	38	63,11	36,89	103
1918078581	5045	1557	76,42	23,58	6602
8115577	4818	1971	70,97	29,03	6789
26444210	5537	911	85,87	14,13	6448
305384601	2896	1208	70,57	29,43	4104
Total	53837	19037			72874

Table IV shows that the number of emojis in this dataset reaches 68%, and the number of emojis in the spam category is higher than in the non-spam category. Table V shows detailed data related to emoji statistics in the dataset. Fig. 2 illustrates the distribution of emoji in the SPAMID-PAIR dataset per IG artist ID and tells us how the emoji is related to the spam or non-spam label. Fig. 3 shows some correlation between some attributes of the SPAMID-PAIR dataset. First, it shows a correlation between the length of comments and spam labels. There is also a correlation between the length of comments and the number of emojis. Lastly, there is a correlation between the length of comments and the post length.

TABLE III. DESCRIPTION OF ATTRIBUTES IN THE SPAMID-PAIR DATASET

Attribute	Description
igid	Account ID
comment	Comment on a post
post	Post from an account ID
emoji	Whether the data contains emojis or not (1 or 0)
spam	Whether the data is spam or not (1 or 0)
lengthcomment	The character length of the comment
lengthpost	The character length of the post
countemojicomment	Number of emoji symbol characters in comments
countemojicommentuniq	Number of emoji symbol characters in comments (unique)
countemojipost	Number of emoji symbol characters in posts
countemojipostuniq	Number of emoji symbol characters in the post (unique)

TABLE IV. NUMBER OF EMOJIS IN THE SPAMID-PAIR DATASET

Category	Count	Percentage (%)
Non-Emoji	22710	31,16
Emoji	50164	68,83

TABLE V. NUMBER OF EMOJI IN THE SPAMID-PAIR DATASET PER ACCOUNT ID

Account ID	Count of Non-Emoji	Count of Emoji	% Non-Emoji	% Emoji	Sub Total
4934196	2085	4731	30,59	69,41	6816
522969993	1679	5141	24,62	75,38	6820
5735890	1013	3075	24,78	75,22	4088
30585021	1142	741	60,65	39,35	1883
2993265	2482	4068	37,89	62,11	6550
1948416	1857	4746	28,12	71,88	6603
361869464	3264	6056	35,02	64,98	9320
225064794	1935	4813	28,68	71,32	6748
24239929	2	101	1,94	98,06	103
1918078581	2052	4550	31,08	68,92	6602
8115577	2126	4663	31,32	68,68	6789
26444210	1592	4856	24,69	75,31	6448
305384601	1481	2623	36,09	63,91	4104
Total	22710	50164			72874

SPAMID-PAIR dataset profile generally has an average comment length of 34.23 characters and an average post length of 252.03 characters. The highest number of emojis (non-unique) in a comment is 359 emojis, and the highest number of unique emojis is 112. In the post data, the highest number of emojis (non-unique) is 32 emojis, and the highest number of unique emojis is 14. Complete statistical details of the comment and post data can be seen in Table VI. The maximum length of the comment is 386, and the post is 3938.

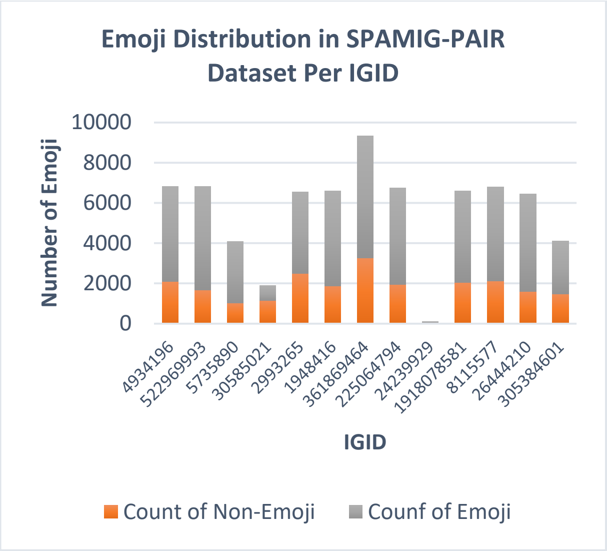


Fig. 2. Distribution of Emoji in the SPAMID-PAIR Dataset.

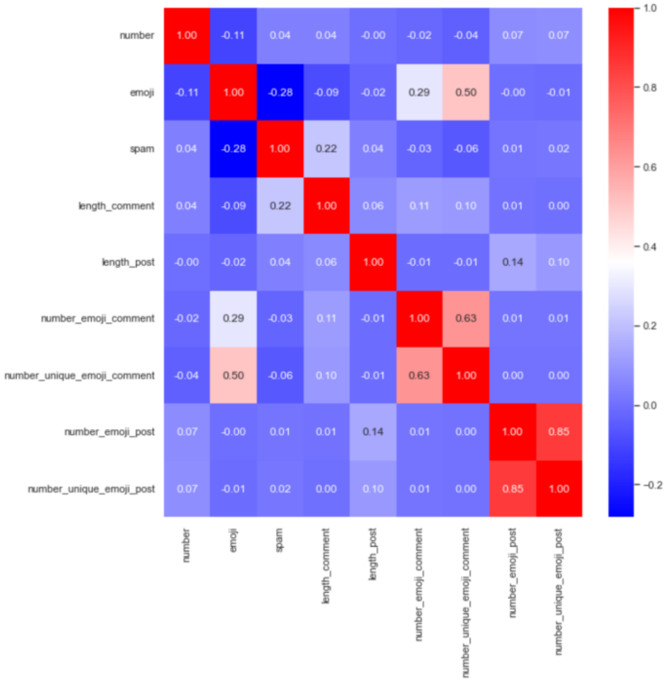


Fig. 3. Attributes Correlation in the SPAMID-PAIR Dataset.

B. Data Profiling and Labelling

After the dataset has been collected, the next step is data profiling, labeling, and validation. Labeling gives each data a "spam" or "not spam" label. The "spam" criterion is given if the post and comment data, text data, or emojis are irrelevant. On the other hand, the "not spam" criteria will be given if the post and comment data are relevant. Two Indonesian labelers carried out the labeling process. Before starting the labeling process, a joint briefing was held between the two native Indonesian labelers to create a common perception of the meaning of "spam" and "non-spam" labels. After that, labeling was done using an excel formatted dataset that was given to each labeler, and there was one additional column, "label," which would be filled with "spam" or "not spam" by each labeler manually. The final label was determined by the final agreement of the two labelers. Based on the Kappa score, the result is the "almost perfect" category with a Kappa score of 0.95, proving that the labeling agreement between the two annotators was relatively easy. The difficulty arises when the comment contains only an emoji, and it is difficult to determine its meaning. However, it can be overcome by looking at the consistency of the type of emoji and the type of "positive" emoji used. Suppose the emojis use "positive" emojis such as expressions of joy, enthusiasm, support, and love. In that case, the label is a high possibility of "not spam." Otherwise, if the post content tends to be "positive" and the comment content tends to be "negative," it is labeled as "spam." Examples of labeling results for data labeled as "spam" and "not spam" can be seen in Table VII.

TABLE VI. STATISTICAL INFORMATION ON COMMENTS AND POSTS DATA IN THE SPAMID-PAIR DATASET

Statistics of Comments	**Average**	**Max**	**Statistics of Post**	**Average**	**Max**
Number of sentences	1,1	29	Number of sentences	2,88	45
Number of characters	34,25	212	Number of characters	2,52	3938
Number of whitespaces	4,4	386	Number of whitespaces	33,13	570
Number of words	4,8	384	Number of words	35,33	602
Number of numbers (as a whole)	0,2	72	Number of numbers (as a whole)	1,1	35
Number of punctuations	1,1	213	Number of punctuations	10,52	192
Number of date format	0,00013	1	Number of date format	0,000618	1

TABLE VII. EXAMPLE OF LABELING RESULTS

Comment	**Post**	**Label**	**Reason**
😢😢😢😢 😢🔥🔥🔥 🔥🔥🔥	Can't argue with the clan 😎! Entertainment Inc presents u @USER. Watch the full version on my video	Spam	The comment contains only emojis that are not consistent, "sad and hot" at the same time about the new post video
cantik bangettt😩 🖤 (in English: veryyy pretty 😩 🖤)	🌸 Outfits custom @USER Styling @USER Makeup @USER Hair @USER Photographer @USER	Not spam	The comment reply a post about how pretty an artist is because the post shows how beautiful the artist in an outfit

After the labeling had been completed and re-validated, a data profiling step was carried out to determine additional data from the dataset using Python NLP Profiler. It analyses whether there are emojis or not in the posts or comments. It does statistical analysis on the number of sentences, the number of characters, the number of whitespaces, the number of words, the number of words in the form of numbers, and the number of signs. It also reads and counts the number date format. Data profiling is used to determine the characteristics of the data and assist in determining the appropriate pre-processing steps later.

C. Pre-processing

The pre-processing process consists of the following steps:

1) Generating manual features such as the length, the number of emojis, the number of unique emojis, the number of digits, the number of hashtags, the number of mentions, the number of uppercase letters, the number of special chars, and the number of links.

2) Changing letters to lowercase.

3) Removing spaces and characters that appear excessively.

4) Removing certain punctuation marks unrelated to hashtags, emails, mentions, and URLs.

5) Doing simple normalization as follow:

a) Repeated words normalization (such as "pergi2" to "pergi-pergi").

b) Slang words normalization using a dictionary.

c) Email, hashtag, number, mention to specific TAG (USER, ANGKA, EMAIL, MENTION)

d) Abbreviation normalization using a dictionary.

e) Some minor spelling corrections using a dictionary.

6) Performing stopwords removal (using combined stopwords from standard and stopwords generated from the dataset based on their frequency).

7) Performing stemming using the Sastrawi Python library.

8) Saving the final output and passing it to the model.

IV. Experiments and Discussion for Baseline Performance

The testing was carried out using the ML method (Nave Bayes, Complement Naïve Bayes, Decision Tree, and Multi-Layer Perceptron) [32], which was partially or fully implemented using the Python Scikit Learn library (Sklearn). The test scenario was carried out in two forms: a dataset with emoji in symbols and emoji in the text. Pre-processing uses tokenization, Indonesian stopwords, and simple normalization and uses the n-gram TF-IDF features, i.e., 1-gram and 2-gram. Table VIII shows the experiment scenario using the machine learning methods. The dataset splits into 80% training data and 20% testing data. The evaluation score used F-measure (F1) with a score between 0-1. The measurement matrix uses the F1 score with 80% training data and 20% testing.

The authors use Naïve Bayes (NB) with an alpha value of 0.01 and other parameters defaulted from sklearn. Complement Naïve Bayes (CNB) was used in the second experiment, which was expected to overcome datasets whose classes are not balanced. Both methods were used as representatives of the probability classifier. The Decision Tree (DT) method was also used, representing the classifier tree with the random_state parameter set and the information gain using the Gini index. Finally, an artificial neural network-based classification method was used: a multi-layer perceptron with a limited iteration of 300. The F-measure (F1 score) was chosen for the performance evaluation because the F1 score value represents a combination of recall and precision values and can also be used in the unbalanced dataset.

Moreover, the accuracy value alone is inappropriate for the SPAMID-PAIR dataset with an unbalanced number of classes. Table X shows the results of the experimental scenarios using the methods. Fig. 4 to 10 display the confusion matrixes of the models, while Fig. 11 and 12 show ROC curves of the models in testing data. From the confusion matrix in Fig 4(a) and 4(b) (EmojiSymbol NB), all the true positives are higher than the others (true negative, false positive, and false negative). The ability to detect spam comments is good enough, but it also can be seen that the accuracy is better on not-spam comments than on spam comment labels. Fig. 5(a) and 5(b) (EmojiText NB) show that the F1 score is higher than the EmojiSymbol, although the true positives are lower than the EmojiSymbol. From this result, the EmojiText performs better because it can detect spam comment properly in a balanced dataset. Fig. 6(a) and 6(b) (EmojiSymbol CNB) show that the F1 score (based on true positive, true negative, false positive, and false negative) is better than the NB method. CNB works better because it can complement the weight of an unbalanced dataset [34]. Fig. 7(a) and 7(b) (EmojiText CNB) show that the CNB in text format outperforms the NB in EmojiSymbol and EmojiText. Fig. 8(a), 8(b), 9(a), and 9(b) shows the performance of the DT method that also has better F1 in EmojiText but not for the EmojiSymbol. Decision Tree can handle the emoji symbol well. The last, in Fig. 10(a) and 10(b), it can be seen that the confusion matrix shows that MLP (a traditional neural network) has close F1 score to NB and CNB but trains slower than them. But, based on Fig. 11(a) and 11(b), it can be seen that EmojiText in MLP works the best from the other methods.

The authors also extract a list of emojis categorized as 'spam' and 'not spam' based on the SPAMID-PAIR dataset. It can be seen in Table IX. It can be seen that list of spam emojis is more than not spam emojis. The intersection between them is also quite a lot, and the emoji only used in the "not spam" category contain very reasonable emojis (clear emoji meaning). Still, on the other hand, the emoji used only in the "spam" category is quite a lot and very random emojis (not clear emoji meaning).

TABLE VIII. The Testing Scenario on SPAMID-PAIR using the Machine Learning Methods

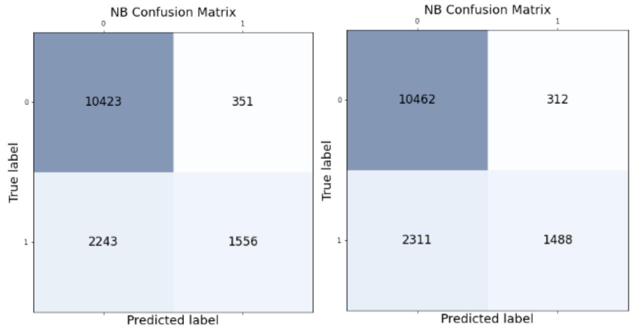
Test Scenario Using Machine Learning Methods		
Emoji Symbol	Pre-processing: tokenization, stopwords, normalization, stemming, feature: TF-IDF 1 gram and 2 gram	Methods: Naïve Bayes (NB) (alpha: 0.01), Complement Naïve Bayes (CNB) (alpha: 0.01, norm: true), Decision Tree (DT) (random_state: 42, gain: Gini), Multi-layer Perceptron (MLP) (random_state: 42, max_iter: 300)
Emoji in Text	Pre-processing: tokenization, stopwords, normalization, stemming, feature: TF-IDF 1 gram and 2 gram	Methods: Naïve Bayes (NB) (alpha: 0.01), Complement Naïve Bayes (CNB) (alpha: 0.01, norm: true), Decision Tree (DT) (random_state: 42, gain: Gini), Multi-Layer Perceptron (MLP) (random_state: 42, max_iter: 300)

TABLE IX. EMOJI LIST (SPAM AND NOT SPAM), INTERSECTION, AND DIFFERENCE IN THE SPAMID-PAIR DATASET

List of Emojis	Category
[illegible]	Spam (S)
[illegible]	Not Spam (NS)
[illegible]	S ∩ NS
[illegible]	NS minus S
[illegible]	S minus NS

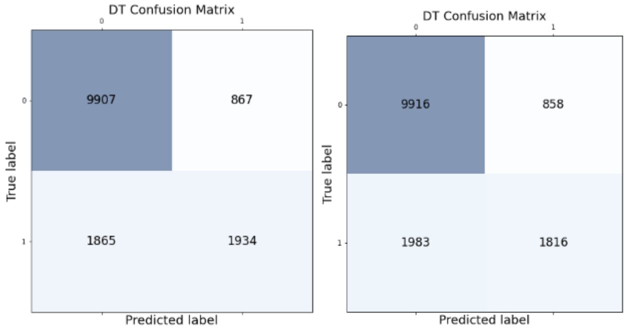
TABLE X. F-MEASURE (F1) SCORE RESULTS OF THE EXPERIMENTAL SCENARIOS

Scenario	NB	CNB	DT	MLP
EmojiSymbol1GramTFIDF	.74	.75	.72	.74
EmojiSymbol2GramTFIDF	.74	.75	.72	.74
EmojiText1GramTFIDF	.77	.78	.78	.80
EmojiText2GramTFIDF	.78	.80	.78	.80



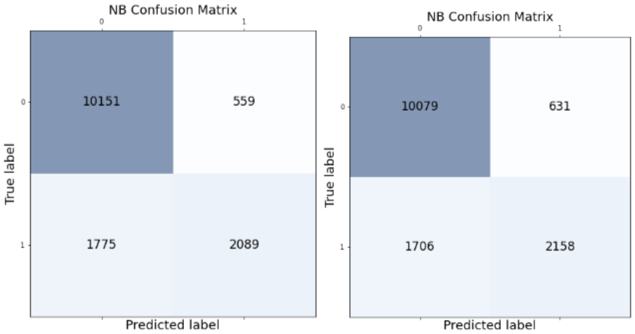
(a) (b)

Fig. 4. Confusion Matrix Naïve Bayes of (a) EmojiSymbol1GramTFIDF (b) EmojiSymbol2GramTFIDF.



(a) (b)

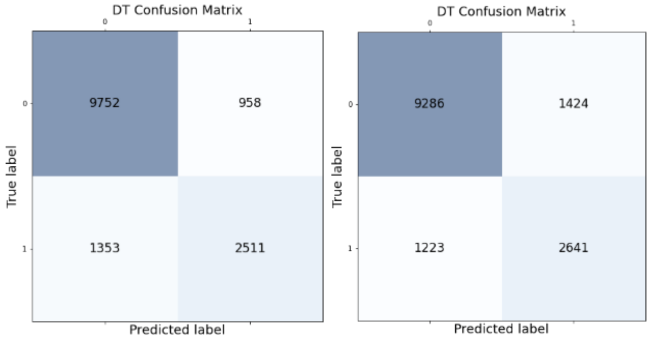
Fig. 8. Confusion Matrix Decision Tree of (a) EmojiSymbol1GramTFIDF (b) EmojiSymbol2GramTFIDF.



(a)

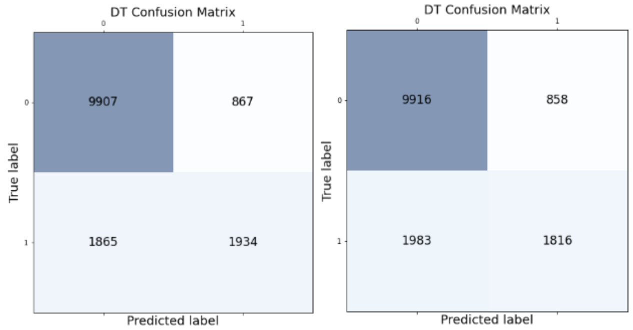
(b)

Fig. 5. Confusion Matrix Naïve Bayes of (a) EmojiText1GramTFIDF (b) EmojiText2GramTFIDF.



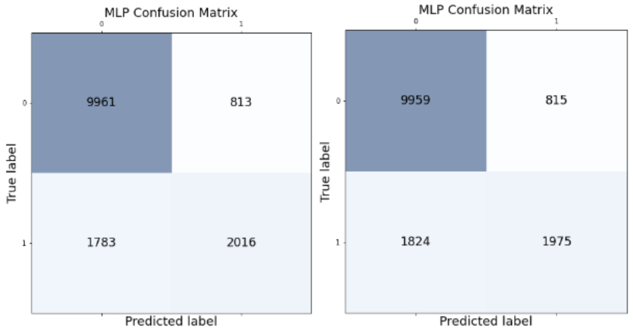
(a) (b)

Fig. 9. Confusion Matrix Decision Tree of (a) EmojiText1GramTFIDF (b) EmojiText2GramTFIDF



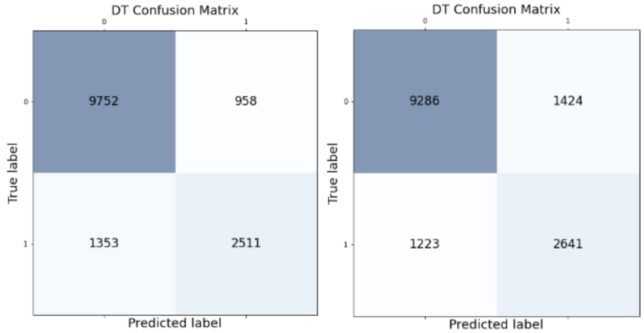
(a) (b)

Fig. 6. Confusion Matrix Complement Naïve Bayes of (a) EmojiSymbol1GramTFIDF (b) EmojiSymbol2GramTFIDF.



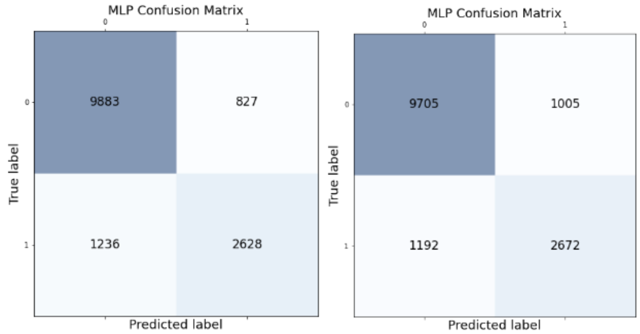
(a) (b)

Fig. 10. Confusion Matrix Multi-layer Perceptron of (a) EmojiSymbol1GramTFIDF (b) EmojiSymbol2GramTFIDF



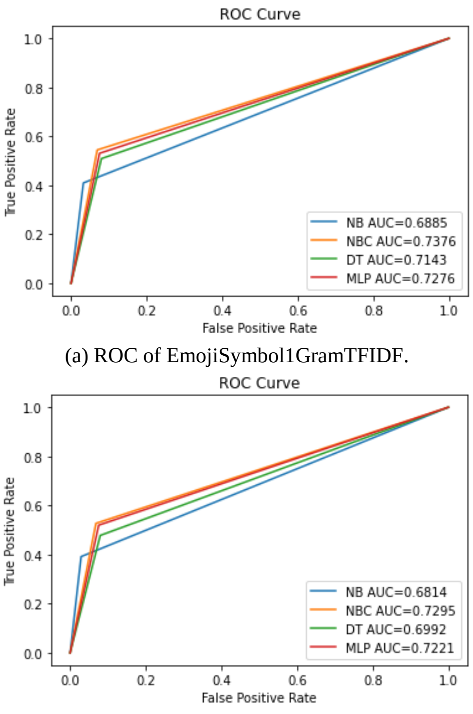
(a) (b)

Fig. 7. Confusion Matrix Complement Naïve Bayes of (a) EmojiText1GramTFIDF (b) EmojiText2GramTFIDF.



(a) (b)

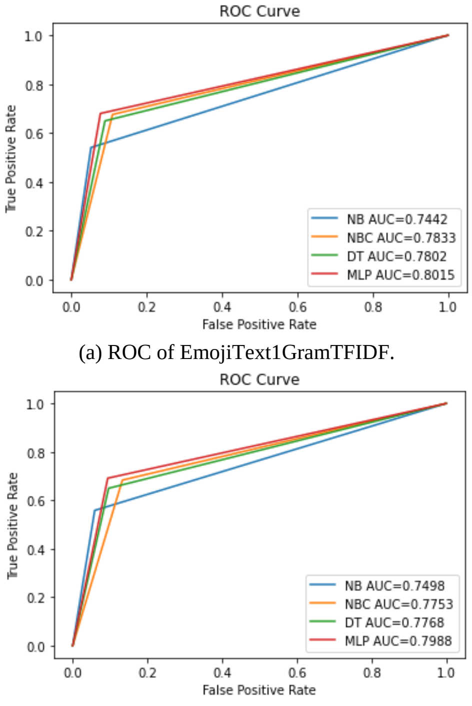
Fig. 11. Confusion Matrix Multi-layer Perceptron of (a) EmojiText1GramTFIDF (b) EmojiText2GramTFIDF



(a) ROC of EmojiSymbol1GramTFIDF.

(b) ROC of EmojiSymbol2GramTFIDF.

Fig. 12. ROC Curve of (a) EmojiSymbol1GramTFIDF (NB, CNB, DT, MLP) and (b) EmojiSymbol2GramTFIDF (NB, CNB, DT, MLP).



(a) ROC of EmojiText1GramTFIDF.

(b) ROC of EmojiText2GramTFIDF.

Fig. 13. ROC Curve of (a) EmojiText1GramTFIDF (NB, CNB, DT, MLP) and (b) EmojiText2GramTFIDF (NB, CNB, DT, MLP)

The results in Table X prove that the SPAMID-PAIR dataset is a dataset that can be used in Indonesian text classification experiments originating from social media. In Fig. 4 to Fig. 10, all the confusion matrixes of the models use 14.573 (20%) data testing. From Table X and Fig. 4 to Fig 12, It can be seen that CNB and MLP are superior to NB and DT. Fig. 13 shows the ROC curve, which explains that the area of the ROC curve in 13(a) is higher than in 13(b). The EmojiText1Gram is better than the EmojiText2Gram because the TFIDF vectors from 1gram have a better weight representing the text's characteristics. The traditional ML can only achieve an F1 score in the range of 0,72-0,78, but a multi-layer perceptron can achieve an F1 score of 0,8. It promises that these results can be improved, such as with the pair context classification approach [35]. Hopefully, this dataset can also be used in other related research and enrich the Indonesian dataset collection, which is still rare. This dataset is also important because it contains pairs of posts and comments that can be related and used in problem sentence pair classification in Indonesian.

V. Conclusion

This research collected post and comment pairs data from 13 selected Indonesian public figures (artists) / public accounts with more than 15 million followers. Two persons labeled all pair data as an expert in 72874 data. The dataset is called SPAMID-PAIR, containing post-comment pairs and label in Unicode text (UTF-8) text containing emojis. The dataset does not include any account information except the ID number. Unlike the other existing sentence pair datasets, the SPAMID-PAIR dataset is specifically used to determine the context between comments and posts that have never been collected in a large enough dataset. The objective of this dataset is as the primary data source in machine learning, especially in the NLP area, for spam comments detection based on the post context. This dataset is intended as one of the Indonesian language datasets that also contains many emoji symbols from social media so that it can be used to understand human expressions using emojis.

SPAMID-PAIR proved that it could be used as a training dataset to detect spam comments based on its post. From the experimental research using some ML methods, it can be seen that ML can only achieve an F1 score in the range of 0,72-0,78, but a multi-layer perceptron (MLP) can achieve an F1 score of 0,8. It significantly promises that these results can be improved in future works. The limitation of this dataset is it includes imbalanced data between not spam and spam categories. This dataset can also be enhanced in the future.

Acknowledgment

KemendikbudRistek DIKTI Indonesia supported this work for the Doctoral Research Grant No. 1897/UN1/DITLIT/Dit-Lit/PT.01.03/2022. The authors thank Gadjah Mada University, especially the Department of Computer Science and Electronics and Information Technology Faculty at Duta Wacana Christian University, for all the support.

References

[1] A. E. Yüksel, Y. A. Türkmen, A. Özgür, and A. B. Altınel, “Turkish tweet classification with transformer encoder,” in International Conference Recent Advances in Natural Language Processing, RANLP, 2019, vol. 2019-Septe, pp. 1380–1387. doi: 10.26615/978-954-452-056-4_158.

[2] T. Alam, A. Khan, and F. Alam, "Bangla Text Classification using Transformers," arXiv, Nov. 2020, [Online]. Available: http://arxiv.org/abs/2011.04446.

[3] C. He and Y. Shi, "Research on Chinese spam comments detection based on Chinese characteristics," in 2018 IEEE 4th International Conference on Computer and Communications, ICCC 2018, 2018, pp. 2608–2612. doi: 10.1109/CompComm.2018.8781051.

[4] A. M. Gaber, A. M. Gaber, and H. Moussa, "SMAD: Text Classification of Arabic Social Media Dataset for News Sources," Int. J. Adv. Comput. Sci. Appl., vol. 12, no. 10, pp. 508–516, 2021, doi: 10.14569/IJACSA.2021.0121058.

[5] S. Mihi, B. A. BEN Ali, I. EL Bazi, S. Arezki, and N. Laachfoubi, "MSTD: Moroccan sentiment twitter dataset," Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 10, pp. 363–372, 2020, doi: 10.14569/IJACSA.2020.0111045.

[6] B. Wilie et al., "IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding," arXiv, Sep. 2020, [Online]. Available: https://www.aclweb.org/anthology/2020.aacl-main.85.

[7] 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.

[8] C. Mus, "10+ Akun Instagram Dengan Followers Terbanyak Di Indonesia," musdeoranje.net, 2015. http://www.musdeoranje.net/2016/08/akun-instagram-dengan-followers-terbanyak-di-indonesia.html (accessed Oct. 13, 2021).

[9] Allstars, "10 Artis Followers Terbanyak di Indonesia pada Instagram di 2021," Allstars.id, 2021. https://www.allstars.id/blog/2021/09/23/artis-indonesia-dengan-followers-instagram-terbanyak/ (accessed Oct. 27, 2021).

[10] R. Wongso, F. A. Luwinda, B. C. Trisnajaya, O. Rusli, and Rudy, "News Article Text Classification in Indonesian Language," Procedia Comput. Sci., vol. 116, pp. 137–143, 2017, doi: 10.1016/j.procs.2017.10.039.

[11] R. Ghanem and H. Erbay, "Context-dependent model for spam detection on social networks," SN Appl. Sci., vol. 2, no. 9, pp. 1–8, 2020, doi: 10.1007/s42452-020-03374-x.

[12] M. Li, B. Wu, and Y. Wang, "Comment Spam Detection via Effective Features Combination," 2019. doi: 10.1109/ICC.2019.8761340.

[13] X. Ban, C. Chen, S. Liu, Y. Wang, and J. Zhang, "Deep-learnt features for Twitter spam detection," 2018 Int. Symp. Secur. Priv. Soc. Networks Big Data, Soc. 2018, pp. 22–26, 2018, doi: 10.1109/SocialSec.2018.8760377.

[14] G. Jain, M. Sharma, and B. Agarwal, "Spam detection in social media using convolutional and long short term memory neural network," Ann. Math. Artif. Intell., vol. 85, no. 1, pp. 21–44, 2019, doi: 10.1007/s10472-018-9612-z.

[15] 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," Indones. J. Electr. Eng. Comput. Sci., vol. 14, no. 3, pp. 1508–1517, 2019, doi: 10.11591/ijeecs.v14.i3.pp1508-1517.

[16] S. Aiyar and N. P. Shetty, "N-Gram Assisted Youtube Spam Comment Detection," Procedia Comput. Sci., vol. 132, pp. 174–182, 2018, doi: 10.1016/j.procs.2018.05.181.

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

[18] R. Abinaya, E. Bertilla Niveda, and P. Naveen, "Spam detection on social media platforms," 2020 7th Int. Conf. Smart Struct. Syst. ICSSS 2020, pp. 31–33, 2020, doi: 10.1109/ICSSS49621.2020.9201948.

[19] IMDB, "IMDb Datasets," IMDb Datasets, 2022. https://www.imdb.com/interfaces/ (accessed Sep. 30, 2022).

[20] J. Jang, Y. Kim, K. Choi, and S. Suh, "Sequential Targeting: an incremental learning approach for data imbalance in text classification," 2020, [Online]. Available: http://arxiv.org/abs/2011.10216.

[21] V. Narayanan, I. Arora, and a Bhatia, "Fast and accurate sentiment classification using an enhanced Naive Bayes model," Int. Data Eng. Autom. Learn. Lect. Notes Comput. Sci., vol. 8206, pp. 194–201, 2013, doi: 10.1007/978-3-642-41278-3_24.

[22] T. A. Almeida and H. JosÃ, "SMS Spam Collection Data Set," UCI Machine Learning Repository, 2012. https://archive.ics.uci.edu/ml/datasets/SMS+Spam+Collection (accessed Sep. 30, 2022).

[23] A. Akbik, T. Bergmann, D. Blythe, K. Rasul, S. Schweter, and R. Vollgraf, "FLAIR: An easy-to-use framework for state-of-the-art NLP," NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Demonstr. Sess., pp. 54–59, 2019.

[24] L. Zhang and D. Moldovan, "Classification of semantic relations between pairs of nominals using transfer learning," Proc. 32nd Int. Florida Artif. Intell. Res. Soc. Conf. FLAIRS 2019, pp. 92–97, 2019.

[25] NSCLab, "Twitter Spam," 2014. http://nsclab.org/nsclab/resources/ (accessed Sep. 30, 2022).

[26] C. Chen, J. Zhang, X. Chen, Y. Xiang, and W. Zhou, "6 million spam tweets: A large ground truth for timely Twitter spam detection," IEEE Int. Conf. Commun., vol. 2015-September, pp. 7065–7070, Sep. 2015, doi: 10.1109/ICC.2015.7249453.

[27] D. Kang et al., "A Dataset of Peer Reviews (PeerRead): Collection, Insights and NLP Applications," in Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), 2018, vol. 1, pp. 1647–1661. doi: 10.18653/v1/N18-1149.

[28] Q. Lhoest et al., "Datasets: A Community Library for Natural Language Processing," in Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2021, pp. 175–184. doi: 10.18653/v1/2021.emnlp-demo.21.

[29] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," in Proceedings of the 28th International Conference on Computational Linguistics, 2020, pp. 757–770. doi: 10.18653/v1/2020.coling-main.66.

[30] A. A. Septiandri and O. Wibisono, "Detecting spam comments on Indonesia's Instagram posts," J. Phys. Conf. Ser., vol. 801, no. 012069, pp. 1–7, 2017, doi: 10.1088/1742-6596/755/1/011001.

[31] W. Zhang and H.-M. Sun, "Instagram Spam Detection," in 2017 IEEE 22nd Pacific Rim International Symposium on Dependable Computing (PRDC), Jan. 2017, pp. 227–228. doi: 10.1109/PRDC.2017.43.

[32] 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, 2017, doi: 10.1016/j.eswa.2017.04.003.

[33] B. Priyoko and A. Yaqin, "Implementation of naive bayes algorithm for spam comments classification on Instagram," in 2019 International Conference on Information and Communications Technology, ICOIACT 2019, 2019, pp. 508–513. doi: 10.1109/ICOIACT46704.2019.8938575.

[34] N. A. Haqimi, N. Rokhman, and S. Priyanta, "Detection Of Spam Comments On Instagram Using Complementary Naïve Bayes," IJCCS (Indonesian J. Comput. Cybern. Syst., vol. 13, no. 3, p. 263, Jul. 2019, doi: 10.22146/ijccs.47046.

[35] R. Yang, J. Zhang, X. Gao, F. Ji, and H. Chen, "Simple and Effective Text Matching with Richer Alignment Features," in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 4699–4709. doi: 10.18653/v1/P19-1465.