Clickbait Detection in Indonesian News Title with Gray Unbalanced Class Based on BERT

Pulung Nurtantio Andono, Pieter Santoso Hadi, Muljono*, and Catur Supriyanto

Informatics Engineering, Dian Nuswantoro University, Semarang, Indonesia;
Email: pulung.nurtantio.andono@dsn.dinus.ac.id (P.N.A.), p31201902285@mhs.dinus.ac.id (P.S.H.),
catur.supriyanto@dsn.dinus.ac.id (C.S.)
*Correspondence: muljono@dsn.dinus.ac.id (M.)

***Abstract*—Bahasa Indonesia is used by about 263 million people in the world but it is classified as an under-resourced language. The problem of clickbait in news analysis has gained attention in recent years. However, for Indonesian, there is still a lack of resources for clickbait tasks. Clickbait attracts the attention of readers, even though the content is not informative and misleading. The imbalance of the clickbait dataset means unequal distribution of classes within the dataset which affects the classification result. In this research, focal loss is proposed to improve classification accuracy without reducing the number of original data. Normally, clickbait data are separated into two classes, namely clickbait, and non-clickbait. However, some titles are difficult to categorize, even by humans. Therefore, this study categorizes the titles into three categories, namely clickbait, non-clickbait, and gray-clickbait. The proposed method achieves an accuracy of 93.4% in the classification of two classes, which is better than previous studies. However, the proposed method achieves an accuracy of 73.3% in the classification of three classes. Our research shows a high similarity between gray-clickbait and clickbait data, making classification more challenging. On the other hand, the use of titles on three categorizations in clickbait is not enough to provide better classification performance.**

***Keywords*—classification, imbalanced data, BERT, focal loss, clickbait, Indonesian**

I. Introduction

In this era of the internet, everyone can easily access news online. The large population, especially the Indonesian, is the target of content publishers to generate advertising revenue using clickbait news titles. Indonesia itself is one of the ten most internet user countries in the world [1–3]. Bahasa Indonesia is used by about 263 million people in the world but it is classified as an under-resourced language. Bahasa Indonesia is still categorized as under-resourced language because the development itself in electronic media is still rare [4].

Clickbait news offers a stunning impression and usually uses hyperbole or exaggerated title. Clickbait headlines have an emphasis on intent. Research states that accessing news with clickbait titles gives readers disappointment because the content is different from the title [5]. However, clickbait news content is not a hoax [6]. Clickbait titles can cause misinformation for internet users who only read the title without reading the news content. Some problems can occur when misinformation turns into a hoax [7].

In a previous study, William and Sari created an Indonesian clickbait dataset containing a collection of news titles that have been classified as clickbait and non-clickbait [8]. Our research uses the dataset created by them. William and Sari used CNN and Bi-LSTM methods in their research. CNN and Bi-LSTM methods provide a relatively low accuracy, which is only about 70% [8]. This happens because several factors, such as Indonesian data structures have multiple meanings, such as "bias" (capable) or "bisa" (poison). Words with multiple meanings affect the classification results [9]. The BERT method can answer this problem. The BERT method uses words and sentences to distinguish the context of words in a sentence. Based on [9], BERT is better than Bi-LSTM, especially on large datasets.

The imbalanced dataset is also the cause of the low accuracy. Fakhruzzaman *et al.* solved the problem of data imbalance by using the under-sampling method, which eliminated a large number of non-clickbait datasets (datasets with a larger number) [9]. It is not appropriate to remove a large number of non-clickbait datasets because it can eliminate important information [10]. The focal loss method was proposed in this research to overcome the problem of dataset imbalance so as not to eliminate a large number of non-clickbait datasets.

There are also data on clickbait and non-clickbait that are difficult to define by humans, so in this study, we separate the categories into three classes, namely clickbait, non-clickbait, and gray-clickbait. Gray-clickbait itself is data that did not receive complete approval at the time of data validation. In the dataset created by William and Sari [8], gray-clickbait can be found in the Doesn't Agree category. These three classes or labels can show more profound and accurate classification results than previous studies, especially in [8, 9]. As mentioned before, clickbait is a form of advertising. However, clickbait isn't always false or harmful. Clickbait isn't necessarily bad and many of them offers useful

Manuscript received July 1, 2022; revised August 21, 2022; accepted October 7, 2022; published March 17, 2023.

information [6]. In this research, we call this kind of clickbait as gray-clickbait. This means gray-clickbait has strong background for real-world use cases. Gray-clickbait will open a new fresh path in clickbait detection and information filtering.

The objective of this research is to test the effectiveness of the proposed method in clickbait detection, which can be compared with the methods of previous research and to develop better understanding of the use of three classes in clickbait detection, which is based on the author's knowledge is the first time this has been done. In the following parts of the paper, we introduce related work, summarize the theory, and provide an overview of the methods for clickbait detection in Section II. Toward Section III, we describe the proposed method. We describe the setups and results of the case study in Section IV, while Section V discusses our summaries and conclusions.

II. Related Works

Bahasa Indonesia, a standardized version of Malay, is the sixth most widely spoken language in the world. The spoken system similar to Malay and the written system is refer to Roman alphabet system. However, it is classified into an under-resourced language. Under-resourced language lacks of a unique writing system or stable orthography, limited presence on the web, lack of linguistic expertise, or lack of electronic resources for speech and language. In this case, Bahasa Indonesia is still categorized as under-resourced language because the development itself in electronic media is still rare [4].

Nowadays, clickbait is used by various media to attract internet users [11]. The use of clickbait has proven to increase ad revenue results. Clickbait news offers a stunning impression and usually has a hyperbole or exaggerated title [12]. Signs like "!", "?", informal words which are hyperbole often appear in clickbait writing. Unfortunately, clickbait creates information problems that are detrimental to internet users. Clickbait titles gives readers disappointment because the content is different from the title. Moreover, clickbait titles can cause misinformation and turns into a hoax [13–16].

Based on previous studies, machine learning is a key for solving the problem, namely by detecting clickbait or non-clickbait [8, 9]. However, some data are difficult to categorize, even by human. William and Sari tries to cope with this problem by removing large numbers of data using validation process [8]. However, Fakhruzzaman *et al.* noted this problem but failed to give satisfactory answer. The reason is the data itself has different features than clickbait or non-clickbait [9]. Hence, the use of three classes, gray-clickbait, is needed in classification [9, 17].

LSTM and Bi-LSTM are capable of learning order dependence in sequence prediction problems. However, Bahasa Indonesia have double meaning in its structure. LSTM and Bi-LSTM can't resolve this kind of problem. Devlin *et al.* shows BERT (Bidirectional Encoder Representations from Transformers) as modelling language and BERT has proven to outperform the other languages modelling [18]. In previous research by Fakhruzzaman *et al.* [9], BERT was proven to overcome the problem of word ambiguity in Indonesian. BERT works by understanding the context of a sentence by using a masked language model and Next Sentence Prediction [18]. BERT randomly takes words in a sentence and changes them to [mask] in the masked language model. In Next Sentence Prediction, BERT takes two sentences and determines the relationship between these two sentences. These works allow BERT to understand the context of a sentence. In this case, BERT works as feature extraction.

The imbalanced dataset is generally solved in two ways, namely cost-sensitive learning and re-sampling methods. In the previous research of [9] solved the problem of data imbalance by using the under-sampling method. However, this method can eliminate large amounts of data which can cause the data to be monotonous. Therefore, cost-sensitive learning is proposed to answer these problems. Cost-sensitive learning works by applying a loss function with different weights for each class. Focal loss as a loss function is proposed to deal with imbalanced class problems [10]. The loss function works by evaluating how well the model is used. If the prediction results deviate far, the loss function will have a high value, while if the prediction is suitable or appropriate, the loss function will have a low value. Therefore, we adopted the concept and the stages for our experiment to increase accuracy.

Generally, the results of pre-trained model are excellent. The research of [9] use BERT with this approach [9]. However, fine-tuning BERT can be done to find better accuracy results in the classification. Fine-tuning BERT is done with backward propagation to find a more suitable weight so that the results of the neural network classification become better [18].

SVM, Random Forest, and Decision Tree has been used as baseline model for many classifications task. SVM, Random Forest, and Decision Tree have been proven to give good accuracy result in classification. Neural Network is a method used in deep learning.

Anand, Chakraborty, and Park [19] has compared the use of neural networks with baseline methods such as SVM, Random Forest, and Decision Tree, where the use of neural networks has been shown to outperform the results of the baseline method. Neural networks consist of neurons. Connections between these neurons are called weights, and biases are connected to each neuron. Neural Network distinguishes between input, hidden, and output layers. Moving forward through the network is called a forward pass. Neural Network iteratively uses a formula to calculate each neuron in the next layer until Neural Network gets an output. Given the result of output, Neural Network goes back and adjusts the weights and biases to optimize the loss function. This is called a backward pass. Backward pass essentially tries to adjust the whole neural network to optimize the output value. In a sense, Neural Network keeps optimizing the output result by running through new observations from the dataset [17–19].

III. METHOD

A. Scheme

In this section, we will explain the methodology and the process steps that will be done in the research. The purpose of our research is to study the effectiveness of the proposed method for clickbait detection of Indonesia online news headlines. The proposed method is compared with previous research to see the effectiveness of result. The scheme is shown in Fig. 1. BERT is used as word embeddings or feature extraction of the dataset. We use cross-fold validation, where the value of k is five. We use Neural Network with focal loss for classification. Focal loss is used as weighting scheme to cope with the problem of imbalance classes. For better accuracy, the model is fine-tuned by changing the hyperparameters on the neural network. Confusion Matrix is used for evaluation. We explain them with more details in next points.

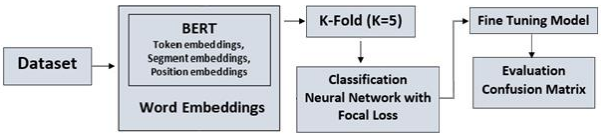

Figure 1. Research method flow.

B. Dataset

This study uses a public dataset in the form of a collection of online news titles, where this dataset is obtained from Mendeley Data [8]. Fig. 2 Shows the framework figure of [8]. The dataset consists of news titles from 12 online local news, namely detikNews, Fimela, Kapanlagi, Kompas, Liputan6, Okezone, Posmetro Medan, Republika, Sindonews, Tempo, Tribunnews, and Wowkeren. The data validation was carried out by three annotators or examiners fluent in Indonesian. The test is carried out based on the title of the news article. The majority of the three examiners' opinions are then considered groundtruth. The study results created three types of datasets, namely the dataset with the main code, the dataset with the All Agree (AG) code, and the dataset with the Doesn't Agree code.

The dataset with the code All Agree (AG) contains a dataset that all three examiners approved in the validation process. The code Doesn't Agree dataset is a collection of data wherein the validation process has difficulty determining the data as clickbait or non-clickbait, namely, where one examiner classifies the data as clickbait and the other two examiners classify the data as non-clickbait. The dataset with the main code contains a combination of the All Agree and Doesn't Agree codes.

Data with code Main and All Agree is used for classification (Bi-LSTM & CNN). Confusion Matrix is used to measure the classification results. The research of only used Accuracy as the value of measurement.

In this study, three experiments were conducted with different data. The first experiment uses the code AG (All-Agree) and the Main dataset. In this experiment, the research's method of [8] was used to compare the proposed method, BERT with focal loss. This experiment has two classes, namely clickbait and non-clickbait.

The second experiment uses the AG dataset with two classes, namely clickbait and non-clickbait. Fakhruzzaman *et al.* used as a comparison or comparison in this Experiment [9]. The third experiment uses the AG dataset and the Doesn't Agree dataset, which was coded Gray in this study. In this experiment, classes are broken down into clickbait, non-clickbait, and gray-clickbait. The gray-clickbait data is obtained from the Doesn't Agree code.

The total amount of data used in this study is data with clickbait classes amounting to 3,316, data with non-clickbait classes amounting to 5,297, and data with gray-clickbait classes amounting to 6,387. The total data used in this study is 15,000 data.

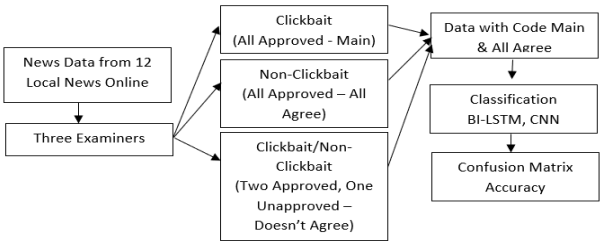

Figure 2. Framework figure of [8].

C. K-Fold Validation (K=5)

K-Fold is used to validate with a K value of 5. The K value of 5 is used because research [8, 9] uses K-Fold with a K value of 5. This means that 80% of the dataset becomes training data and 20% becomes test data. K-Fold (*K*=5) validation works with, for example, 100 data, then data 1 to 20 will be test data and data 21 to 100 become training data, then iterate over with data 21 until data 40 become test data. The rest becomes training data, then data 41 to 60 becomes test data and the rest is training data. This continues until data 81 to 100 become test data and the rest is training data.

D. Embedding Layer

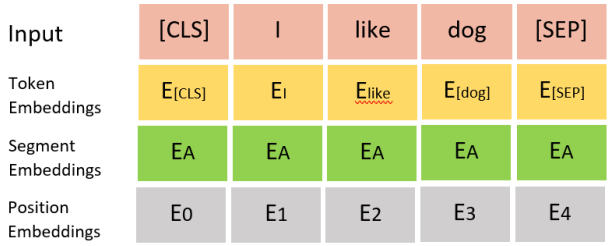

Figure 3. Embeddings layer.

Fig. 3 describes the work of BERT in layer embedding, especially Token Embeddings, Segment Embeddings, and Position Embeddings. The Bidirectional Encoder Representations from Transformers (BERT) used in this study has 12-layer, 768-hidden, 12-heads, and 110M

parameters. BERT works with Python ≥ 3.5 and TensorFlow ≥ 1.10. Generally, BERT works using the GPU for a faster work process.

Token embeddings work by splitting words in a sentence. In this study, BERT uses the light/bert-base-indonesian-522M model for Indonesian. For example, in the sentence "I like strawberries" then the embeddings token will appear as "[CLS]", "I", "like", "Stra", "##w", "##berry", "[SEP]". Extra tokens are added at the beginning of the sentence, namely "[CLS]" and at the end of the sentence, namely "[SEP]." This is done as an input representation in classifying. Tokenization is based on the WordPiece tokenization method and the existing word dictionary. In this research, this dictionary is based on the hugging-face library dictionary.

Segment embeddings work to find out token ownership by a sentence. For example, in two sentences, namely (1) "I like dog" and (2) "I like cat", then after going through the token embedding process, the results will appear as '"[CLS] (0)", "I (0)", "like (0)", "dog (0)", "[SEP] (0)", "[CLS] (1)", "I (1)", "like (1)", "cat (1)"," [SEP] (1)"'. These results display two codes, namely code 0 and code 1. This shows that each token has an ownership code from which the token sentence originated. Because of these embedding segments, the token in BERT has information that the token comes from sentence (1) or sentence (2).

Position embeddings show the order of tokens in a sentence. For example, in the sentence of Bahasa Indonesia, "Saya suka anjing saya" which translated to English means I like my dog, the position embeddings will appear as "[CLS]," "Saya (1)", "suka (2)", "anjing (3)", "saya (4)", "[SEP]". These results show that the word "saya" at the beginning of the sentence is different from "saya" at the end of the sentence. This position of tokens show how BERT can understand the context of words in a sentence. The results of the token embeddings, segment embeddings, and position embeddings will be in the form of numbers. This result will become the feature used in calculating neural network.

E. Neural Network with Focal Loss

The neural network method is used in the classification process, namely the feedforward network. In a feedforward network, information moves in one direction. This feedforward network is an integral part of the BERT method. In this classification method, the focal loss is used to overcome the class imbalance problem. Focal Loss (FL) is an improved version of Cross-Entropy Loss (CE) that tries to handle the class imbalance problem by assigning more weights to hard or easily misclassified examples and to down-weight easy examples. Focal Loss reduces the loss contribution from easy examples and increases the importance of correcting misclassified examples.

Focal loss is an extension of the cross-entropy loss function that would down-weight easy examples and focus training on hard negatives. Focal loss for is written as follows:

$$FL(p_t) = -\alpha_t (1 - p_t)^r \log(p_t) \quad (1)$$

The focal loss itself is a development of the cross-entropy loss method. In this method, Adam Optimizer is also used. In addition, to get better accuracy results, the BERT is fine-tuned by applying backward propagation. The results are evaluated with the Confusion Matrix.

F. Fine Tuning Model

The fine-tuning Model is done to produce better accuracy results. The fine-tuning Model is done by changing the hyperparameters on the neural network.

TABLE I. HYPERPARAMETERS SETTINGS

Hyperparameter	Value
Batch size	16
Epoch	4
Optimizer	Adam Optimizer
Drop Out	0.1
Learning rate	3e-4, 1e-4, 5e-5, 3e-5

Table I shows hyperparameters carried out in this study. The fine-tuning hyperparameters are batch sizes with a value of 16, the learning rate on the Adam optimizer, namely 3e-4, 1e-4, 5e-5, 3e-5 where e-5 is worth 10-5, and dropout value is 0.1. The use of these hyperparameters is in accordance with research [18] as the primary reference for using the BERT method.

G. Evaluation

Confusion Matrix is commonly used to measure text classification results [1]. Table II shows the form of the Confusion Matrix.

TABLE II. CONFUSION MATRIX

	Predicted: Yes	Predicted: No
Actual: Yes	True Positive	False Negative
Actual: No	False Positive	False Positive

Accuracy is used to evaluate the overall effectiveness of the classification algorithm. In this classification process, several types of tests are carried out. The test was carried out by comparing the proposed method with the methods of previous studies, namely research methods [8, 9].

At this testing stage, there is a comparative test of focal loss and cross-entropy loss and a comparison of the Bi-LSTM, BERT (standard), and BERT and Focal Loss methods proposed in this study. The methods being compared are the methods in previous studies. These tests will be used to answer the formulation of the problem in this study.

IV. RESULT

The research was conducted using Python 3.6.9. The GPU is used to help speed up the calculation process in classifying. The hyperparameters applied in this research are learning rate of 3e-5, dropout rate of 0.1, and batch size of 16. Using a learning rate of 3e-5 is the most

optimal learning rate compared to the learning rate of 3e-4, 1e-4, and 5e-5. The experiment was carried out using K-fold validation where K is worth 5. The value of K is five following the research used as a comparison [8, 9].

William and Sari are the original creator of the dataset [8]. In this experiment, the results of the method proposed in this study (BERT with focal loss) were compared with the research method used by William and Sari (Bi-LSTM and CNN). This study uses a public dataset created by William and Sari. The data used in this experiment is data with code All-Agree and Main. The Main totals are 15,000, with 6,290 clickbait and 8,710 labeled non-clickbait. Meanwhile, the All-Agree totals are 8,613, with 3,316 labeled clickbait and 5297 labeled non-clickbait. Fig. 4 shows the label distribution of dataset Main and All-Agree as an imbalanced dataset.

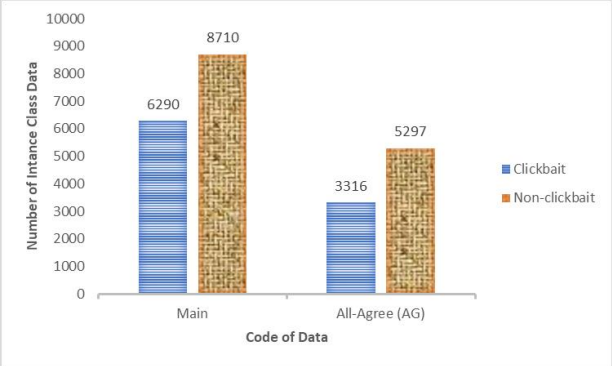

Figure 4. Label distribution of dataset main and all-agree.

However, William and Saridid not change the nature of the imbalance in the dataset, such as using under-sampling or over-sampling methods, before classifying using CNN and Bi-LSTM [8]. On the other hand, William and Sari used symbol such as '?', '!', '.', etc. as comparison. William and Sari only use accuracy as the value of measurement. Table III shows the results of the comparison of the proposed methods (BERT and focal loss with fine tuning) with research [8]. Table III shows the use of BERT and focal loss with fine tuning outperforms CNN and Bi-LSTM methods.

TABLE III. COMPARISON RESULTS OF ACCURACY ON MAIN AND AG

Dataset	CNN [8]	Bi-LSTM [8]	BERT and FL with Fine Tuning
Main with symbols	76.3%	76.9%	81.4%
Main without symbols	75.7%	77.7%	79.6%
AG with symbols	85.7%	88.3%	93.4%
AG without symbols	87.8%	86.7%	90.1%

Based on Table III, it can be concluded four things, namely:

- The proposed method has been successfully outperformed the original paper or creator of dataset.
- There is a decrease in accuracy when the dataset uses symbols and does not use symbols. This is in accordance with research [18] which states that BERT uses symbols to get the context of a sentence so that there is a decrease in accuracy when not using symbols.
- There is a decrease in accuracy in the Main dataset compared to when using the All-Agree dataset. This happens because the Main dataset has a large amount of data, namely 6,387 data, which has the characteristics of two clickbait and non-clickbait labels at once. Research [8] revealed that 6,387 data itself is difficult to distinguish by humans [12]. Clickbait news titles use hyperbole and stunning impressions to attract readers' interest. The characteristics of clickbait news headlines are the symbols '!', '?' and the use of informal words in writing news headlines. However, the 6,387 data has these characteristics but is categorized as non-clickbait by two annotators (examiners) in research [8].
- Research [12] reveals the use of symbols '!', '?' and informal words as a unique feature of clickbait news headlines. This is true and appropriate for the All-Agree data code but not for the Main data code. This is because the Main data code is a mixture of the All-Agree and Doesn't Agree data codes. Our research points out the use of the Main code dataset for classification of clickbait or non-clickbait news titles is not correct. This means the Main code dataset is bad data.

As mentioned before, Fakhruzzaman *et al*., noted the problem of bad data but failed to give satisfactory answer [9]. This research chooses to not use the Main code dataset in his experiment. Fakhruzzaman *et al*. uses the All-Agree data code in his research. Fakhruzzaman *et al*. also uses undersampling to overcome the problem of dataset imbalance. As a comparison, Fakhruzzaman *et al.* [9] uses cross-entropy loss, and this study uses focal loss.

In this experiment, the dataset used for the proposed method is not transformed using the method of undersampling. This is in accordance with the formulation of the problem where the undersampling can eliminate important information and produce monotonous data. Focal loss is used to overcome the problem of unbalanced data. Fig. 5 shows label Distribution of All-Agree (Normal) and All-Agree (Under-sampling).

Figure 5. Label distribution of AG (normal) and (under-sampling).

Table IV shows a slight increase in the results. The results of the fine-tuning BERT method with focal loss show that under-sampling is not required for classification. The results show that the proposed method outperforms the previous method of [9].

TABLE IV. COMPARISON RESULTS OF CLASSIFICATION ON AG DATASET

	BERT with Undersampling [9]	Proposed method
Accuracy	91.5%	93.4%
Precision	91.6%	92%
Recall	91.4%	91.0%
F1	91.4%	91.6%

The proposed method has shown to outperforms any previous studies. However, the problem of vague class, data which hard to distinguish even by human, in the Main dataset code still exists. In this research, we propose the use of three classes. Based on the author's knowledge, the use of three classes, namely clickbait, non-clickbait, and gray-clickbait, for classification is the first to be conducted. The total dataset used in this experiment is 15,000 data. In this experiment, the data was arranged into data with clickbait labels amounting to 3,316, data with non-clickbait labels amounting to 5,297, and data with gray-clickbait labels amounting to 6,387.

The data with the gray-clickbait label comes from the Doesn't Agree data code. Fig. 5 shows an unbalanced data distribution between clickbait, non-clickbait, and gray-clickbait. However, our experiment shows that the use of focal loss can overcome the problem of unbalanced data. Based on research [12], clickbait generally has longer sentence than non-clickbait. However, in this study, it can be seen that gray-clickbait has a sentence length above non-clickbait and below clickbait. According to research [12], clickbait news titles use hyperbole and stunning impressions to attract readers' interest. The characteristics of clickbait news headlines are the symbols '!', '?' and the use of informal words in writing news headlines. However, the examples show that the gray-clickbait category also has these characteristics.

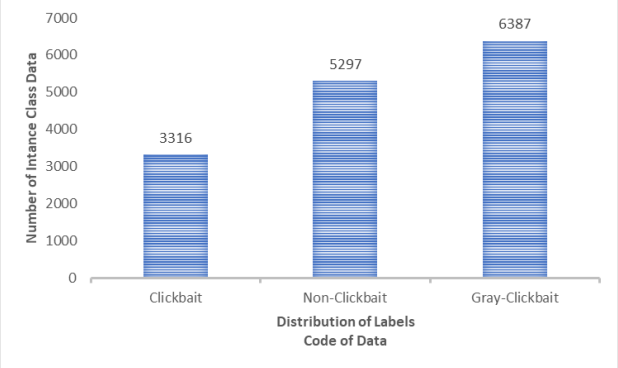

Figure 6. Label distribution of clickbait, non-clickbait, gray-clickbait.

Fig. 7 shows that gray-clickbait has characteristics of clickbait and non-clickbait. Some examples of such data from [8] dataset are as follows:

- Sri Mulyani is sad to see the building made with the people's money has been destroyed by riotous mobs.
- Members of DPRD Banten Pawn his SK, Commission II DPR: Closing Campaign Debt?
- Heartbreaking! 15 Family Killed in Saudi Arabia Attack in Yemen.

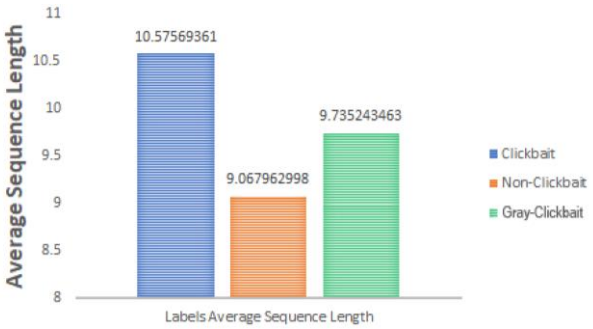

Figure 7. Labels average sequence length.

Table V shows the classification results using three labels: clickbait, non-clickbait, and gray-clickbait. There is a decrease in results compared to this experiment. Table V shows the classification results using the initial three classes of data before the removal using text similarity. Table VI shows the Confusion Matrix of classification in Table V. Based on Table VI, we can see 27% (286) of non-clickbait data are predicted as gray-clickbait and 24% (166) of clickbait data are predicted as gray-clickbait. Table VI shows the difficulty in classifying gray-clickbait.

TABLE V. CLASSIFICATION RESULTS OF THREE LABELS (BEFORE USING TEXT-SIMILARITY METHOD)

Label	Precision	Recall	F1-Score
Non-Clickbait	65.0%	74.6%	69.4%
Clickbait	72.0%	71.0%	71.4%
Gray-Clickbait	61.2%	54.2%	57.6%
Accuracy: 65.08%			

TABLE VI. CONFUSION MATRIX OF CLASSIFICATION (BEFORE USING TEXT-SIMILARITY METHOD)

	Predicted: Non-Clickbait	Predicted: Clickbait	Predicted: Gray-Clickbait
Actual: Non-Clickbait	749	9	286
	Predicted: Non-Clickbait	Predicted: Clickbait	Predicted: Gray-Clickbait
Actual: Clickbait	15	510	166
Actual: Gray-Clickbait	394	194	677

Gray-clickbait was obtained from the [8] dataset, previously coded as Doesn't Agree. The Doesn't Agree code has two labels, clickbait, and non-clickbait, which were not fully approved by the three annotators (examiners). The majority of answers of the two annotators were used to determine the label. This means that data that is more directed towards non-clickbait needs to be removed.

In this research, the removal of these data is done using the text-similarity method. The text-similarity method is used in order to find topic efficiently. Text-similarity method is used to improve the accuracy of the classifiers, especially by combining Cosine Similarity with Classifier [20]. Using the BERT and Cosine-Similarity methods, the gray-clickbait data are compared with the clickbait data from the All-Agree code dataset. A threshold of 0.5 is used so that when a data has a similarity value above 0.5, it will be labelled as gray-clickbait. In contrast, if the data has a similarity value below 0.5, then the data is gray non-clickbait. Gray non-clickbait data has a similar value to non-clickbait data, where the purpose of this study is to detect gray-clickbait so the data will not be used because the data is non-clickbait data. The formula of Cosine Similarity is written as follow:

$$\cos(\theta) = \frac{\sum_{i=1}^{n} AiBi}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \quad (2)$$

Fig. 8 shows the distribution of dataset labels after the text-similarity method is used to test the similarity value of the gray-clickbait data. In Fig. 8, there is a decrease in the number of gray-clickbait data from 6.387 to 2.861. Thus, the total dataset used is 11,474 with 3,316 labeled clickbait, 5,297 labeled non-clickbait, and 2,861 labeled gray-clickbait. The classification results of 11,474 data using BERT with focal loss are shown in Table VII. Table VIII shows the Confusion Matrix of classification in Table VII.

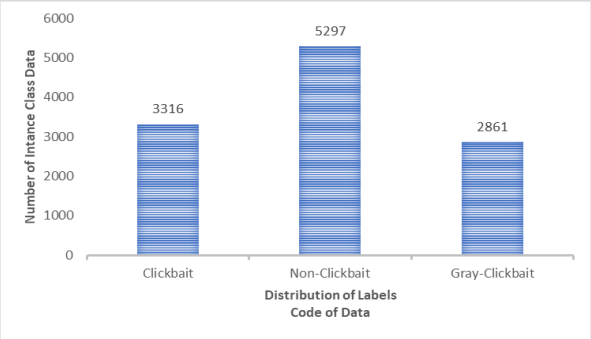

Figure 8. Label distribution after text-similarity.

TABLE VII. CLASSIFICATION RESULTS OF THREE LABELS (AFTER USING TEXT-SIMILARITY METHOD)

Label	Precision	Recall	F1-Score
Non-Clickbait	80.6%	86.8%	83.6%
Clickbait	75.4%	80.8%	77.8%
Gray-Clickbait	52.0%	39.8%	44.8%
Accuracy: 73.3%			

There was an increase in the accuracy of 73.3% from the original 65.08% (Table V). Comparison with Table VI shows an increase in precision, recall, and F1 scores except for the gray-clickbait label. In the precision section, it can be seen that the precision of non-clickbait labels is greater than the precision of clickbait labels. In Table VII, the precision of clickbait labels is greater than the precision of non-clickbait labels. This is influenced by the gray-clickbait label, which currently has similar value to the clickbait label. This similarity affects the precision of the gray-clickbait label, which showed a precision value of 52.0%. Based on Table VIII, we can see 12% (132) of non-clickbait data are predicted as gray-clickbait and 21% (146) of clickbait data are predicted as gray-clickbait. We can see the reduction of non-clickbait data that are predicted as gray-clickbait, from 27% into 12%. However, the precision, recall, and F1 score of gray-clickbait is reduced as the result of that. This is because data that is more directed towards non-clickbait is removed. We can see the precision, recall, and F1 score of clickbait and non-clickbait are increased and not reduced. Table VIII shows the difficulty in classifying gray-clickbait due to the more similarity of gray-clickbait data with clickbait.

TABLE VIII. CONFUSION MATRIX OF CLASSIFICATION (AFTER USING TEXT-SIMILARITY METHOD)

	Predicted: Non-Clickbait	Predicted: Clickbait	Predicted: Gray-Clickbait
Actual: Non-Clickbait	922	12	132
Actual: Clickbait	25	520	146
Actual: Gray-Clickbait	163	113	262

V. CONCLUSION

Our experiments have answered the problem formulation in this study. Based on the research that has been done, it can be concluded several things. First, the proposed method is better than the Bi-LSTM and CNN in classifying clickbait and non-clickbait. The BERT method uses symbols to get the context of a sentence so that there is a decrease in accuracy when not using symbols. This shows the need to use symbols when filtering the pre-processing text.

The use of the Click-ID dataset created by William and Sari with the All-Agree code shows a much better classification result than the Click-ID dataset with the Main code [8]. This is because the dataset with the Main code is a mixture of the dataset with the All-Agree code and the dataset with the Doesn't Agree code, where the data in the Doesn't Agree code has clickbait and non-clickbait characteristics so that the use of the Main dataset for the classification of clickbait news titles is not correct.

The result of the proposed method also shows that the use of under sampling method is not required in conducting the classification. The use of focal loss is

enough to cope with problem of data imbalance. Based on research of Hadi *et al.*, clickbait generally has longer sentence than non-clickbait [12]. However, in this study, it can be seen that gray-clickbait has a sentence length above non-clickbait and below clickbait. In addition, clickbait news titles have a hyperbole and stunning impression to attract readers' interest. The characteristics of clickbait news headlines are the symbols '!', '?' and the use of informal words in writing news headlines. However, research shows that the gray-clickbait category also has these characteristics. The use of three classes, namely clickbait, non-clickbait, and gray-clickbait, is important because there is data with clickbait and non-clickbait properties, which will affect the classification. Classification with three labels, namely clickbait, non-clickbait, and gray-clickbait, showed a decrease in results compared to classification with two labels, clickbait and non-clickbait.

As shown in Table VII, there is difficulty in classifying gray-clickbait due to the similarity of gray-clickbait data with clickbait. This points out the use of title alone is not enough in the future of clickbait classification. Table VII shows the results of the gray-clickbait precision of 52%, which shows the difficulty in classifying the data as gray-clickbait. Based on the author's knowledge, this research using the gray-clickbait label in the clickbait and non-clickbait classifications is the first to be conducted. In real-world use case, gray-clickbait can become optional choice for internet users to be shown or not to be shown. Clickbait detection and gray-clickbait detection are very different in real-world use case. Since, gray-clickbait still has useful information for its reader but clickbait doesn't have any useful information.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Conceptualization of this research was done by Pulung Nurtantio Andono and Muljono. The methodology of this research was done by Pulung Nurtantio Andono, Pieter Santoso Hadi and Muljono. Software was prepared by Pieter Santoso Hadi and Muljono. Validation was done by Pulung Nurtantio Andono and Catur Supriyanto. Formal analysis was analyzed by Pieter Santoso Hadi and Muljono. Resources was prepared by Pulung Nurtantio Andono. Investigation was worked by Pulung Nurtantio Andono and Muljono. Data curation was worked by Pieter Santoso Hadi and Catur Supriyanto. The writing of the original draft preparation was done by Pulung Nurtantio Andono. The review and editing was done by Muljono, Pieter Santoso Hadi, and Catur Supriyanto. The visualization was done by Pulung Nurtantio Andono and Pieter Santoso Hadi. Supervision was managed by Muljono and Catur Supriyanto. The project administration was managed by Muljono. The funding acquisition was prepared by Pulung Nurtantio Andono. All authors read and approved the final manuscript.

REFERENCES

[1] A. W. Haryanto, E. K. Mawardi, and Muljono, "Influence of word normalization and chi-squared feature selection on Support Vector Machine (SVM) text classification," in *Proc. International Seminar on Application for Technology of Information and Communication: Creative Technology for Human Life*, 2018, pp. 229–233.

[2] P. K. Dimpas, R. V. Po, and M. J. Sabellano, "Filipino and English clickbait detection using a long short term memory recurrent neural network," in *Proc. International Conference on Asian Language Processing*, 2018, pp. 276–280.

[3] P. Klairith and S. Tanachutiwat, "Thai clickbait detection algorithms using natural language processing with machine learning techniques," in *Proc. 4th International Conference on Engineering, Applied Sciences and Technology: Exploring Innovative Solutions for Smart Society*, 2018, pp. 1–4.

[4] E. Cahyaningtyas and D. Arifianto, "Development of under-resourced Bahasa Indonesia speech corpus," in *Proc. 9th Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. APSIPA ASC 2017*, 2017, pp. 1097–1101.

[5] S. Chawda, A. Patil, A. Singh, and A. Save, "A novel approach for clickbait detection," in *Proc. Int. Conf. Trends Electron. Informatics*, 2019, pp. 1318–1321.

[6] D. Palau-Sampio, "Reference press metamorphosis in the digital context: Clickbait and tabloid strategies in Elpais.com," *Commun. Soc.*, vol. 29, no. 2, pp. 63–71, 2016.

[7] W. Zhang, W. Du, Y. Bian, C. H. Peng, and Q. Jiang, "Seeing is not always believing: An exploratory study of clickbait in WeChat," *Internet Res.*, vol. 30, no. 3, pp. 1043–1058, 2020.

[8] Y. William and A. Sari, "CLICK-ID: A novel dataset for indonesian clickbait headlines," *Data Br.*, 106231, 2020.

[9] M. N. Fakhruzzaman, S. Z. Jannah, R. A. Ningrum, and I. Fahmiyah. (2021). Clickbait headline detection in indonesian news sites using multilingual bidirectional encoder representations from transformers (M-BERT). [Online]. Available: http://arxiv.org/abs/2102.01497

[10] R. Iikura, M. Okada, and N. Mori, "Improving BERT with focal loss for paragraph segmentation of novels," in *Proc. 17th International Conference Distributed Computing and Artificial Intelligence*, 2021, p. 38.

[11] S. Pengnate, "Shocking secret you won't believe! Emotional arousal in clickbait headlines: An eye-tracking analysis," *Online Inf. Rev.*, vol. 43, no. 7, pp. 1136–1150, 2019.

[12] P. S. Hadi, Mujono, A. Z. Fanani, G. F. Shidik, Purwanto, and F. Alzami, "Using extra weight in machine learning algorithms for clickbait detection of indonesia online news headlines," in *Proc. International Seminar on Application for Technology of Information and Communication (iSemantic)*, 2021, pp. 37–41.

[13] A. Kuriakose, D. Sebastian, E. M. Mathew, H. Mathew, and G. E. Gokulnath, "ALIKAH—A clickbait and fake news detection system using natural language processing," in *Proc. Int. Conf. Trends Electron. Informatics*, 2019, pp. 1203–1206.

[14] F. Wei and U. T. Nguyen, "A neural attentive model using human semantic knowledge for clickbait detection," in *Proc. Int. Symp. Parallel Distrib. Process. with Appl. 2020 IEEE Int. Conf. Big Data Cloud Comput. 2020 IEEE Int. Symp. Soc. Comput. Netw*, 2020, pp. 770–776.

[15] O. Papadopoulou, M. Zampoglou, S. Papadopoulos, and I. Kompatsiaris. (2017). A two-level classification approach for detecting clickbait posts using text-based features. [Online]. Available: http://arxiv.org/abs/1710.08528

[16] H. Ahmed, I. Traore, and S. Saad, "Detection of online fake news using n-gram analysis and machine learning techniques," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 10618, pp. 127–138, 2017.

[17] M. N. Fakhruzzaman and S. W. Gunawan. (2021). Web-based application for detecting Indonesian clickbait headlines using IndoBERT. [Online]. Available: http://arxiv.org/abs/2102.10601

[18] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol*, 2019, vol. 1, no. Mlm, pp. 4171–4186.

[19] A. Anand, T. Chakraborty, and N. Park, "We used neural networks to detect clickbaits: You won't believe what happened next!" *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 10193, pp. 541–547, 2017.
[20] K. Park, J. S. Hong, and W. Kim, "A methodology combining cosine similarity with classifier for text classification," *Appl. Artif. Intell.*, vol. 34, no. 5, pp. 396–411, 2020.

Pulung Nurtantio Andono is currently an associate professor of informatics engineering. He received a master of computer science from Universitas Dian Nuswantoro, Semarang, Indonesia, and a doctoral degree in computer science from Institut Teknologi Sepuluh Nopember (ITS) Surabaya. He is currently a lecturer at Universitas Dian Nuswantoro, Semarang, Indonesia. His research interests include machine learning and computer vision.

Pieter Santoso Hadi is currently a master's student. His major is in text and data mining. He is a graduate student from the School of Computer Science at Dian Nuswantoro University, Semarang, Indonesia.

Muljono is with Informatics Engineering Department, Dian Nuswantoro University, Semarang, Indonesia. He obtained his doctoral degree from at Electrical Engineer-ing Department, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia. He also was joined by collaboration research at School of Media Science, Tokyo University of Technology Japan. His research interests are artificial intelligence and natural language processing

Catur Supriyanto received a master of computer science from Universiti Teknikal Malaysia Melaka, Malaysia, in 2011 and a doctoral degree in Computer Science from Universitas Gajah Mada, Yogyakarta, Indonesia, in 2021. He is currently a lecturer at Universitas Dian Nuswantoro, Semarang, Indonesia. His research interests include information or image retrieval.