ICIC Express Letters
Volume **16**, Number **3**, March **2022**
 ISSN 1881-803X
pp. **289–297**

INDOBERT FOR INDONESIAN FAKE NEWS DETECTION

Sani Muhamad Isa*, Gary Nico and Mikhael Permana

BINUS Graduate Program – Master of Computer Science
Bina Nusantara University
Jl. K. H. Syahdan No. 9, Kemanggisan, Palmerah, Jakarta 11480, Indonesia
{ gary.nico; mikhael.permana }@binus.ac.id; *Corresponding author: sani.m.isa@binus.ac.id

Received April 2021; accepted July 2021

Abstract. *Fake news has been known as a deceptive information. In the digital era, especially in COVID-19 pandemic outbreak, fake news is used to mislead society for certain purposes. Social media such as Facebook, Twitter and WhatsApp are used as a platform to spread false news. To identify fake news, we have to manually verify the news with legitimate sources. However, this takes some effort and time rather than using fake news detection systems. A good fake news detection system is needed to reduce the spread of misleading information and those side effects of it. Most of the recent research in English fake news detection systems has already used deep learning models especially transformer models. However, research in Indonesian fake news detection systems is still using old machine learning approaches. In this study, we proposed IndoBERT, an Indonesian Bidirectional Encoder Representations from Transformers (BERT) based transformer model that focuses on the context and attention of the input sentences. For the experiment, we fine-tuned the proposed model with the dataset that was collected from turnbackhoax.id and adjusted hyperparameter to get the best result. Afterwards, we evaluated our model and achieved 94.66% score on precision, recall, and F1-score.*
Keywords: Fake news detection, BERT, IndoBERT, Classification

1. **Introduction.** Information is one of the critical things that should be delivered properly and wisely in this COVID-19 pandemic situation. Since digital technology became the part of human beings, any kind of information or news could have been spread rapidly. The problem is that during this COVID-19 pandemic situation, there are even more fake news spread on the Internet. Based on the data from the Ministry of Communication and Information Technology Indonesia, there is 1,028 fake news about COVID-19 as of August 8, 2020.

Fake news is a term used to represent false news, misleading information or propaganda. The general motive to spread such news is to mislead the readers, damage reputation or to gain from sensationalism [1]. Meanwhile, real news is authentic content for providing real information, coming from legitimate sources such as trusted journalists or reputable news agencies.

With the increasing number of Internet users and affordability of smartphones, most of Indonesian people have access to social media. This means, having better literacy is helpful in fighting the fear and stress related to the pandemic [2]. The conventional way to classify real or fake news is to manually verify the news with other sources; however, this takes some efforts and a lot of time.

Nowadays, transformer model is used for most of the classification tasks in Natural Language Processing (NLP). The reason is because transformer model has the attention mechanism [3], which will perform better than the other machine learning model. Unfortunately, there is still no research in Indonesian fake news detection using transformer-based model.

DOI: 10.24507/icicel.16.03.289

This paper is the first research using IndoBERT transformer-based model for fake news classification. We fine-tune IndoBERT pre-trained model and evaluate its performance with other machine learning models.

The remainder of this paper is organized as follows. Section 2 reviews the literature in the domain of hoax news classification. The proposed model is introduced in Section 3. Section 4 describes the dataset. In Section 5, the results of the proposed method are presented and discussed. Conclusion and future work are presented in the last section.

2. **Related Work.**

Ahmed et al. in [4] proposed n-gram modelling to test the n-gram length on the accuracy of six different classifiers (stochastic gradient descent, support vector machines, linear support vector machines, k-nearest neighbour, logistic regression, and decision trees), extracting word by word with Term Frequency-Inverse Document Frequency (TF-IDF) and Term Frequency (TF). It is turned out the proposed model achieved its highest accuracy when using unigram features and linear SVM classifier. The highest accuracy score is 92%.

Kudari et al. in [5] discussed passive-aggressive classifiers which are similar to the perceptron that does not require a learning rate and use TF-IDF vectorizer. For comparison, using the same dataset, the authors used count vectorizer and Naïve Bayes classifier and cross-paired each vectorizer with the models. The experiment showed that 90% of accuracy was obtained by using passive aggressive and TF-IDF vectorizer.

Rahutomo et al. in [6] proposed TF-IDF word embedding and Naive Bayes classifier for fake news detection in Indonesian. The dataset contained 600 news consisting of 372 real news and 228 fake news. The accuracy score is 82.6%.

From the previous section, we have addressed the increasing number of fake news due to COVID-19 pandemic situation. Here are some of the fake news detection research using COVID-19 fake news dataset.

Wani et al. [7] evaluated the performance of various deep learning models using the Constraint@AAA1 2021 COVID-19 fake news dataset. The chosen models are Convolutional Neural Networks (CNN), Long Short Term Memory (LSTM) and BERT. The result showed that BERT, a pre-trained transformer model, outperforms other deep learning models with 3%-4% differential in accuracy. The maximum accuracy achieved with BERT is 98.41%.

Glazkova et al. [8] focused on detecting fake news in social media experimented with three transformer-based models: BERT, RoBERTa, COVID-TWITTER-BERT (CT-BERT). The best model is CT-BERT with 10%-30% marginal improvement compared to its base model, BERT. CT-BERT is pre-trained on a large corpus of COVID-19 related text from Twitter. This model performed well at fake news classification with 98.69% of accuracy.

Based on all these research, we understand that transformer models like BERT outperform other traditional approaches. However, the recent research of fake news detection in Indonesian still uses traditional approaches with TF-IDF as the feature extractor and SVM or Naïve Bayes as the classifier. We developed fake news detection in Indonesian using a transformer model.

3. **Methodology.**

3.1. **Data preprocessing.**

The aim of data preprocessing is to use various NLP techniques to pre-process and prepare the data for the next step which is feature extraction. Because our news dataset came from various social media sources, we need to eliminate the noise by removing or normalizing the unnecessary tokens. Here are the data preprocessing pipeline.

- Remove Emojis: Emojis need to be removed from the dataset [9], because emojis are not readable and they are neither alphabets nor numbers.

- Remove Multiple Punctuations: We do not need multiple punctuations like '??', '!!', '???', and '!!!' to get the context of the sentence. So we can remove multiple punctuations into single punctuations using regex.
- Data Resampling: Our dataset suffered imbalanced class with 3,465 fake news and only 766 real news with a ratio of 5 : 1; this can cause problems in the later training process like low predictive accuracy for the infrequent class. Resampling is also known as Bootstrapping, a method that consists of drawing repeated samples from the original data sample [10]. In this case, we resampled the real news class to get a balanced number of 3,465. The final class ratio is 1 : 1.
- Split Data: For evaluating the performance of a machine learning model, the dataset will be split into training, validation and testing data [11]. The proposition of training data is 80% and testing data is 20%, while 20% of training data is used for validation data.
- Removing too many aspects from the dataset will decrease the variety of the word. We do not remove slang words; therefore, the model can process and learn from various kinds of input in the later training process.

3.2. **Machine learning approaches.**

3.2.1. *TF-IDF. Term Frequency-Inverse Document Frequency* is a weighting algorithm that is widely applied into language models to building NLP systems and used for information retrieval and text mining. Terms can be words or phrases. The product of TF-IDF is statistical measurement used as an indicator of the importance of a term [12].

TF(t) = (Number of times term t appears in a document)/(Total number of terms in the document).

IDF(t) = log_e(Total number of documents/Number of documents with term t in it).

The weight of the term = TF $*$ IDF.

After weighing the terms, it has to be converted into numbers, also called vectorization and will be treated as an input for the model. Vectorizing is the process of converting text into vectors. The numbers of the vectors represent the content of the text. TF-IDF gives us a way to associate each word in a document with a number that represents how relevant each word is in the document. Documents with similar relevant words will have similar vectors.

3.2.2. *SVM.* Support vector machine is a supervised machine learning algorithm [13], performed by separating a given set of binary labelled training data with maximum hyperplane that represents a separator between two classes in the input space. SVM does not use pure probability values such as Naïve Bayes, but uses margin or distance. The farther testing point from the hyperplane, the higher probability of that point can be classified [14].

$$P(y|x) = \frac{1}{1 + e^{(-yw^T x)}} \tag{1}$$

where $y = \pm 1$ as class label, x is data and $w \in Rn$ is a weight vector. T is a hyperparameter of prior distribution.

We combined TF-IDF for feature extraction and SVM for the classification. We need to maximize the margin between the classes, so SVM can find a hyperplane to divide the two classes, fake news and real news. We used a Radial Basis Function (RBF) kernel with a number of degree of one. RBF kernel is better than linear kernel function for predictive performance tasks [15].

3.2.3. *Naïve Bayes.* Naïve Bayes classifiers are known as very simple and fast yet effective linear classifiers [16]. This algorithm is based on Bayes' theorem with an assumption of independence among the features. Bayes' theorem describes probability of an event based

on its prior probability. The advantage of Naïve Bayes is that it performs well on small datasets.

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)} \tag{2}$$

where $P(c|x)$ is the posterior probability of class c given predictor, $P(c)$ is the prior probability of class, $P(x|c)$ is the likelihood probability of the predictor given class c, and $P(x)$ is the prior probability of the predictor.

We also combined TF-IDF for feature extraction and Naïve Bayes for classification. We used multinomial Naïve Bayes that are suitable for text classifications. The multinomial distribution requires fractional counts that we can get from the output of TF-IDF [17].

3.3. **IndoBERT.** IndoBERT is an Indonesian BERT based model which is trained on the Indo4B dataset [18]. This means both IndoBERT and BERT are transformer-based models. Transformer is the first sequence transduction model based entirely on attention, replacing the recurrent layers most commonly used in encoder-decoder architectures with multi-headed self-attention [3]. Transformer model consists of six layers of encoder and six layers of decoder. This transformer model is suitable for solving machine translation problems, like translating English to German or English to French. The encoder generates word embedding for the English sentence and the decoder uses the word embedding from the encoder to translate English word to the destination language. In simpler terms, the encoder learns the context and attention from the sentence given while the decoder translates language into another. Both encoder and decoder are trying to understand the language, this is what makes the transformer model special.

IndoBERT uses the same training strategies, Masked Language Model (MLM) and Next Sentence Prediction (NSP), just like the original Bidirectional Encoder Representations from Transformers (BERT) model [19], making it also a pre-trained model. The process of pre-training BERT is to train the model to perform MLM and NSP. MLM is like a fill-in-the-blank task, in which the model uses the surrounding masked words to predict what the masked word in the sentence should be. BERT is bidirectional so it can learn both left-to-right and right-to-left words at the same time, this helps the model to predict the masked word. NSP aims to train the model to understand the correlation between two sentences. Both pre-training processes that consist of MLM and NSP help BERT to understand the context and meaning of the language. We just simply need to fine tune the pre-trained model for the chosen NLP tasks.

Research shows that syntax-aware attention already exists in BERT, which may be one of the reasons for its success [20]. BERT can have a deeper sense in the language of the sentence because it is bidirectionally trained. BERT is a pre-trained model and it furthermore shows the importance and ease of use of transfer learning.

Indo4B, the dataset of IndoBERT, is a large-scale dataset collected around four billion words from Indonesian pre-processed text data. The dataset contains text from various sources like local online news, social media, Wikipedia, online articles, subtitles from video recordings, and parallel datasets. Indo4B covers both formal and casual Indonesian sentences, and compiled from two Indonesian casual sentence datasets, eight Indonesian formal sentence datasets and the rest have mixed both formal and casual sentences.

There are four IndoBERT pre-trained models based on the hyperparameter size: IndoBERTBASE, IndoBERTLARGE, IndoBERT-liteBASE and IndoBERT-liteLARGE, and the bigger hyperparameter size means also longer training duration. In this work we use IndoBERTBASE with a total parameter of 124.5 M, 12 numbers of layers, 12 self-attention heads per layer, and a hidden size of 768.

To use the pre-trained IndoBERT model, we needed to encode our input text from the dataset into a tensor format, and to do so we used BertTokenizer, inherited from Hugging Face PreTrainedTokenizer. IndoBERT input can take a single sentence or pair sentences

as a sequence. Each sequence (sentence) contains multiple tokens (words). The first token of every sequence is always [CLS] and the last token of a single sentence is [SEP]. If it is a pair of sentences, [SEP] is the boundary token for separating the pair sentences.

Example of single sentence: ['[CLS]', 'Pria', 'itu', 'pergi', 'ke', 'toko', '.', '[SEP]'].

Example of pair sentences: ['[CLS]', 'Pria', 'itu', 'pergi', 'ke', 'toko', '.', '[SEP]', 'Dia', 'membeli', 'satu', 'galon', 'susu', '.', '[SEP]'].

Each token was given a unique ID in the process of pre-training the IndoBERT model. We used BertTokenizer to assign each token into their unique ID in a tensor format based on the IndoBERTBASE vocabulary. Even though tensor is a format number, it needs to be converted into a vector for the input of the IndoBERT model. All this process is called word embedding.

The next step was to fine-tune the IndoBERT model by optimizing the hyperparameter using Adam Optimizer for the training process. Adam Optimizer is a method for efficient stochastic optimization that only requires first-order gradients with little memory requirement [21].

Adam uses the squared gradients to scale the learning rate and improve the use of momentum by using moving average of the gradient. We chose Adam Optimizer because it suits our huge hyperparameters model.

To evaluate how well our model learns the dataset, we used binary cross entropy, a loss function that is used in binary classification tasks [22]. Binary cross entropy is very convenient to solve classification problems; moreover, classification can be reduced to a binary choice. Since our classes are fake news and real news (binary option), binary cross entropy was the best option for our loss function.

$$Loss = \sum -t_i \log(y_i) - (1 - t_i) \log(1 - y_i) \tag{3}$$

where t is the target, y is the output, and i is the number of iterations of the output size.

For the classification model, we used the BertForSequenceClassification, which is a classification layer on top of the 12 layers of IndoBERT. This classification model is a Pytorch torch.nn.Module subclass.

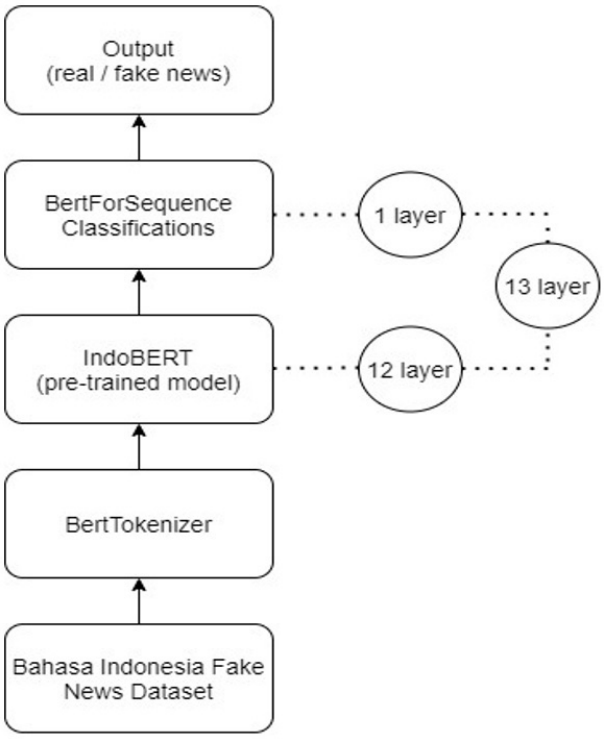

FIGURE 1. Model architecture

Based on Figure 1 we can conclude that our model architecture consists of 13 layers. 12 hidden layers of IndoBERT and 1 layer of classifier on the very top. Every layer does some multi-headed attention on the word embedding from the previous layer. The dimension in every layer is [number of tokens $\times$ 768 $\times$ batch size].

After several experiments, based on training loss function, we found that the best number of epochs was 8, batch size with 8 samples, and 2e-5 for the learning rate. We limited the maximum number of the tokens to 200 for each sequence.

4. **Data.** The dataset was scraped from turnbackhoax.id, a website that contains Indonesian news from July of 2015 until now. The news has labelled as true or fake news. All the news were collected from various sources like online news and social media like Facebook, Twitter and WhatsApp. Each news in turnbackhoax.id is updated daily and the labelling process was done manually by verifying the news with trusted sources.

From the scraped data, we could get title, narration, date, URL, media content, and news category. For this research, we used narration as the feature and news category as the label. We were able to collect 3,465 fake news and 766 real news. Table 1 shows the example of both fake and real news in the dataset.

Table 1. Narration and label example from the dataset

Narration	Label
Perjuangan FPI tahun 1940 melawan penjajah dari Belanda. Subhanallah	Fake
Penulis, pengirim, dan atau penyebar berita bohong atau hoax harus berhati-hati. Pelaku bisa diancam pidana penjara enam tahun dan atau denda Rp 1 miliar	Real

5. **Results and Discussion.** In this section, we compared the performance of several machine learning models and the IndoBERT model using confusion matrix. For binary classification problem, the discrimination evaluation of the best (optimal) solution during the classification training can be defined based on the confusion matrix [23].

From the confusion matrix, we obtained the precision, recall, F1-score and accuracy of the model. All three measures distinguish the correct classification of labels within different classes. They concentrate on one class (positive examples) [24].

In the testing process using the TF-IDF + SVM model shown in Table 2, we got 644 fake news out of 693 fake news data and 602 real news out of 693 real news data successfully classified.

Table 2. TF-IDF + SVM confusion matrix

FAKE	644	49
REAL	91	602
	FAKE	REAL

The experiment in Table 3 using another machine learning model, TF-IDF + Naïve Bayes successfully identified 657 fake news out of 693 fake news data and 497 real news out of 693 real news data.

Table 3. TF-IDF + Naïve Bayes confusion matrix

FAKE	657	36
REAL	196	497
	FAKE	REAL

IndoBERT has the best confusion matrix among SVM and Naïve Bayes shown in Table 4, with successfully classified 657 real news out of 693 real news data and 655 fake news out of 693 fake news data.

Table 4. IndoBERT confusion matrix

FAKE	655	38
REAL	36	657
	FAKE	REAL

From Table 5, we notice that TF-IDF + SVM and TF-IDF + Naïve Bayes are performing well and getting good accuracy scores, but comparing between the two models, TF-IDF + SVM performs better. TF-IDF + SVM took 5 minutes for training time and TF-IDF + Naïve Bayes only took less than 1 minute.

Table 5. Comparison of models performance

Model	Precision	Recall	F1-score	Accuracy
TF-IDF + SVM	90%	90%	90%	90%
TF-IDF + Naïve Bayes	83%	85%	83%	83%
IndoBERT	94.66%	94.665%	94.66%	94.66%

IndoBERT unsurprisingly outperforms those two models, and this is because IndoBERT is a pre-trained model. BERT is trained on Mask Language Model (MLM), which allows the model to have a deeper sense of understanding a language context from bidirectionally trained. IndoBERT has 124.5 M hyperparameters, and it means the model has high transfer learning capabilities to memorize the target class (pattern) of each training sample.

Meanwhile TF-IDF, as feature extraction, works on the number of appearances of the word rather than BertTokenizer that works on attention of the word. However, the disadvantages of the IndoBERT transformer model are that it needs a lot of data and takes longer process time than the other two models. IndoBERT took 15 minutes for training time, and this means three times longer than TF-IDF + SVM and fifteen times longer than TF-IDF + Naïve Bayes.

Figure 2 shows the training and validation loss of the IndoBERT model, where validation loss is increasing after the training process. This might happen when the model with

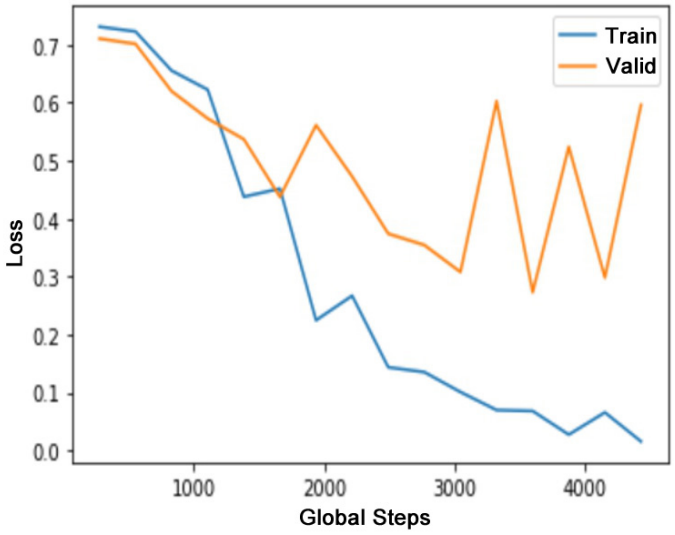

Figure 2. Training/validation loss

a huge amount of hyperparameters trained with a small amount of data, causes the model to overfitting. In other words, the model is too complex so we must reduce the number of layers. However, since IndoBERT is a pre-trained transformer model, we cannot modify the model architecture. One of the options is to increase the number of data to match the IndoBERT model [25].

6. **Conclusions.** Fake news has spread rapidly especially in the pandemic situation, and this can lead to misleading information and much worse scenarios. To reduce the spread of Indonesian fake news, we proposed the IndoBERT transformer model as the solution.

The dataset was gathered from turnbackhoax.id with 3,465 fake news and 766 real news. From the experiment we have proved that building Indonesian fake news detection using the IndoBERT transformer model is possible with the accuracy score of 94.66%. We also found that our model is overfitting because the dataset was not big enough to train our model that has a huge number of hyperparameters.

In the future work, we can experiment with more data and explore another transformer model like BERT Multilingual and DistilBERT Multilingual for fake news detection. Moreover, we can explore hybrid models combining IndoBERT with another transformer model and use other NLP methods [26,27].

REFERENCES

[1] A. Thota, P. Tilak, S. Ahluwalia and N. Lohia, Fake news detection: A deep learning approach, *SMU Data Science Review*, vol.1, p.1, 2018.

[2] A. Moscadelli, G. Albora, M. A. Biamonte, D. Giorgetti, M. Innocenzio, M. S. Paoli, C. Lorini, P. Bonanni and G. Bonaccorsi, Fake news and COVID-19 in Italy: Results of a quantitative observational study, *International Journal of Environmental Research and Public Health*, 2020.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin, Attention is all you need, *The 31st Conference on Neural Information Processing Systems (NIPS2017)*, Long Beach, CA, USA, 2017.

[4] H. Ahmed, I. Traore and S. Saad, Detection of online fake news using n-gram analysis and machine learning techniques, in *Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments. ISDDC 2017. Lecture Notes in Computer Science*, I. Traore, I. Woungang and A. Awad (eds.), Cham, Springer, DOI: 10.1007/978-3-319-69155-8_9, 2017.

[5] J. M. Kudari, V. Varsha, B. G. Monica and R. Archana, Fake news detection using passive aggressive and TF-IDF vectorizer, *International Research Journal of Engineering and Technology (IRJET)*, vol.7, no.9, pp.1-3, 2020.

[6] F. Rahutomo, I. Y. R. Pratiwi and D. M. Ramadhani, Naïve Bayes experiment for Indonesian hoax news detection, *JPKOP*, pp.3-12, 2019.

[7] A. Wani, I. Joshi, S. Khandve, V. Wagh and R. Joshi, Evaluating deep learning approaches for COVID-19 fake news detection, *arXiv.org*, arXiv: 2101.04012, 2021.

[8] A. Glazkova, M. Glazkov and T. Trifonov, Exploiting CT-BERT and ensembling learning for COVID-19 fake news detection, *arXiv.org*, arXiv: 2012.11967, 2021.

[9] M. Redmond, S. Salesi and G. Cosma, A novel approach based on an extended cuckoo search algorithm for the classification of tweets which contain emoticon and emoji, *2017 2nd International Conference on Knowledge Engineering and Applications (ICKEA)*, pp.13-19, DOI: 10.1109/ICKEA.2017.8169894, 2017.

[10] P. M. Dixon, Bootstrap resampling, *Encyclopedia of Environmetrics*, pp.212-220, 2006.

[11] Z. Reitermanová, *Data Splitting*, Charles University, Czech Republic, 2010.

[12] T. Xia and Y. Chai, An improvement to TF-IDF: Term distribution based term weight algorithm, *Journal of Software*, vol.6, pp.413-420, 2011.

[13] T. S. Furey, N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer and D. Haussler, Support vector machine classification and validation of cancer tissue samples using microarray expression data, *Bioinformatics*, vol.16, pp.906-914, 2000.

[14] A. B. Prasetijo, R. R. Isnanto, D. Eridani, Y. A. Adi Soetrisno, M. Arfan and A. Sofwan, Hoax detection system on Indonesian news sites based on text classification using SVM and SGD, *ICITACEE*, p.47, 2017.

[15] T. Senechal, D. McDuff and R. E. Kaliouby, Facial action unit detection using active learning and an efficient non-linear kernel approximation, *2015 IEEE International Conference on Computer Vision Workshop*, 2015.

[16] I. Rish, An empirical study of the Naïve Bayes classifier, *Workshop on Empirical Methods in Artificial Intelligence*, pp.41-46, 2001.

[17] S. Xu, Y. Li and Z. Wang, Bayesian multinomial Naïve Bayes classifier to text classification, in *Advanced Multimedia and Ubiquitous Engineering. FutureTech 2017, MUE 2017. Lecture Notes in Electrical Engineering*, J. Park, S. C. Chen and K. K. R. Choo (eds.), Singapore, Springer, DOI: 10.1007/978-981-10-5041-1_57, 2017.

[18] B. Wilie, K. Vincentio, G. I. Winata, S. Cahyawijaya, X. Li, Z. Y. Lim, S. Soleman, R. Mahendra, P. Fung, S. Bahar and A. Purwarianti, IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding, *AACL-IJCNLP*, 2020.

[19] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, *2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol.1, pp.4171-4186, 2019.

[20] K. Clark, U. Khandelwal, O. Levy and C. D. Manning, What does BERT look at? An analysis of BERT's attention, *2019 ACL Workshop BlackBoxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp.276-286, 2019.

[21] D. P. Kingma and J. Ba, Adam: A method for stochastic optimization, *The 3rd International Conference for Learning Representations*, San Diego, 2015.

[22] D. U. Ruby, P. Theerthagiri, D. J. Jacob and D. Vamsidhar, Binary cross entropy with deep learning technique for image classification, *International Journal of Advanced Trends in Computer Science and Engineering*, vol.9, p.4, 2020.

[23] M. Hossin and M. N. Sulaiman, A review on evaluation metrics for data classification evaluations, *International Journal of Data Mining & Knowledge Management Process (IJDKP)*, vol.5, 2015.

[24] M. Sokolova, N. Japkowicz and S. Szpakowicz, Beyond accuracy, F-score and ROC: A family of discriminant measures for performance evaluation, *Advances in Artificial Intelligence*, pp.2-3, 2006.

[25] X. Ying, An overview of overfitting and its solutions, *Journal of Physics: Conf. Series*, DOI: 10.1088/1742-6596/1168/2/022022, 2019.

[26] L. Tang, Combining BERT with WordNet sense embeddings to predict graded word similarity changes, *Proc. of the 14th Workshop on Semantic Evaluation*, pp.166-170, 2020.

[27] N. Peinelt, D. Nguyen and M. Liakata, tBERT: Topic models and BERT joining forces for semantic similarity detection, *The 58th Annual Meeting of the Association for Computational Linguistics*, pp.7047-7055, 2020.