International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Indonesian News Classification Using IndoBert

Budi Juarto*[1], Yulianto[2]

Submitted: 12/11/2022 **Accepted**: 14/02/2023

Abstract: In 2021 there was an increase in the number of people who have internet access, and the number of users increased from 175 million users to 202 million users. News classification in general still uses traditional techniques such as word embedding with TF-IDF and machine learning. The latest development for technology that can classify NLP news using the BERT model is a state-of-the-art pre-trained model. However, the pre-train model of BERT is only limited to use in English. So in this study, IndoBERT will be used in making news recommendations based on the category. This dataset uses an Indonesian news dataset that has 5 categories, including football, news, business, technology, and automotive. The IndoBERT method will be compared with other pre-train models, such as XLNET, BERT multilingual, XLMRoberta. Meanwhile, the machine learning method with TF-IDF word embedding was compared using the XGBoost method, LGB, and random forest. In this study, we see that the classification method using IndoBERT gives the best results with an accuracy value of 94% and also provides the smallest computation time compared to other methods with a time of one minute 56 seconds and a validation time of 10 seconds. BERT can give the best results because BERT is a type of pre-trained model that is trained from various kinds of Indonesian words such as news and several website sources to add to the corpus of vocabulary sources in the model. In the future research will be carried out to implement the dual IndoBERT model and the Siamese IndoBERT.

Keywords: *Natural Language Processing, News Classification, IndoBERT, Multiclass Classification*

1. Introduction

The number of people using the internet in Indonesia has consistently shown steady growth from one year to the next in their usage. There has been an increase in the number of people using the internet from 175 million users to 202 million users; this statistic has increased by eleven percent or as many as 27 million users since 2021. The number of people using the internet is now 202 million users [1]. There are 175 million people who use the internet, and of those people, there are 160 million active users of social media who also use the internet. The present population of Indonesia is 272 million [2], and within that number, there are types of individuals who do not actively use the internet, such as toddlers and the elderly. Therefore, this figure is a significant amount. People that use the internet in big quantities may find that they can obtain news information via the internet in a way that is both simple and rapid, increasing the likelihood that they will do so.

According to a source from The Wall Street Journal, the circulation of news online runs quickly and a lot every day, and it has reached an average number of three million and five hundred thousand readers by the year 2021. This number is derived from the typical amount of news that is published in a given year. six days a week. If we extrapolate this data further, we find that the daily readership of print publications is currently at 734 thousand copies, while the daily circulation of digital publications has reached over 2.7 million [3]. The categories of news that are held by news organizations are typically somewhat varied; for example, the Wall Street Journal covers politics, business, and economics in its reporting. Because of the vast volume of news, it will be challenging to categorize each piece of news according to its respective topic. The incorporation of machine learning and deep learning algorithms will make it simpler for editors of news publications to classify stories following the appropriate headings.

Previous research has been carried out in the instance of Indonesian news using classification using machine learning XGB utilizing the TF-IDF or a method that counts the number of words in a document. [4] In addition, the XGB approach is utilized to categorize the severity of drought in the Boyolali and Purwerjo regions of Indonesia. These regions are located in a part of the country where the results of the XGB method are superior to those of random forests [5]. LightGBM is another machine learning method that is employed. This study found that methods such as random forest and also SVM or also called support vector machines have advantages when dealing with large datasets and are suitable for classifying land cover cases in this study [6]

[1] *Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480*
ORCID ID : 0000-0002-5134-2989
[2] *Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480*
ORCID ID : 0000-0001-9668-3819
** Corresponding Author Email: budi.juarto@binus.ac.id*

Classification in the field of machine learning in the field of natural language processing still mostly uses TD-IDF, which has the disadvantage of being context-independent in making word representations, whereas word vectorization using BERT is more because it is context dependent so that it can provide better word vector representation [7]. Methods such as TF-IDF are still widely used in the field of text classification by combining them with several machine learning models where TF-IDF is more focused on the process of converting text into vectors so that it can be understood by models in machine learning. When it comes to text categorization, the BERT technique is superior to both TF-IDF and machine learning [8]. In addition to this, the BERT method also performs better. The previous study also continues to use BERT as the classification system for English news [9]. In this investigation, we will apply IndoBERT to the task of news categorization and evaluate its performance relative to that of other pre-trained transformer models.

2. Related Works

Machine learning methods have been widely used since the beginning of research in the field of news classification. After conducting a literature study, it can be seen that several machine learning methods that are often used are SVM, naive Bayes, KNN, random forest, and others. Research about comparative study of the SVM, KNN, and naive-Bayes algorithms in classifying Myanmar language news. The dataset used has 4 classes, namely politics, business, entertainment, and sports with the best performance obtained by the SVM method[10].

Extreme Gradient Boosting method in classifying Indonesian news. The classification task carried out is to distinguish news that is hoax news and valid news. The dataset carried out is a news dataset collected from several online news sites. The dataset was collected from 2015 to 2020 with the number of studies conducted in the study was 500. The distribution of training and test data with a proportion of 80% training data and 20% testing data. So that there are 400 news that becomes training data and 100 that become testing data. In this study, converting each word into a vector using the TF-IDF method. The results obtained in the study are an accuracy of 92%[4].

Research conducted by (Prasetyo et al, 2019) entitled "Data Analysis of Landsat 8 OLI Imagery as Drought Prediction Index Using Machine Learning in Boyolali and Purworejo Regencies. The results of this research can be used as a consideration for the government to regulate land and also water availability to overcome the problems of climate change, drought, and also the problem of supply and demand of water needed by humans to meet their daily needs. In this study, the data used were Landsat 8 OLI satellite imagery in Boyolali and Purworejo districts which were then analyzed using two methods, namely XGBoost and Random Forest methods. After evaluating the XGBoost and Random Forest methods, an accuracy value of 82% and a kappa value of 65% were obtained, while the random forest method yielded much lower results with an accuracy value of 68% and a kappa value of 37%. This shows that the XGBoost method is better to use than the Random Forest method because it has a higher accuracy and kappa value[5].

Shweta D. Mahajan (2021) conducted research on news classification using the TF-IDF method for word vectorization and naive-bayes algorithm as a classifier. The dataset used is an English dataset with three classes which is travel, politics, and criminal classes[11].

Shah et al conducted a study on news classification using several machine learning algorithms such as logistic regression, random forest, and also KNN [12]. Metric measurements carried out in this study is to use accuracy as an evaluation metric. In this study it was measured using an accuracy metric to measure how well the model gave results in a prediction, the first method using logistic regression gave results of how much accuracy was 97%, and another method was followed with an accuracy of 93%, namely using the random forest method. While the method that gives the lowest accuracy results is KNN with an accuracy value of only 92%.

Deep learning methods have increased in popularity in recent times due to their performance which can rival and beat traditional machine learning methods. Sari et al conducted a study on the detection of AGNews news using the LSTM (Long-Short Term Memory) method using word2vec embeddings[13]. The dataset used in this study is the AGNews dataset which has 4 labels with a total of 496.835 data. The results obtained in this study are the LSTM model with word2vec produces an accuracy of 95.38% with an average value of precision, recall, and F1-score of 95%.

The traditional method of machine learning has the disadvantage that it requires large labeled data to train it. Chenbin Li et al., (2018) used the Bi-LSTM deep learning method combined with CNN to perform text classification tasks [14]. This study uses the word2vec embedding method to generate input vectors that will be processed by the model. The developed model was compared with the TF-IDF, SVM, LSTM and CNN methods and it was concluded that the designed model had the best performance with an F1 of 0.99.

Berfu Buy¨uk¨ oz e al., (2020) classified socio-political news data using the ELMo and DistilBERT methods[15]. As comparisons for the designed method, the Naive Bayes multinomial and linear support vector are employed. The research concludes that the DistilBERT technique

outperforms ELMo in the field of semantic data value processing. Additionally, the DistilBERT model is 30% lower in size and 83% quicker in processing speed than ELMo.

3. Methodology

3.1. Research Data

The dataset used in this investigation is the secondary dataset retrieved from GitHub by A.Chandra (2020)[16]. News information is gathered from a variety of Indonesian news outlets. There are 6127 lines of training data and 2627 lines of testing data in the dataset. The dataset has five categories: news, business, technology, football, and automobiles. We keep the pre-processing basic by merely removing stop words and transforming lowercase letters. Before data is ready to be used, there are some pre-processing processes before the sentences are used in the IndoBERT. After pre-processing data the sentence will be tokenized with IndoBERT tokenizer and then load the IndoBERT model for the training of the model to evaluate in data evaluation and data testing. The process in this study is described in Fig 1.

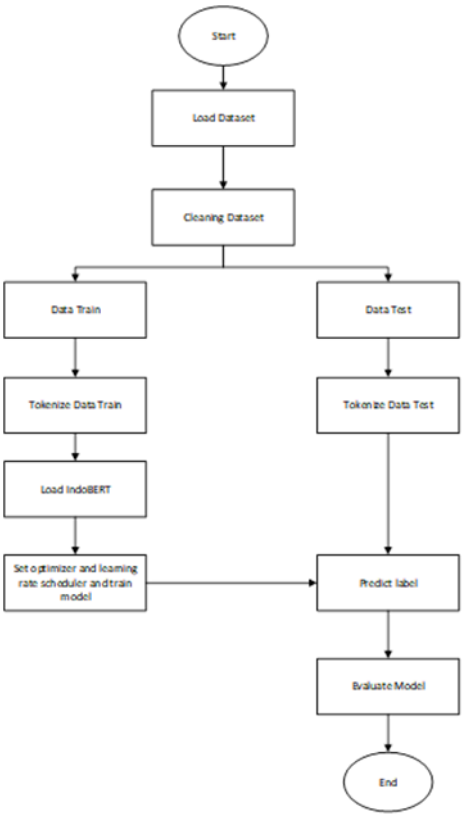

Fig. 1 Flowchart for IndoBERT Research Process

3.2. Bidirectional Encoder Representations from Transformers (BERT)

BERT is an architectural model developed from Transformer by taking transformer encoder layers and stacking them to build a new architecture [17]. The original BERT consists of 2 types, namely BERT-base with 12 encoder layers and BERT-large with 24 encoder layers. Like transformers, BERT is transfer learning and can be trained with unlabelled data. 2 types of BERT training methods are carried out simultaneously, namely mask language prediction and next sentence prediction. The results of training with large data without labels will produce a model that has general knowledge about the relationship between words and sentences. Furthermore, the model can be fine-tuned with a specific dataset that is small and has been labeled according to the task to be completed. BERT Architecture is shown in Figure 2.

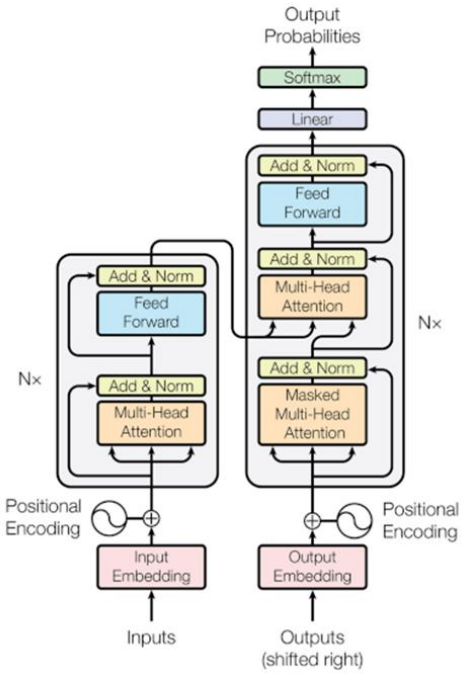

Fig. 2 IndoBERT Architecture

BERT uses Transformers which can learn the relationship between words by using a mechanism called attention. BERT has an encoder and a decoder in its architecture. The difference lies in the encoder whose function is to receive input and turn it into a word vector while the decoder functions to make predictions from various tasks. To be able to make words into a vector in the encoder, BERT has three kinds of embeddings. This is because the BERT model creates a language model that can be represented

properly in vector form so that it can provide good predictive model results. Some of the embeddings that are carried out in the BERT input include:

1. Token embeddings: tokens that are at the beginning of a sentence represented by [CLS] while at the end of a sentence represented by tokens by [SEP].
2. embeddings of segments: This embedding is useful to be used to distinguish between the two sentences by marking each sentence.
3. Positional embeddings: These embeddings provide a token that is useful for providing positional information on an embedding in a sentence.

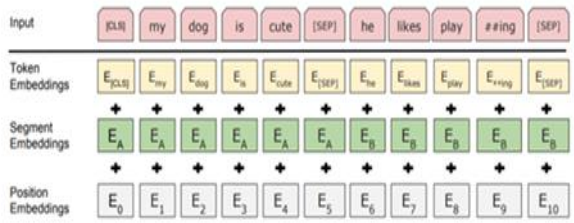

Fig. 3 IndoBERT Architecture

The Transformer, in essence, applies a layer that maps sequences to sequences, producing an output that is also a series of vectors with input and output tokens corresponding 1:1 at the same index. And as we already know, BERT does not attempt to predict the next word that will be spoken. Two strategies are used during training.

3.3. IndoBERT

IndoBert is one of the state-of-the-art in conducting text analysis using Indonesian. The architecture that is built is using the transformer model on BERT in general which uses English. As in BERT, this method also uses twelve hidden layers where each hidden layer is limited to a dimension of 786 and also uses 12 attention heads. IndoBERT was also built using a large number of Indonesian vocabulary with a total of more than two hundred twenty million words where the sources obtained are language sources that use good and correct Indonesian such as from online newspapers and also from the Indonesian Web Corpus and other sources. other. IndoBERT itself is a pre-trained model that has been developed previously using 2.4 million steps or 180 epochs. The training process takes up to two months. So that IndoBERT has good performance for several tasks in NLP cases. In testing the IndoLEM dataset, IndoBERT has provided better results than MalayBERT.

3.4. XL-NET

Yang et al. (2019) proposed a new model dubbed XLNet that takes into account the position of masked words on the objective of masked language modeling [19]. XLNet aimed to optimize auto-regressing and auto-encoding models while avoiding their respective limitations. This research also proposed a new objective referred to as permutation language modeling, which permits the model to retain its auto-regression modeling advantage while also capturing bidirectional context. XLNet's neural architecture is designed to work seamlessly with the autoregressive objective, including the incorporation of Transformer-XL and the meticulous design of a two-stream attention mechanism. Yang et al. observe that both BERT and XLNet perform partial prediction. In the BERT case, if all tokens are masked, it is impossible to make predictions with any degree of accuracy. In addition, for both BERT and XL-NET, partial prediction reduces the difficulty of optimization by predicting only tokens with sufficient context.

3.5. XLM

Cross-lingual language models (XLMs) are models for multiple languages that have been trained using two distinct methodologies. One unsupervised, which uses solely monolingual data, and one supervised, which uses parallel data and a novel cross-lingual language model objective[20]. Without the use of a single parallel sentence, a cross-lingual language model trained on the XNLI cross-lingual classification benchmark exceeds the previous supervised state of the art by an average of 1.3% accuracy.

3.6. ROBERTA

The ROBERTA model is intended to optimize the BERT model by analyzing the effect of parameters and training data size[21]. The modification consists of training the model with a larger batch but fewer data points. Remove the prediction aim for the next sentence. The training uses a longer sequence and dynamically modifies the masking pattern applied to the data.

3.7. Multilingual BERT

Multilingual BERT is a pre-trained model which is continued from BERT which was previously only trained in English but now is also trained in various languages. There are 104 languages trained in multilingual BERT which can provide good annotations between various languages. Just like the BERT model, this model uses 12 layers of the transformer model and uses Wikipedia data from various languages, not only from English. BERT has good results for various languages and Indonesian is one of the languages that has been included in this pre-trained model. So in this study, we will also try the Multilingual BERT model to compare its performance to the IndoBERT model.

4. Results and Discussion

This study uses accuracy, precision, and recall evaluation metrics for the evaluation process to be compared to other

deep learning and machine learning algorithm. These evaluation metrics are used to predict whether each piece of news will be grouped in the correct class. These evaluation metrics are some formulas are shown in equation 1, equation 2, and equation 3. Some of the terms used in this equation, such as:

- True Positives (TP_i) – condition if the prediction result is positive then the actual value is also positive so that the classification is correct.
- True Negatives (TN_i) – condition if the prediction result gives a negative label and the actual label is negative so that the classification is correct.
- False Positives (FP_i) – condition if the prediction results give a negative label even though the actual value is positive so that the classification gives the wrong label.
- False Negatives (FN_i) – conditions where the prediction results give positive results but the actual label is negative so that the classification gives the wrong label.

$$Accuracy = \frac{\sum_{i=1}^{l} \frac{TP_i+TN_i}{TP_i+TN_i+FP_i+FN_i}}{l} x100\% \quad (1)$$

$$Precision = \frac{\sum_{i=1}^{l} TP_i}{\sum_{i=1}^{l} (FP_{i+}TP_i)} x100\% \quad (2)$$

$$Recall = \frac{\sum_{i=1}^{l} TP_i}{\sum_{i=1}^{l} (TP_{i+}FN_i)} x100\% \quad (3)$$

In this case, the classification uses 5 news labels so that the evaluation uses the Confusion Matrix Multiclass. The multi-class confusion matrix table used in this study is shown in Table 1.

TABLE 1 CONFUSION MATRIX FOR NEWS CLASSIFICATION

	Predicted Soccer	Predicted News	Predicted Business	Predicted Technology	Predicted Automotive
Actual Soccer	TP Soccer				
Actual news		TP News			
Actual business			TP Business		
Actual technology				TP Technology	
Actual automotive					TP Automotive

Based on formulas 1, 2, and 3, we can evaluate the result of IndoBERT compared to machine learning and deep learning model. The result of the evaluation is shown in Table 2 and Table 3.

TABLE 2. COMPARISON OF THE EVALUATION RESULTS OF THE MACHINE LEARNING METHOD WITH INDOBERT

Method	**Evaluation Metric**		
	Accuracy	**Precision**	**Recall**
XGB	0.89	0.89	0.89
LGBMC	0.92	0.92	0.92
Random Forest	0.88	0.89	0.88
IndoBert	0.95		

TABLE 3. COMPARATIVE EVALUATION OF VARIOUS PRE-TRAINED MODELS

Model	**Accuracy**	**Training Time**	**Validation Time**
XLNet	88.4	00:03:10	00:00:20
BERT Multilingual	91.7	00:01:59	00:00:10
XLM-RoBERTa	91.9	00:02:01	00:00:10
IndoBERT	94.5	00:01:56	00:00:10

Table 3 shows the result between IndoBERT and other machine learning method. IndoBERT has higher accuracy than other machine learning methods. In the Table, II IndoBERT is compared to another deep learning method for multilingual language models, such as XLNet, BERT Multilingual, IndoBERT, and XLM-RoBERTa. IndoBERT has the lowest training loss, validation loss, and validation time with the highest accuracy than other deep learning methods. The training and validation loss for each epoch is shown in Figure 4, Figrue 5, Figure 6. and Figure 7. Table 2 and Table 3 show the accuracy summary and IndoBERT has the highest accuracy with a score of 94.5%.

Based on the results of this study it can be seen that the IndoBERT method has the best results of all methods because IndoBERT is trained using more than two hundred

million words obtained from news sources, Wikipedia, and also from Indonesian Web Corpus website sources so that to carry out the classification task IndoBERT has a better pre-trained model than some machine learning classification models and other deep learning pre-trained models.

Fig 4. Training and validation loss XL-Net

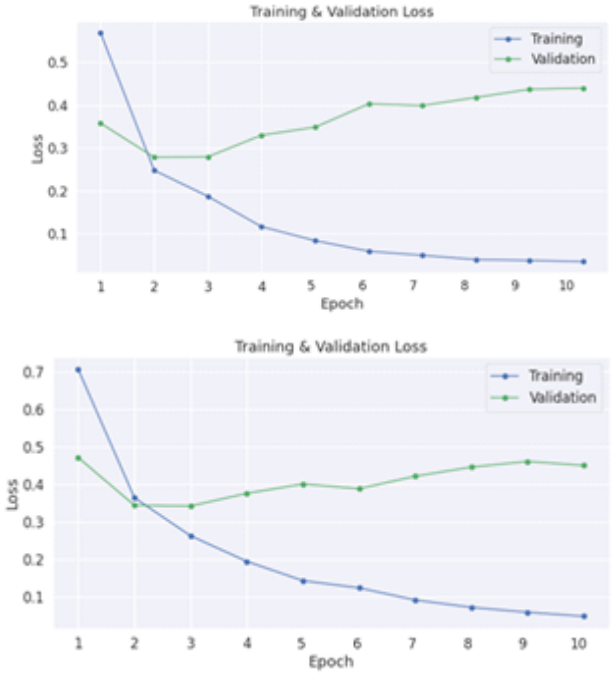

Fig 5. Training and validation loss BERT Multilingual

Fig 6. Training and validation loss IndoBERT

Fig 7. Training and validation loss XLM ROBERTA

5. Conclusions

We compare IndoBERT with machine learning models such as XGB, random forest, and light gradient boosting. The results show that the accuracy of the IndoBERT model provides the highest accuracy value, which is 94.5%, while the machine learning model which has the highest accuracy in the light gradient boosting model is 92%. Next, IndoBERT is compared with models such as XLMNet, XLM Roberta, and multi-language BERT that support multiple languages such as Indonesian. The results show that IndoBERT produces the highest accuracy value compared to the XLMNet, XLM Roberta, and multi-language BERT models. The IndoBERT model also produces the smallest training loss and validation loss values compared to other pre-trained models. The shortest computation time is in the IndoBERT model compared to other pre-trained models. IndoBERT provides the best classification results because IndoBERT's pre-trained models are trained from Indonesian news.

For future research, the Siamese BERT model can be developed using dual BERT. The first BERT is used for news headlines and the second BERT is used for news content included in the Siamese BERT. Another further improvement is the IndoBert model can use other BERT models in Indonesian.

References

[1] A. T. Haryanto, "Riset: Ada 175,4 Juta Pengguna Internet di Indonesia," https://inet.detik.com/, 2020. https://inet.detik.com/cyberlife/d-4907674/riset-ada-1752-juta-pengguna-internet-di-indonesia.

[2] P. Agustini, "Warganet Meningkat, Indonesia Perlu Tingkatkan Nilai Budaya di Internet," aptika.kominfo.go.id, 2021. https://aptika.kominfo.go.id/2021/09/warganet-meningkat-indonesia-perlu-tingkatkan-nilai-budaya-di-internet/ (accessed Sep. 12, 2021).

[3] A. Watson, "Average circulation of the Wall Street Journal from 2018 to 2020," www.statista.com, 2020. https://www.statista.com/statistics/193788/average-paid-circulation-of-the-wall-street-journal/ (accessed Jul. 07, 2021).

[4] J. P. Haumahu, S. D. H. Permana, and Y. Yaddarabullah, "Fake news classification for Indonesian news using Extreme Gradient Boosting (XGBoost)," IOP Conf. Ser. Mater. Sci. Eng., vol. 1098, no. 5, p. 052081, 2021, doi: 10.1088/1757-899x/1098/5/052081.

[5] S. Y. J. Prasetyo, Y. B. Christianto, and K. D. Hartomo, "Analisis Data Citra Landsat 8 OLI Sebagai Indeks Prediksi Kekeringan Menggunakan Machine

Learning di Wilayah Kabupaten Boyolali dan Purworejo," Indones. J. Model. Comput., vol. 2, no. 2, pp. 25–36, 2019, [Online]. Available: https://ejournal.uksw.edu/icm/article/view/2954.

[6] D. A. McCarty, H. W. Kim, and H. K. Lee, "Evaluation of light gradient boosted machine learning technique in large scale land use and land cover classification," Environ. - MDPI, vol. 7, no. 10, pp. 1–22, 2020, doi: 10.3390/environments7100084.

[7] A. K. Singh and M. Shashi, "Vectorization of text documents for identifying unifiable news articles," Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 7, pp. 305–310, 2019, doi: 10.14569/ijacsa.2019.0100742.

[8] S. González-Carvajal and E. C. Garrido-Merchán, "Comparing BERT against traditional machine learning text classification," no. Ml, 2020, [Online]. Available: http://arxiv.org/abs/2005.13012.

[9] K. S. Nugroho, A. Y. Sukmadewa, and N. Yudistira, "Large-Scale News Classification using BERT Language Model: Spark NLP Approach," 2021, [Online]. Available: http://arxiv.org/abs/2107.06785.

[10] K. Thandar Nwet, "Machine Learning Algorithms for Myanmar News Classification," Int. J. Nat. Lang. Comput., vol. 8, no. 4, pp. 17–24, 2019, doi: 10.5121/ijnlc.2019.8402.

[11] S. MAHAJAN, "News Classification Using Machine Learning," Int. J. Recent Innov. Trends Comput. Commun., vol. 9, no. 5, pp. 23–27, 2021, doi: 10.17762/ijritcc.v9i5.5464.

[12] K. Shah, H. Patel, D. Sanghvi, and M. Shah, "A Comparative Analysis of Logistic Regression, Random Forest and KNN Models for the Text Classification," Augment. Hum. Res., vol. 5, no. 1, 2020, doi: 10.1007/s41133-020-00032-0.

[13] W. K. Sari, D. P. Rini, R. F. Malik, and I. S. B. Azhar, "Klasifikasi Teks Multilabel pada Artikel Berita Menggunakan Long Short- Term Memory dengan Word2Vec," Resti, vol. 1, no. 10, pp. 276–285, 2017.

[14] [C. Li, G. Zhan, and Z. Li, "News Text Classification Based on Improved Bi-LSTM-CNN," Proc. - 9th Int. Conf. Inf. Technol. Med. Educ. ITME 2018, pp. 890–893, 2018, doi: 10.1109/ITME.2018.00199.

[15] B. Büyüköz, A. Hürriyetoğlu, and A. Özgür, "Analyzing ELMo and DistilBERT on Socio-political News Classification," Proc. Work. Autom. Extr. Socio-political Events from News 2020, no. May, pp. 9–18, 2020, [Online]. Available: https://www.aclweb.org/anthology/2020.aespen-1.4.

[16] A. Chandra, "Indonesian News Dataset," https://github.com/, 2020. https://github.com/andreaschandra/indonesian-news.

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

[18] [F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," pp. 757–770, 2021, doi: 10.18653/v1/2020.coling-main.66.

[19] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized autoregressive pretraining for language understanding," Adv. Neural Inf. Process. Syst., vol. 32, no. NeurIPS, pp. 1–11, 2019.

[20] [A. Conneau and G. Lample, "Cross-lingual language model pretraining," Adv. Neural Inf. Process. Syst., vol. 32, 2019.

[21] J. Libovický, R. Rosa, and A. Fraser, "On the Language Neutrality of Pre-trained Multilingual Representations," pp. 1663–1674, 2020, doi: 10.18653/v1/2020.findings-emnlp.150.