JURNAL UPI

Journal of Software Engineering, Information and Communication Technology (SEICT)

Journal homepage: https://ejournal.upi.edu/index.php/SEICT

SEICT

The Investigation of Convolution Layer Structure on BERT-C-LSTM for Topic Classification of Indonesian News Headlines

Dzakira Fabillah[1], Rizka Auliarahmi[2], Siti Dwi Setiarini[3], Trisna Gelar[4]*

[1,2,3,4]Department of Computer and Informatics Engineering, Politeknik Negeri Bandung, Indonesia
Correspondence: E-mail: trisna.gelar@polban.ac.id

ABSTRACT

An efficient and accurate method for classifying news articles based on their topics is essential for various applications, such as personalized news recommendation systems and market research. Manual classification methods are tedious, prompting deep learning techniques in this study to automate the process. The developed model, BERT-C-LSTM, combines BERT, the convolutional layer from CNN, and LSTM, leveraging their individual strengths. BERT excels at transforming text into context-dependent vector representations. The design of the classification model employs a blend of convolutional layers and LSTM referred to as C-LSTM. The convolutional layer possesses the capability to extract salient elements, including keywords and phrases, from input data.

On the other hand, the Long Short-Term Memory (LSTM) model exhibits the ability to comprehend the temporal context present in sequential data. This study investigates the influence of the convolutional layer structure in BERT-C-LSTM on classifying Indonesian news headlines into eight topics. The results indicate no significant differences in accuracy between BERT-C-LSTM model architectures with a single convolutional layer and multiple parallel convolutional layers and the models using various filter sizes. Furthermore, the BERT-C-LSTM model achieves an accuracy that is not much different from the BERT-LSTM and BERT-CNN models, with accuracies reaching 92.6%, 92.1%, and 92.7%, respectively.

ARTICLE INFO

Article History:
Submitted/Received 12 Oct 2023
First Revised 01 Nov 2023
Accepted 18 Nov 2023
First Available online 01 Dec 2023
Publication Date 01 Dec 2023

Keyword:
BERT,
Convolution Layer,
IndoBERT,
LSTM,
News,
Word Embedding.

1. INTRODUCTION

As the volume of online news articles grows, manually classifying them becomes increasingly challenging. The need for an efficient and accurate way to classify news articles based on their topics is crucial for various applications, including personalized news recommendation systems and market research.

To solve this problem, researchers and data scientists are leveraging deep learning to automate the classification process. One popular method is classifying news articles based on their headlines because using title data requires less computational costs than news content. In addition, the topic of a news story tends to be described in the headlines. The widely used deep learning methods for text classification include LSTM, CNN, and RNN.

In the field of headline news classification, three distinct methodologies were examined, namely LSTM, CNN, and RNN. Notably, LSTM exhibited the highest level of accuracy when compared to the other two methods. The study revealed that the Long Short-Term Memory (LSTM) model exhibited a 2% increase in accuracy compared to the Recurrent Neural Network (RNN) model and a 3% increase in accuracy compared to the Convolutional Neural Network (CNN) model (Kandhro et al., 2020). In addition, a separate study conducted a comparative analysis of several methodologies, wherein the Long Short-Term Memory (LSTM) approach showed superior performance in comparison to Convolutional Neural Networks (CNN) and other neural network techniques, with a notable margin of 3-6% (Chowdhury et al., 2022).

While it has been demonstrated that LSTM models tend to exhibit higher accuracy compared to CNN models, it is worth noting that CNN models possess a distinct advantage in their ability to extract features effectively. Researchers leverage this opportunity by integrating Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) models to harness the respective advantages offered by each methodology Li et al., 2019). One research attempt involved the utilization of Convolutional Neural Networks (CNN) in conjunction with Long Short-Term Memory (LSTM) models to tackle the task of news text classification. The research achieved an accuracy rate of 87%, surpassing the LSTM approach by 13% and the CNN method by 27% (Chowdhury et al., 2022). Another study also employed the CNN-LSTM model. The results of the study demonstrated an accuracy rate of 91.17%. In comparison, the LSTM model achieved an accuracy of 84.17%, while the CNN model achieved an accuracy of 88.03% (Ingole et al., 2018).

To be interpreted by computers, the data for the C-LSTM model must be transformed into a set of numbers (vectors). This data transformation technique is also known as word embedding and can be accomplished with several methods such as Word2Vec, TF-IDF, Glove, ELMo, BERT, and other programs. Furthermore, this data processing may affect the accuracy (Fauzi, 2018).

The study demonstrated that the ELMo approach had superior performance to the Word2Vec method in accuracy, with improvements of up to 10% (Maslennikova, 2019). Furthermore, a comparative analysis was conducted on Glove, ELMo, and BERT in word embedding. It was observed that both BERT and ELMo exhibited a similar accuracy rate of 82%, which demonstrated a 1% improvement in comparison to the utilization of Glove for word embedding (Pogiatzis, 2019). Besides word embedding, BERT can classify text Fields (González-Carvajal and Garrido-Merchán, 2020). In contrast, using BERT as the classification model was not employed in this study due to its reliance on a substantial volume of data to achieve satisfactory accuracy (Ezen-Can, 2020).

A study utilizing 50,000 data points to compare BERT and LSTM in addressing a movie review categorization task indicated that BERT exhibited a 4.53% reduced error rate compared to LSTM. In addition, the investigation employed a dataset including 120,000 data points to ascertain news themes. The findings revealed that BERT exhibited a 1.7% reduced error rate compared to LSTM (González-Carvajal and Garrido-Merchán, 2020).

The present study examined the impact of varying data quantities on the classification performance of BERT and LSTM models. The investigation encompassed data sizes ranging from 3750 to 15,000. In the dataset comprising 3750 instances, the Long Short-Term Memory (LSTM) model exhibited a 10% higher accuracy than the BERT model. Similarly, in the dataset consisting of 15,000 instances, the LSTM model demonstrated a 5% superior accuracy compared to the BERT model (Ezen-Can, 2020).

This research explores how the shape of the CNN structure affects the overall architecture of the BERT-C-LSTM model when building a headline news text classification model depending on its topic. The model is divided into eight classes, including "Politics", "Economy", "Culture", "Defence and Security", "Sports", "Technology", "Automotive", and "Health". This study's premise is that an architecture consisting of more than one convolutional layer constructed in parallel with varying kernel sizes delivers the highest performance compared to simpler configurations.

2. METHODS

The research process consists of three steps, as seen in Figure 1.

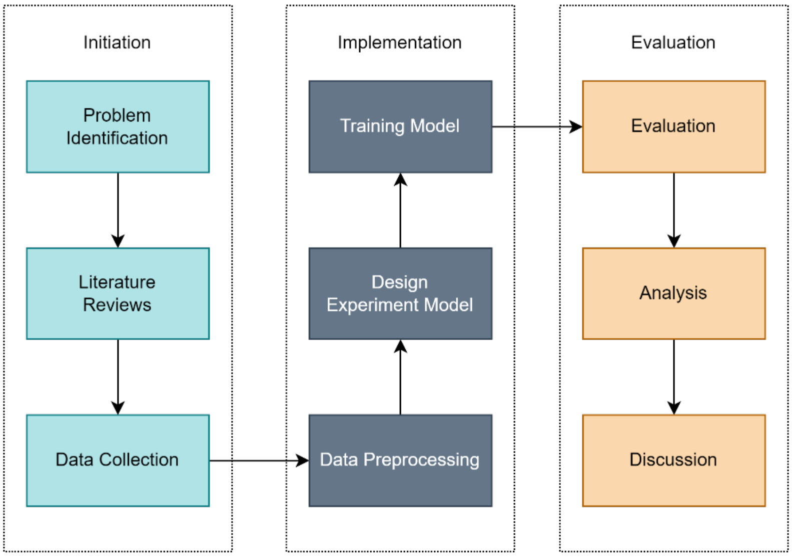

Figure 1. Research Method.

2.1. InitiationP

2.1.1. Problem Identificaiton

Problem identification is crucial in research, as it helps establish the most suitable solution and the subsequent stages. This study will construct a model that implements one of the pre-trained BERT models, IndoBERT, for word embedding and a combination of CNN and LSTM architectures for text classification. By using BERT, which possesses contextualized word embedding characteristics, the model can influence linguistic comprehension (Zhang et al., 2020). Additionally, the effect of combining CNN and LSTM with various configurations of convolution layers will be evaluated.The treatment in this station goes through several phases shown schematically **Figure 1**.

2.1.1 Literature Reviews

A. Related Work

The performance of classification models in text processing is significantly influenced by the strategies employed for extracting text into vector fields (Chen et al., 2013). One approach involves the utilization of BERT, a language representation model capable of encoding data by using the contextual and positional information of words inside sentences (Murfi et al., 2022). BERT provides a better representation in representing words as vectors and can improve accuracy compared to context-independ methods such as GloVe and Word2vec (James Mutinda et al., 2023).

Based on several previous studies, in many cases, to build a classification model with fine-tuning datasets, BERT requires a large amount of data to produce good accuracy (González-Carvajal and Garrido-Merchán, 2020). It is no better than LSTM on a small number of datasets (Ezen-Can, 2020). Consequently, numerous scholars have employed the BERT pre-trained model for word embedding, integrating it with additional deep learning methods, like the approach. For example, combine it with CNN, and the results are proven to be better than if not combined (Safaya et al., 2020) and (Kaur and Kaur, 2023).

Due to the support of a convolution layer that can perform feature extraction, CNN has an advantage in this area (Widhiyasana et al., 2021), thereby reducing the number of classification parameters. However, CNN cannot study sequential correlation (C. Zhou et al., 2015) or needs to improve in processing related data such as text (Widhiyasana et al., 2021). In overcoming this problem, CNN is usually combined with a Recurrent Neural Network (RNN) or its derivative, Long Short Term Memory (LSTM).

Therefore, some researchers combine CNN and LSTM to get better results. Research [18] shows that the CNN-LSTM architecture can have better performance (93.2%) compared to LSTM (91.9%), bi-LSTM (91.6%), and LSTM + Attention (91.8%) for short texts. As for long texts, bi-LSTM is still superior. The architecture is built sequentially, where the input vector will go through the CNN layer to extract features, and then the results will be entered into the LSTM. Reference explained the advantages of this combination of architectures, namely that CNN is primarily designed to acquire and process local information. In contrast, LSTM is specifically designed to gather and process global information. Furthermore, it is proposed to enhance the precision by improving the vector representation that serves as the input for the neural networks (H. Zhou, 2022).

B. Convolution Layers

The CNN architecture consists of two main components, namely the convolutional layer and the pooling layer. In this study, the exclusive utilization of the convolution layer is employed as the feature extractor for the contextual analysis. The convolutional layer performs a convolution operation on semantic vectors, utilizing N randomly generated filters to convolve the input data. (Kaur and Kaur, 2023).

The input layer receives the news headline text vector output by the input. The representation layer utilizes the convolutional layer to perform convolution operations, enabling the precise extraction of local sentiment information included within the news headline language. The resulting output of the convolutional layer is represented by Equation 1 (Dong et al., 2020).

$$y_{ij}^{I} = f\left(\sum_{m=1}^{M} w_{m,j}^{I} x_{m-1,j}^{I-1} + b_j^I\right) \quad (1)$$

The calculation of y_{ij}^{I} involves the utilization of the output vector x_{ij}^{I} at the input representation layer. The bias term b_j^I is applied to the feature j. The weight of the convolution kernel is denoted as w, while m represents the index value of the filter. The activation function is denoted as f. The resulting output of this layer is a novel matrix referred to as a feature maps.

The size of the input vector to the convolution layer in this study is 768 since it corresponds to the output of the word embedding step utilizing the IndoBERT pre-trained model. Previous studies have revealed that the use of filters of varying sizes can yield diverse sorts of characteristics (Wang et al., 2017). Hence, the primary objective of this work is to examine the impact of the structure and size of the kernel convolution layer on the overall architecture of the proposed model.

2.1.1 Data Collection

The data to be used in this study is data on Indonesian-language news headlines. The data to be used consists of 8 classes, with details in Table 1. Data was obtained from the data crawling process on several online news portals, including cnnindonesia.com, detik.com, tribunnews.com, suara.com, jawapos.com, sindonews.com, elshinta.com, medcom.id, viva.co.id, and suara.com.

Table 1 Indonesian Headline News Dataset

No	Topic	Amount of Data
1	Politics	3183
2	Economy	3145
3	Culture	1643
4	Defense and Security	2681
5	Sports	2988
6	Technology	2987
7	Automotive	2990
8	Health	3000
	Total	22608

Crawling data is collected on eight predetermined topics from various online news portals. The data obtained in the process of removing duplicates is processed first before entering the next process.

2.2. Implementation

Data preprocessing aims to manipulate and/or clean data from information that is not important. Consequently, the acquired data attains greater significance and has the potential to enhance the efficacy of the resultant model. The stages of data preprocessing that will be implemented in this study include deleting symbols and numbers and case folding. In the application of BERT for word embedding, many text preprocessing processes are not applied, such as removing stop words, stemming, and lemmatization, because these features play a role in providing context for each existing word.

The process of data splitting involves the partitioning of data into three distinct subsets: training data, validation data, and test data. The data is partitioned into three subsets, with proportions of 80%, 10%, and 10% respectively. The training data was utilized to construct the model, while the validation data was employed to assess the performance of the trained networks and validate the model. Finally, the testing data was utilized to evaluate the model that had been constructed.

Model training was carried out for all experimental designs to see the effect of the independent variables on model performance, which is the dependent variable of the study.

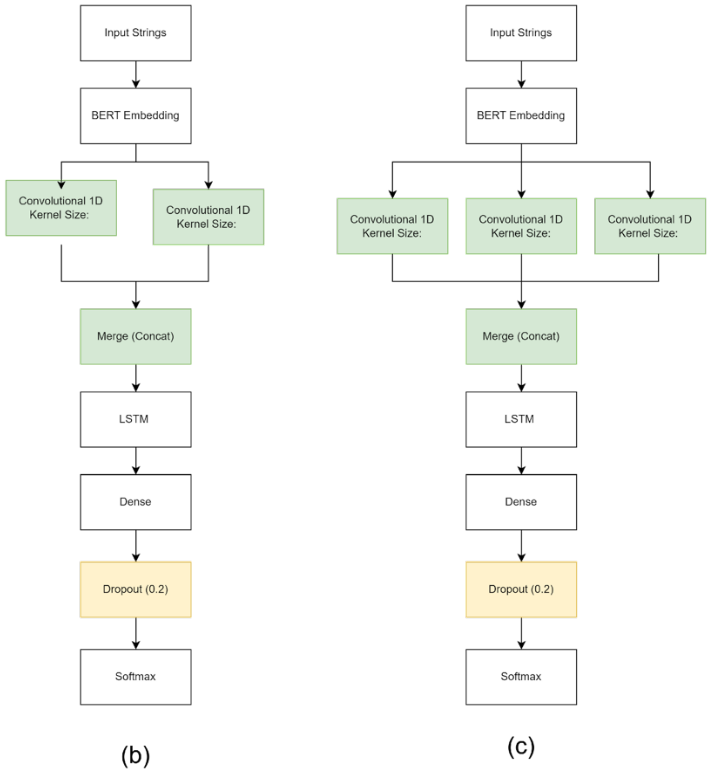

Figure 2. Three Convolution Layer Structure on BERT-C-LSTM.

The architectural framework implemented in this research comprises multiple sequential steps. The initial step involves performing word embedding by utilizing the pre-trained indoBERT (Koto et al., 2020) model on the provided input data. The outcome of the process yielded a 768-dimensional embedding vector, which then serves as the input for the convolutional layer. The topology of the convolutional layer varies based on the specific experiment. Following the convolutional layer, the input is passed through the LSTM layer, the dense layer, the dropout layer, and finally the output layer, which employs the softmax activation function. Figures 2 (a), (b), and (c) depict an architectural structure consisting of a sequential convolution layer arrangement, two parallel layers, and three parallel layers, respectively.

2.5. Evaluation

The process of model evaluation is then conducted for each model generated throughout the experiment. To assess the efficacy of the suggested model, the expected outcome was categorized into four distinct classes.

- True positive (TP) refers to the count of positive classes that have been correctly projected as positive.
- The true negative (TN) refers to the count of instances where the model correctly predicts negative classes as negative.
- False positive (FP) refers to the count of positive classes that are incorrectly projected as negative.
- False negative (FN) refers to the count of instances where negative classes are incorrectly classified as positive.

The evaluation model metrics that will be examined in this study include accuracy and F1 score. The F1 score is derived from the harmonic mean of precision and recall values. The metrics were computed based on the true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), as previously described.

Moreover, once all the tests have been conducted, the analysis will be performed, with particular attention given to examining the impact of the convolution layer configuration and the independent variable values on the performance of the IndoBERT-C-LSTM model. The present study employed the Kruskal-Wallis test to examine the presence of a statistically significant disparity in accuracy across many independent groups. Furthermore, the Kruskal-Wallis test was employed to assess the disparity in average scores between the groups.

3. RESULTS AND DISCUSSION

3.1. Experiment Setup

In the conducted tests, the seven convolution layer architectures in the overall architecture were compared, as presented in Table 2. Throughout all the scenarios, consistent parameters were employed to ensure the attainment of reliable and reproducible outcomes. The selection of hyperparameter values, including maximum length, number of neurons, learning rates, batch sizes, and epochs, was informed by prior studies in the field.

In addition, the Rectified Linear Unit (ReLU) activation function was applied to the hidden layer, specifically the convolutional layer, while the SoftMax activation function was employed on the output layer. The choice of categorical cross-entropy as the loss function was motivated by the fact that the target label classification problem involves more than two classes. To mitigate the risk of overfitting, the model was subjected to regularization by implementing a dropout value of 0.2 on the final hidden layer preceding the output layer. Table 2 displays a summary of the hyperparameter values.

Table 2 Model Hyperparameter

Parameter	**Parameter Value**
Pretrained Model	indobenchmark/indobert-base-p2
Max Length	33
Batchsize	16, 32, 64, 128
Epoch	10
Learning Rate	0.01, 0.001, 0.0001
Dropout	0.2
Conv. Layer Act. Fun.	ReLU
Output Layer Act. Func.	SoftMax
Loss Function	categorical cross-entropy
Optimization	Adam

3.2 Experimental Result

To investigate the impact of the convolution layer structure's configuration on the performance of the BERT-C-LSTM model, we compare seven deep learning architectural models, where the difference in each architecture lies in the convolution layer configuration. Configuration means the structure and the number of kernels. The findings of this comparison are presented in Table 3.

Table 3 Model Performance of Each Convolution Layer Configuration On bert-c-LSTM model

Conv. Layer Configuration	**Kernel Size**	**Acc**	**F1 Score**
Single Layer	2	89%	88,8%
Single Layer	3	88,9%	88,7%
Single Layer	4	88,6%	88,4%
Multiple Parallel (2)	2, 3	88,9%	88,8%
Multiple Parallel (2)	2, 4	88,8%	88,7%
Multiple Parallel (2)	3, 4	88,7%	88,5%
Multiple Parallel (3)	2, 3, 4	89,2%	89,1%

Table 3 shows the mean accuracy and mean f1-score obtained from each BERT-C-LSTM model with different configurations of the convolution layer. The results show that the model with three convolution layers and varying kernel sizes performs better in terms of both accuracy and f1-score compared to other configurations. However, the performance difference is not significantly large, indicating that models with more convolutional layers and varied kernel sizes cannot be considered better than other model configurations.

3.3 Kruskal-Wallis Test Result

In order to strengthen the resilience of our research findings, we utilized the Kruskal-Wallis test. The Kruskal-Wallis test was employed to strengthen the reliability of our findings by examining whether there were statistically significant differences in accuracy across the various models. Before performing the Kruskal-Wallis test, the dataset was divided into three distinct and independent groups: the BERT-C-LSTM architecture with a single convolutional layer, multiple convolutional layers arranged in parallel with two layers, and multiple convolutional layers arranged in parallel with three layers. The distribution of data is depicted in Table 4.

Table 4 Kruskal-Wallis test independent group

Group	Number of Data
Single Conv Layer	36
2 Multiple Conv Layer	36
3 Multiple Conv Layer	12

The results of the Kruskal-Wallis test in the SPSS application show the results attached in Table 5.

Table 5 Kruskal Wallis test result

Result	Number of Data
Kruskal-Wallis	0,454
Df	2
Asymptote Significance	0,797

Based on the results of the Kruskal-Wallis test, the asymptotic significance value obtained is 0.797, which is greater than (>) 0.05. This indicates that there is no significant difference among the experimental groups in terms of the architecture, namely single convolution layer, 2 multiple convolution layers, and 3 multiple convolution layers.

3.4 Comparison with BERT-LSTM and BERT-CNN

Table 6 Comparison with others

Model	Accuracy
BERT-LSTM	91,51%
BERT-CNN	90,63%
BERT-CLSTM	93,19%

The BERT-C-LSTM model achieved the best accuracy of 92.6% out of all the experimental scenarios run. This value is not significantly different from the BERT-CNN and BERT-LSTM models, which have accuracies of 92.7% and 92.1%, respectively, as shown in Table 6. The integration of convolutional layers into the BERT-LSTM architecture, denoted as BERT-C-LSTM, has the potential to enhance the model's proficiency in news topic classification. In most of the specified training parameter settings, BERT-C-LSTM demonstrates superior accuracy compared to BERT-LSTM. The incorporation of convolutional layers enhances the process of extracting spatial features from news titles, thereby collecting localized patterns and significant information within the textual content. The inclusion of convolutional layers leads to an enhancement in the overall performance of the model.

In contrast, a comparative analysis between BERT-C-LSTM and BERT-CNN reveals that the incorporation of LSTM into the architecture generally does not yield a substantial enhancement in accuracy. The limited complexity of data features and the presence of weak temporal dependency patterns contribute to the challenges in defining news themes. While

LSTM could capture sequential patterns, in the context of news topic determination, these patterns do not exert a substantial influence. Therefore, the relative importance of the LSTM and convolutional layers combination diminishes when compared to BERT-CNN in the context of this study subject.

4. CONCLUSION

In summary, the hypothesis that using diverse filter sizes improves model performance in classification is accepted. Although the Kruskal-Wallis test did not show significant differences between the data, the mean rank values indicate a slight difference, suggesting improved model performance with the use of diverse filter sizes. Based on the findings, it can be inferred that BERT-C-LSTM exhibits marginally superior performance compared to BERT-LSTM within the specific problem domain under consideration. However, in comparison to BERT-CNN, BERT-C-LSTM does not demonstrate a comparable level of effectiveness.

5. ACKNOWLEDGMENT

The Research Program receives financial assistance from the Bandung State Polytechnic through DIPA funding for the Final Project. I would like to express my gratitude to all individuals who have made valuable contributions towards the accomplishment of this initiative.

6. AUTHORS' NOTE

The authors declare that there is no conflict of interest regarding the publication of this article. Authors confirmed that the paper was free of plagiarism.

7. REFERENCES

Chen, Y., Perozzi, B., Al-Rfou, R., and Skiena, S. (2013). *The Expressive Power of Word Embeddings*. *28*.

Chowdhury, P., Eumi, E. M., Sarkar, O., and Ahamed, M. F. (2022). Bangla News Classification Using GloVe Vectorization, LSTM, and CNN. *Lecture Notes on Data Engineering and Communications Technologies*, *95*(December 2021), 723–731. https://doi.org/10.1007/978-981-16-6636-0_54

Dong, J., He, F., Guo, Y., and Zhang, H. (2020). A commodity review sentiment analysis based on BERT-CNN model. *2020 5th International Conference on Computer and Communication Systems, ICCCS 2020*, 143–147. https://doi.org/10.1109/ICCCS49078.2020.9118434

Ezen-Can, A. (2020). *A Comparison of LSTM and BERT for Small Corpus*. 1–12.

Fauzi, M. A. (2018). Automatic Complaint Classification System Using Classifier Ensembles. *Telfor Journal*, *10*(2), 123–128. https://doi.org/10.5937/telfor1802123A

González-Carvajal, S., and Garrido-Merchán, E. C. (2020). *Comparing BERT against traditional machine learning text classification*. *Ml*.

Ingole, P., Bhoir, S., and Vidhate, A. V. (2018). Hybrid Model for Text Classification. *Proceedings of the 2nd International Conference on Electronics, Communication and Aerospace Technology, ICECA 2018*, *Iaeac*, 450–458. https://doi.org/10.1109/ICECA.2018.8474920

James Mutinda, Mwangi, W., and Okeyo, G. (2023). Sentiment Analysis of Text Reviews Using Lexicon-Enhanced Bert Embedding (LeBERT) Model with Convolutional Neural Network. *Appl. Sci*, *13*(1445). https://doi.org/10.3390/app13031445

Kandhro, I. A., Jumani, S. Z., Kumar, K., Hafeez, A., and Ali, F. (2020). Roman Urdu Headline News Text Classification Using RNN, LSTM and CNN. *Advances in Data Science and Adaptive Analysis*, *12*(02), 2050008. https://doi.org/10.1142/s2424922x20500084

Kaur, K., and Kaur, P. (2023). BERT-CNN: Improving BERT for Requirements Classification using CNN. *Procedia Computer Science*, *218*(2022), 2604–2611. https://doi.org/10.1016/j.procs.2023.01.234

Koto, F., Rahimi, A., Lau, J. H., and Baldwin, T. (2020). *IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP*. 757–770. https://doi.org/10.18653/V1/2020.COLING-MAIN.66

Li, W., Gao, S., Zhou, H., Huang, Z., Zhang, K., and Li, W. (2019). The automatic text classification method based on bert and feature union. *Proceedings of the International Conference on Parallel and Distributed Systems - ICPADS*, *2019-Decem*, 774–777. https://doi.org/10.1109/ICPADS47876.2019.00114

Maslennikova, E. (2019). *ELMo Word Representation For News Protection*.

Murfi, H., Gowandi, T., Ardaneswari, G., and Nur-, S. (2022). *BERT-Based Combination of Convolutional and Recurrent Neural Network for Indonesian Sentiment Analysis*. 1–15.

Pogiatzis, A. (2019). *NLP: Contextualized word embeddings from BERT*.

Safaya, A., Abdullatif, M., and Yuret, D. (2020). KUISAIL at SemEval-2020 Task 12: BERT-CNN for Offensive Speech Identification in Social Media. *14th International Workshops on Semantic Evaluation, SemEval 2020 - Co-Located 28th International Conference on Computational Linguistics, COLING 2020, Proceedings*, 2054–2059. https://doi.org/10.18653/v1/2020.semeval-1.271

Wang, J., Wang, Z., Zhang, D., and Yan, J. (2017). Combining knowledge with deep convolutional neural networks for short text classification. *IJCAI International Joint Conference on Artificial Intelligence*, *0*, 2915–2921. https://doi.org/10.24963/ijcai.2017/406

Widhiyasana, Y., Semiawan, T., Gibran, I., Mudzakir, A., and Noor, M. R. (2021). Penerapan Convolutional Long Short-Term Memory untuk Klasifikasi Teks Berita Bahasa Indonesia (Convolutional Long Short-Term Memory Implementation for Indonesian News Classification). *Jurnal Nasional Teknik Elektro Dan Teknologi Informasi |*, *10*(4), 354–361.

Zhang, Z., Zhao, H., and Wang, R. (2020). *Machine Reading Comprehension: The Role of Contextualized Language Models and Beyond*.

Zhou, C., Sun, C., Liu, Z., and Lau, F. C. M. (2015). *A C-LSTM Neural Network for Text Classification*. *October 2017*.

Zhou, H. (2022). Research of Text Classification Based on TF-IDF and CNN-LSTM. *Journal of Physics: Conference Series*, *2171*(1). https://doi.org/10.1088/1742-6596/2171/1/012021