Tuned bidirectional encoder representations from transformers for fake news detection

Amsal Pardamean[1], Hilman F. Pardede[2]
[1,2]Graduate School of Computer Science, STMIK Nusa Mandiri, Indonesia
[2]Research Center for Informatics, Indonesian Institute of Sciences, Indonesia

Article Info

Article history:

Received Jan 20, 2021
Revised Apr 10, 2021
Accepted Apr 12, 2021

Keywords:

BERT
Fake news
Fine-tuning
Natural language processing

ABSTRACT

Online medias are currently the dominant source of Information due to not being limited by time and place, fast and wide distributions. However, inaccurate news, or often referred as fake news is a major problem in news dissemination for online medias. Inaccurate news is information that is not true, that is engineered to cover the real information and has no factual basis. Usually, inaccurate news is made in the form of news that has mass appeal and is presented in the guise of genuine and legitimate news nuances to deceive or change the reader's mind or opinion. Identification of inaccurate news from real news can be done with natural language processing (NLP) technologies. In this paper, we proposed bidirectional encoder representations from transformers (BERT) for inaccurate news identification. BERT is a language model based on deep learning technologies and it has found effective for many NLP tasks. In this study, we use transfer learning and fine-tuning to adapt BERT for inaccurate news identification. The experiments show that our method could achieve accuracy of 99.23%, recall 99.46%, precision 98.86%, and F-Score of 99.15%. It is largely better than traditional method for the same tasks.

Corresponding Author:

Amsal Pardamean
STMIK Nusa Mandiri
Graduate School of Computer Science
Jakarta, Indonesia
Email: amsalpardamean@gmail.com

1. INTRODUCTION

Online medias are currently the dominant source of information due to not being limited by time and place, fast and wide distributions [1], [2]. Inaccurate news or often referred as fake news is a major problem in news dissemination for online medias today. Inaccurate news is information that is not true that is engineered to cover actual information. It has no factual basis which is made in the form of news that has mass appeal and is presented in the guise of genuine and legitimate news nuances to deceive or change the reader's mind. Inaccurate news is dangerous because the writings or contents on the news is disseminated by irresponsible sources. Inaccurate news could change the mass opinions and hence affecting the society on making wrong decisions or actions that can later harm individuals or other groups [3]-[6]. Inaccurate news also more often appears on political elections to seek sympathy and increase the number of votes even to attack political opponents [7].

Therefore, it is important to have tools to detect inaccurate news. By doing so, the news that is consumed by the public can be verified for truth, and hence reduce the expansion of inaccurate news and impose penalties on the creators or sites that widen and initiate hoaxes [3]. Natural language processing (NLP) technologies, a part of artificial intelligence which centralize learning on natural language processing or human language to communicate [8], [9] could be implemented for this purpose.

Fake news detection is a quite active studies in NLP and several studies have explored algorithms in solving inaccurate news detection. In [10], term frequency-inverse document frequency of bi-grams and probabilistic context-free grammar is used as features and Support vector machines, stochastic gradient descent, gradient boosting, bounded decision trees, and random forests are used as classifiers. The study found that term frequency-inverse document frequency (TF-IDF) and the stochastic gradient achieved accuracy of 77.2%. In [7], naïve bayes (NB) is compared with hybrid convolution neural network and recurrent neural network models. The comparison results state that research using the deep learning model gets 82% accuracy. Various features such as count vectors, TF-IDF by applying word level, N-gram level and character level, and word embedding are evaluated in [6] for inaccurate news detection in Twitter. The results show that support vector machine (SVM) has an accuracy of 89.34% with TF-IDF and Word2Vec. feed-forward (FF) and back-propagation (BP), neural networks classification algorithms are used in [4]. The study achieved of 78.76%. In [5], TF-IDF with bag of world and n-gram are used as feature extraction and deep multi-layer perceptron (MLP) are used as classification on 600 datasets of inaccurate news in Indonesia. The study achieved 83% accuracy, precision, 84%, recall 0.73, and F1-score 78%. In [3], count vectorizer, TF-IDF vectorizer, and word embedding followed are usd with SVM, logistic regression, decision tree, random forest, XG-Boost, gradient boosting, and deep learning neural networks. The best results are achieved for TF-IDF with SVM with 94% accuracy results.

One recent technology in NLP called bidirectional encoder representations from transformers (BERT) [11]. BERT introduces a two-ways training for a transformer, a deep learning architectures that fits to deal with sequence data. BERT has been used in various NLP problems such as the stanford sentiment treebank (SST) sentiment classification [12], the classification of people's livelihood texts [13], relation extraction in chinese medical texts [14] and website category classification [15]. These studies shows that the implementation of BERT achieve most recent results to solve many problems in text classification. In this study, we propose to use BERT with fine tuning for classification of inaccurate news.

2. PROPOSE METHOD

Bidirectional encoder representation of transformers (BERT) is a proposed model to perform natural language processing (NLP) tested with bidirectional representation and is based on pre-training neural network techniques [11], [16]. BERT is determined from the transformer methodology and uses the attention mechanism. Attention mechanism is obtained by adopting encoder-decoder architectures to transformers architectures with the aim to find the "summary" of the data with encoder and then the decoder translate them. In this sense, the transformers try to find mindfulness and find a way of looking at the relationships between words in a particular sentence [17], [18]. Currently, BERT is one of the most powerful representations of context and word.

BERT is designed to be able to distinguish a process that has a different meaning. BERT uses unlabeled in designing deep bidirectional representations by moving the context left and right across all layers [11], [19]. The results of the previously trained BERT model can then be tuned with one output layer to create a model for performing various tasks including answering questions, language inference without substantial modification of the task-specific architecture. BERT has the latest results on eleven problem solving tasks in natural language. the BERT model is conceptually simple and empirically strong [20]-[22].

In this study, we use pre-training BERT and then fine tune it for inaccurate news detection. The pre-training modeling is carried out on data that does not have a label and fine-tuning initializes the parameters for training and in fine-tuning modeling the data used are labeled data [11], [17]. The model of BERT is shown in Figure 1.

First, we adopt google BERT pre-training model. The encoder layer of BERT consists of token embedding layer, segment embedding layer and position embedding layer [17], [23]. BERT uses transformer, an attention to the mechanism in studying the relationship between the meaning or context contained in the text between one word and another or in sub words. Transformer consists of two different mechanisms which are predictive encoders for tasks [24], [25]. BERT transformer encoder reads the entire word to build a model in learning to understand the meaning of words from the words in the surroundings. We retrain the BERT model to our inaccurate news task. By doing Pre-trained and fine tune we could adapt BERT model only using a small corpus for inaccurate news detection [12], [17]. The representation of BERT is shown in Figure 2.

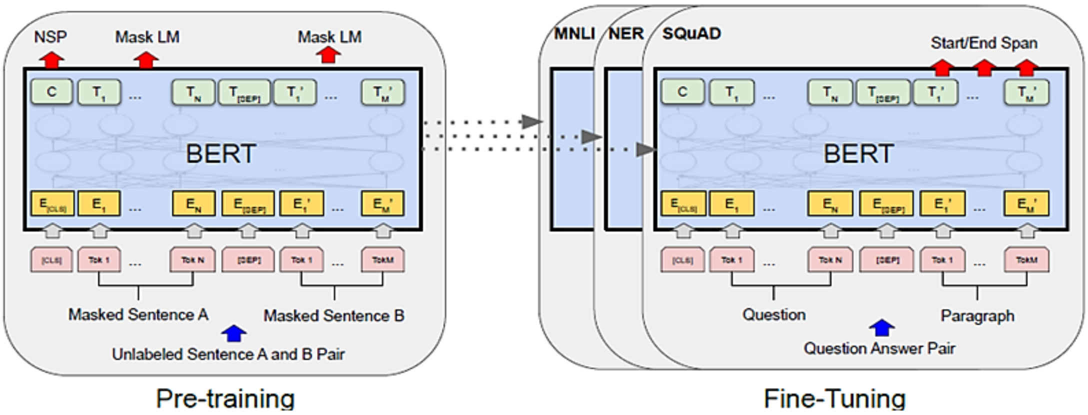

Figure 1. BERT model [11]

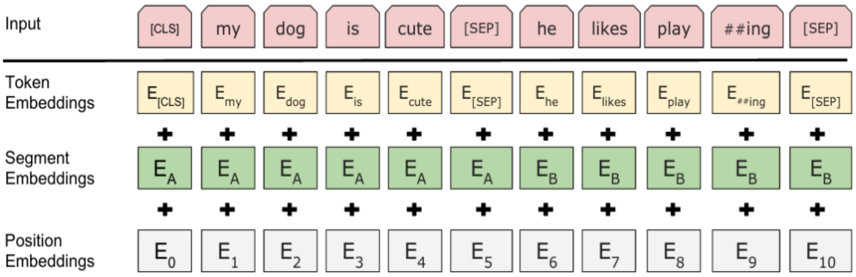

Figure 2. BERT representation [11]

3. RESEARCH METHOD

We evaluate our method on inaccurate news data taken from the kaggle website. The dataset has 28,711 news data which comprise of 12,999 inaccurate news and 15,712 true news. The data set is news from 100percentfedup, 21stcenturywire, abcnews, abeldanger, abovetopsecret, activistpost, addictinginfo, blacklistednews, collective-evolution, counterpunch, dailywire, New York Times, Cable News Network (CNN), Atlantic, Fox News, National Review, Guardian, Reuters, Washington Post and Vox [26].

We divide the data into 2 parts: training data and testing data randomly with ration of 80% training data and 20% testing data. The next stage is data are passed through the BERT Tokenizer process. Hyperparameter settings in data pre-processing are 350 words for each column and a maximum of 35,000 features. BERT Tokenizer specifically for pre-processing BERT model data and all vocabulary is available. We apply learning rate 2e-5 and epoch value=3. This is based on our empirical observations that only small number of epoch is required to re-train the BERT model. We also apply Naive Bayes-SVM (NBSVM) as reference methods in the study.

4. RESULTS AND DISCUSSION

The comparisons of BERT and NBSVM is shown in Figure 3. It is clear that BERT achieves better results than NBSVM, confirming the effectiveness of our method. Need to be noted that the results is obtained when we use only 3 epochs for BERT while we need 25 epochs for NBSVM. In this study the results of the Performance Measure on BERT fine-tuning in the form of accuracy of 99.23%, recall 99.46%, precision 98.86%, and F1-score of 99.15% are results that have been proven to provide good performance.

The progression of BERT fine-tuning from epoch 1-3 is shown in Figure 4. The bar chart shows an increase in the performance measure which is getting higher with an increase in the epoch value. For the epoch value of 1, the method we propose gets a higher value with the best value on the NB-SVM method. Then the increase in epoch 2 gives a quite different difference and finally with the epoch 3 value the increase still occurs but is not so different from the epoch 2 value.

The progression of NBSVM from epoch 3-25 is shown in Figure 5. The bar chart shows an increase in performance measure when the epoch value is further increased in the NBSVM model. However, it can be seen after epoch 22, the progression of the performance is stable, indicating that the best performance is achieved when epoch is around 20. Even with more epochs, NB-SVM is still worse than our method.

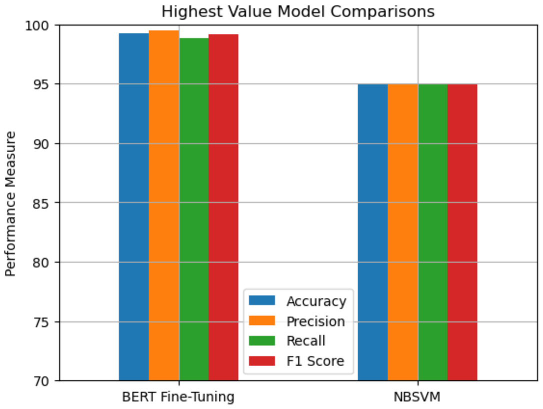

Figure 3. Comparisons of BERT and NBSVM

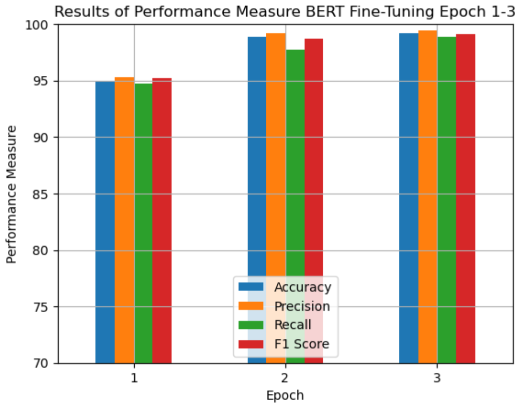

Figure 4. Performance measure BERT fine-tuning epoch 1-3

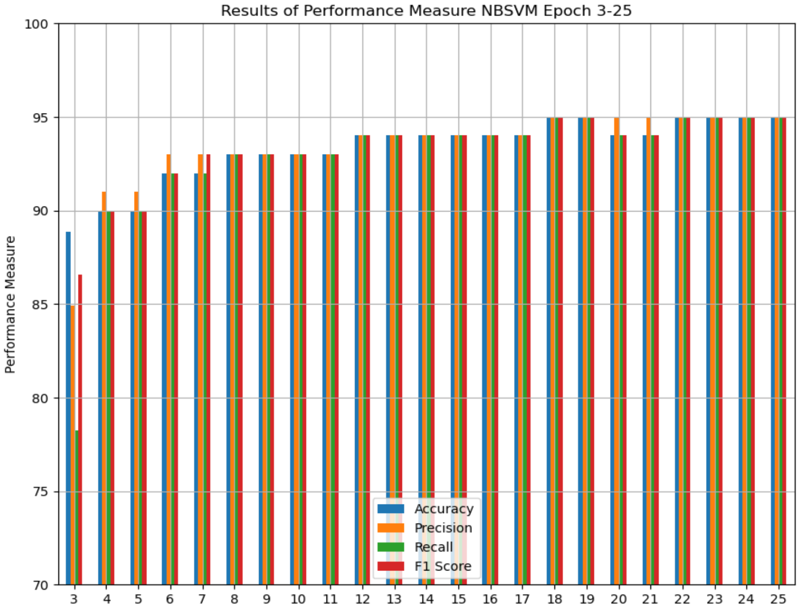

Figure 5. Performance measure NBSVM epoch 3-25

5. RESULTS AND DISCUSSION

In this paper, we propose pre-trained BERT and fine tuning for inaccurate news detection. We compare it with NBSVM. Our experiments confirm that our method achieve better performance when with smaller number of epoch. However, need to be noted that BERT is highly computational models and efforts to reduce computational load of BERT is needed.

In the future, we plan to use combination models (hybrid) to further improve the results of the performance measure accuracy, precision, recall and f-score on the detection of inaccurate news. Combinations of features and the use of feature learning as inputs for BERT is also our future plan. We also plan to use BERT for indonesian inaccurate news detection. It is also interesting to see how BERT perform for other natural language processing problems such as spam detection in SMS or e-mail, text classification, sentiment analysis, detection of emotions from text for Indonesian data.

ACKNOWLEDGEMENTS

This Research is supported by department computer science of STMIK Nusa Mandiri Jakarta.

REFERENCES

[1] A. Fronzetti Colladon, “Forecasting election results by studying brand importance in online news,” *International Journal of Forecasting*, vol. 36, no. 2, pp. 414–427, 2020, doi: 10.1016/j.ijforecast.2019.05.013.

[2] X. Zhang and A. A. Ghorbani, “An overview of online fake news: Characterization, detection, and discussion,” *Information Processing and Management*, vol. 57, no. 2, p. 102025, 2020, doi: 10.1016/j.ipm.2019.03.004.

[3] N. Smitha, “Performance Comparison of Machine Learning Classifiers for Fake News Detection,” in *2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA)*, pp. 696–700, 2020, doi: 10.1109/ICIRCA48905.2020.9183072.

[4] C. W. Kencana, E. B. Setiawan, and I. Kurniawan, “Hoax Detection System on Twitter using Feed-Forward and Back-Propagation Neural Networks Classification Method,” *RESTI*, vol. 4, no. 4, pp. 655-663, Aug. 2020, doi: 10.29207/resti.v4i4.2038.

[5] A. Rusli, J. C. Young, and N. M. S. Iswari, “Identifying fake news in indonesian via supervised binary text classification,” *2020 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT)*, 2020, pp. 86-90, doi: 10.1109/IAICT50021.2020.9172020.

[6] A. A.-Tanvir, E. M. Mahir, S. Akhter, and M. R. Huq, “Detecting Fake News using Machine Learning and Deep Learning Algorithms*,” 2019 7th International Conference on Smart Computing and Communications (ICSCC)*, 2019, pp. 1-5, doi: 10.1109/ICSCC.2019.8843612.

[7] W. Han and V. Mehta, “Fake news detection in social networks using machine learning and deep learning: Performance evaluation,” *2019 IEEE International Conference on Industrial Internet (ICII)*, 2019, pp. 375–380, doi: 10.1109/ICII.2019.00070.

[8] L. Zhao *et al.*, “Natural Language Processing (NLP) for requirements engineering: A systematic mapping study,” *arXiv*, no. v, 2020.

[9] A. Ly, B. Uthayasooriyar, and T. Wang, “A survey on natural language processing (nlp) and applications in insurance,” *arXiv*, pp. 1-34, 2020.

[10] S. Gilda, “Evaluating machine learning algorithms for fake news detection,” *IEEE Student Conference on Research and Development: Inspiring Technology for Humanity (SCOReD)*, vol. 2018, 2018, pp. 110–115, doi: 10.1109/SCORED.2017.8305411.

[11] 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, 2019, pp. 4171–4186.

[12] M. Munikar, S. Shakya and A. Shrestha, “Fine-grained sentiment classification using BERT,” *arXiv*, vol. 1, pp. 1–5, 2019.

[13] S. Liu, H. Tao, and S. Feng, “Text Classification Research Based on Bert Model and Bayesian Network,” *2019 Chinese Automation Congress (CAC)*, 2019, pp. 5842–5846, 2019, doi: 10.1109/CAC48633.2019.8996183.

[14] K. Xue, Y. Zhou, Z. Ma, T. Ruan, H. Zhang, and P. He, “Fine-tuning BERT for joint entity and relation extraction in chinese medical text,” *arXiv*, pp. 892–897, 2019.

[15] F. Demirkıran, A. Çayır, U. Ünal, and H. Dağ., “Website Category Classification Using Fine-tuned BERT Language Model,” *2020 5th International Conference on Computer Science and Engineering (UBMK)*, 2020, pp. 333-336, doi: 10.1109/UBMK50275.2020.9219384.

[16] T. Wang, K. Lu, K. P. Chow, and Q. Zhu, “COVID-19 Sensing: Negative Sentiment Analysis on Social Media in China via BERT Model,” *IEEE Access*, vol. 8, pp. 138162-138169, 2020, doi: 10.1109/ACCESS.2020.3012595.

[17] A. G. D’Sa, I. Illina, and D. Fohr, “BERT and fastText Embeddings for Automatic Detection of Toxic Speech,” *2020 International Multi-Conference on: Organization of Knowledge and Advanced Technologies (OCTA)*, 2020, pp. 1-5, doi: 10.1109/OCTA49274.2020.9151853.

[18] Cai, Ren *et al.*, “Sentiment Analysis About Investors and Consumers in Energy Market Based on BERT-BiLSTM,” *IEEE Access*, vol. 8, pp. 171408-171415, doi: 10.1109/ACCESS.2020.3024750.

[19] Dong, Junchao, F. He, Y. Guo, and H. Zhang, “A Commodity Review Sentiment Analysis Based on BERT-CNN Model,” *2020 5th International Conference on Computer and Communication Systems (ICCCS)*, pp. 143–147, 2020, doi: 10.1109/ICCCS49078.2020.9118434.

[20] W. Li, S. Gao, H. Zhou, Z. Huang, K. Zhang, and W. Li., “The automatic text classification method based on bert and feature union,” *2019 IEEE 25th International Conference on Parallel and Distributed Systems (ICPADS)*, 2019, pp. 774–777, doi: 10.1109/ICPADS47876.2019.00114.

[21] C. J. Lin, C. H. Huang, and C. H. Wu., “Using BERT to process chinese ellipsis and coreference in clinic dialogues,” *2019 IEEE 20th International Conference on Information Reuse and Integration for Data Science (IRI)*, 2019, pp. 414–418, doi: 10.1109/IRI.2019.00070.

[22] I. Annamoradnejad, M. Fazli, and J. Habibi, “Predicting Subjective Features from Questions on QA Websites using BERT,” *arXiv*, pp. 240–244, 2020.

[23] J. Yadav, D. Kumar, and D. Chauhan., “Cyberbullying Detection using Pre-Trained BERT Model,” *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*, 2020, pp. 1096–1100, doi: 10.1109/ICESC48915.2020.9155700.

[24] W. Maharani., “Sentiment Analysis during Jakarta Flood for Emergency Responses and Situational Awareness in Disaster Management using BERT,” *2020 8th International Conference on Information and Communication Technology (ICoICT)*, 2020, doi: 10.1109/ICoICT49345.2020.9166407.

[25] Gao, Zhengjie, Ao Feng, Xinyu Song, and Xi Wu., “Target-Dependent Sentiment Classification with BERT,” *IEEE Access*, vol. 7, pp. 154290 – 154299, doi: 10.1109/ACCESS.2019.2946594 .

[26] Kaggle, Fake News Classifier - Final Project, 2018. [Online]. Availbale: https://www.kaggle.com/anthonyc1/fake-news-classifier-final-project.