Objectivity and Subjectivity Classification with BERT for Bahasa Melayu

Wing Kin Chong[1], Hu Ng[1(✉)], Timothy Tzen Vun Yap[1], Wooi King Soo[1], Vik Tor Goh[2], and Dong Theng Cher[3]

[1] Faculty of Computing and Informatics, Multimedia University, Cyberjaya, Malaysia
nghu@mmu.edu.my
[2] Faculty of Engineering, Multimedia University, Cyberjaya, Malaysia
[3] SIRIM Berhad, Shah Alam, Malaysia

Abstract. This research present the notion of subjectivity and objectivity in Bahasa Melayu language. Word2Vec and BERT word embedding models are created for the purpose of subjectivity classification and sentiment classification. Two types of embeddings are developed (Word2Vec and BERT) with Wikipedia data as objectivity dataset, Twitter data as subjectivity dataset and combination of both datasets. A pre-trained BERT embedding model called Bert-Base-Bahasa-Cased is used as a reference. First, the datasets are fed into every embedding model to be embedded as vectors. The subjectivity classification and sentiment classification are carried out via 70:30 train-test splits. Both classification tasks are carried out using Logistic Regression, Random Forest, and Double Layer Neural Network classifiers. Logistic Regression on Bert-Base-Bahasa-Cased model achieved the highest result of 99.95% in subjectivity classification and 74.30% in sentiment classification.

Keywords: Objectivity · Word2Vec · Subjectivity classification · BERT · Sentiment classification

1 Introduction

Natural language is a way to communicate with each other. Language is a general, abstract aspect and a sum of organisation skills and principles. It is the system that governs any concrete act of communication [1]. Natural language can be expressed in many forms such as text, speech, visual language, light signals, smoke signals and programming language. In this era of big data, language processing is needed to perform a variety of tasks. Web search engine likes Google, uses language processing to process a large scale of queries. Furthermore, spelling and grammar checking system are also examples of application in NLP.

According to Wikipedia [2], Natural language processing (NLP) is a sub-field of linguistics, computer science, and artificial intelligence concerned with the interactions between computers and human language. The goal of NLP is to let computer to'understand' the contextual nuances and thus extracting useful information and even communicate with human.

S.-C. Haw and K. Sonai Muthu (Eds.): CITIC 2022, 10, pp. 246–257, 2022.
https://doi.org/10.2991/978-94-6463-094-7_20

Researchers nowadays have shown great interest in Natural Language Processing (NLP). One of the reasons is because of machine learning and deep learning techniques are act as tools in NLP. An example of machine learning application in NLP is text classification [3]. Text classification is a process of assigning categories to a text depending on its content.

Bahasa Melayu (BM), called Malay language in English, is an Austronesian language spoken in Indonesia, Brunei, Malaysia and Singapore. BM is spoken by 290 million people (around 260 million in Indonesia alone in its own literary standard named "Indonesian") across the Malay world. It is also the Bahasa Kebangsaan ("national language") of Malaysia and Indonesia. In Malaysia, it is designated as Bahasa Malaysia ("Malaysian language"), where there are a significant number of users who are using Malay to express their opinions on social media. However, due to low resource of BM comparing to the other languages such as English, very limited research has been attributed to NLP in BM.

The objectives of this research paper are to benchmark and compare classical word embedding method called Word2Vec and the state-of-the-art embedding model called Bidirectional Encoder Representations from Transformers (BERT) on BM. By training with various classification models on subjectivity classification and sentiment classification using word vectors embedded by Word2Vec and BERT embedding methods. The performance of the two embedded methods is evaluated and analysed.

2 Literature Review

Word embedding or word vector, is the representation of word in vector. Words that have similar meaning are encoded closer in the vector space. There are generally two types of word embedding, the contextual and non-contextual. Contextual embedding assigns each word in a sentence or a vector representation of its context where it considers the consequences of the ordering of the words while non-contextual embedding does not require this.

Word2Vec by Mikolov et al. [4] is an example of non-contextual word embedding. Before Word2Vec was introduced, researchers [5–7] were using Neural Net Language Model (NNLM) for word embedding. NNLM uses the concept of neural network, which consists of an input layer, linear projection layer, hidden layer, and the output layer. Although those researchers managed to obtain a good result by applying NNLM, the computation is very dense and complex due to the hidden layer are computed by each of every input and transfer to projection layer. Word2Vec overcomes the problem of NLM and improves the performance of word embedding architectures by removing the hidden layer. It connects the input layer with the projection layer so that the input can share all the weight. Word2Vec also introduces two new model which are Skip-gram model architectures and Continuous Bag of Words (CBOW) model architectures. The high efficiency of Word2Vec in large-scale corpus and shorter time consume have make it become a very popular model in word embedding architectures.

In the pass, researchers were applying deep learning model such as Markov Decision Processes (MDPs) [8], Recurrent Neural Networks (RNNs) [9], Convolutional Neural Network (CNN) [9] and Long Short-term Memory (LSTM) [10] to perform NLP tasks.

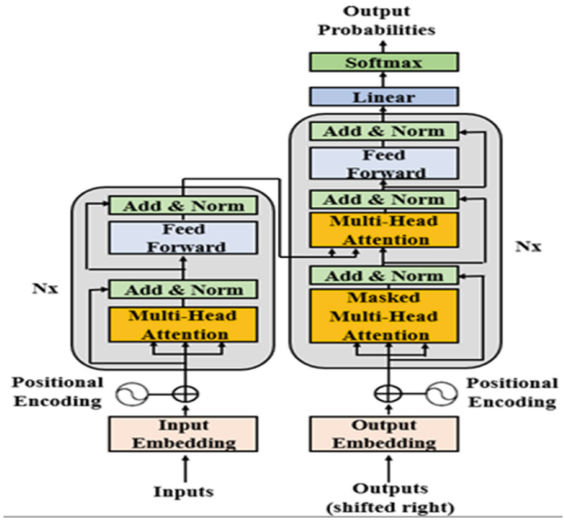

Fig. 1. The architecture of transformer [12]

Those were the existing state-of-the-art NLP model and rely on the recurrent architecture, but they seem to be relatively old and outdated. There is no huge improvement and innovation over thirty years. Those models also come at a tremendous cost in terms of calculations and machine power compared to the big data in this era. In 2017, Vaswani et al. [11] introduced the concept of the original transformer model which uses an encoder-decoder architecture. Figure 1 shows the architecture of transformer. The transformer consists of 6-layers encoder stack on the left and another 6-layers decoder stack.

Bidirectional Encoder Representation from Transformer (BERT) is the state-of-the-art embedding model published by Devlin et al. [13] in 2019. It is a contextual embedding model that achieve a breakthrough in the field of NLP by providing better accuracy results in many NLP tasks. BERT is based on the transformer model with encoder. The term Bidirectional means that it can read a sentence in both directions. It returns the representation for each word in the sentence when sentence is feed as input to the encoder. It can differentiate same word with different context. The encoder understands the context of each word by using the multi-head attention mechanism. It relates each word in the sentence to all the other words in the sentence, so that it can learn the relationship and contextual meaning of the words. It can be stacked up to N number by referring to the size of the encoder layer.

Two supervised machine learning models and one deep learning model are considered in this research to perform classification. The two machine learning models are Logistic Regression (LR) [14] and Random Forest (RF) [15]. The deep learning model, namely Double Layer Neural Network (DNN) is only applied to vectors embedded using BERT since it can directly attach a classification neural network layer. LR uses a logistic function to model a binary dependent variable and produces a logistic curve which limit the outcome values between 0 and 1. By setting a threshold, the LR model can classify

the data into their corresponding classes after calculating obtained estimated probability and compared with the threshold.

RF is a machine learning technique that is used to solve regression and classification problems. It is a large group of decision trees which utilizes the ensemble learning which is a technique that combines many decision tree classifiers where each classifier will provide a class prediction and the class with the most votes from the decision trees will be selected as the model prediction. This approach can eradicate the limitations of a single decision tree algorithm.

3 Methodology

There are three stages in this research project. The first stage begins with training different word embedding models, which are Word2Vec and BERT. Second stage is performing subjectivity classification using the trained embedding models. The final stage is to perform sentiment analysis.

3.1 Datasets

To train embedding models and subjectivity classification models, a huge number of datasets are needed. Two large corpora with objectivity and subjectivity representation are chosen to train embedding models. The first one is Wikipedia dataset [16] to represent objectivity dataset. Wikipedia dataset consists of 1.2 million rows of sentence. Second is the Twitter dataset [17] which represent subjectivity dataset. Twitter dataset consists of 3 million rows of tweets. Both datasets are in Bahasa Melayu language and are large enough. For sentiment analysis, Malay Twitter dataset from Malaya [18] is chosen as it has been labeled (positive and negative) by the contributor.

According to Wikipedia Policy, all Wikipedia content must be written from a neutral point of view (NPOV), which means that the content must always be fair. Editors who write and upload the contents in Wikipedia must comply with these rules. The Twitter dataset is selected because it contains tweets and comments from users which are non-neutral point of view and is judgmental.

In this research, an assumption has been made where the sentences are purely factual or non-factual in the Wikipedia dataset and Twitter dataset. The datasets are gone through the pre-processing stages before converting into the word embedding model. An overview of the dataset is shown in Fig. 2.

3.1.1 Data Acquisition

All the datasets use in this research is downloaded from Malaya dataset GitHub repository [16–18]. There are three dataset available, which are the Wikipedia dataset [16], Twitter dataset [17] and Malay Twitter dataset [18]. The Wikipedia [16] and Twitter dataset [17] are unlabeled dataset and only contain sentences, while the Malay Twitter dataset [18] is pre-labeled with positive and negative.

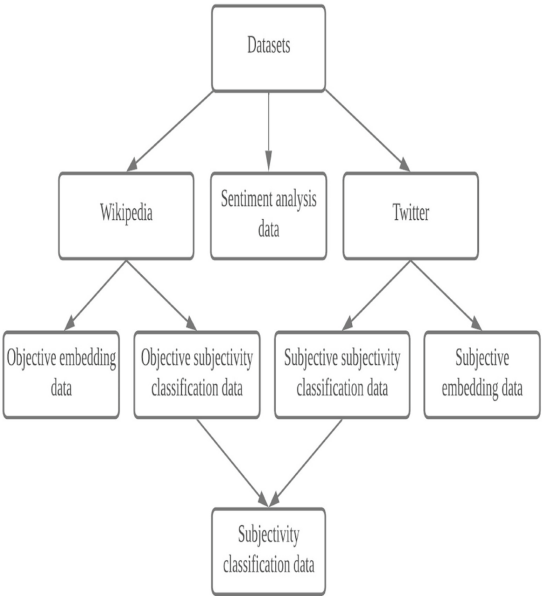

Fig. 2. Overview of the dataset

3.1.2 Data Labelling

Wikipedia dataset and Twitter dataset are originally unlabeled, thus manual labelling is required for the later subjectivity classification purpose. In nature, all the Wikipedia data is in neutral point of view, which means that all the sentences are labeled as objectivity. While for Twitter dataset, all sentences are labeled subjectivity.

3.1.3 Data Integration

One million of sentences are sampled out without replacement from each of Wikipedia and Twitter datasets. Another 200k of sentences from each of Wikipedia and Twitter dataset are sampled out and being merged to become the subjectivity classification dataset.

3.1.4 Data Cleaning

The sentences in all the dataset are gone through a cleaning pipeline to improve the quality of the data. The cleaning pipeline includes four processes as listed below with examples:

(a) Replace special words into tokens of tags

(i) "rm10k" becomes " <money>",
(ii) "#drmahathir" becomes "<hashtag>drmahathir </hashtag>"

(b) Translate English word to Bahasa Melayu words.

(i) "happy" become "gembira"

(c) Lowering cases for uppercase text

(i) "MAKAN" becomes "makan"
(ii) "Makan" becomes "makan"

(d) Stemming and lemmatization of words

(i) "menarik, "menarikan" become "tarik"
(ii) "menyeru, menyerukanlah" become "\seru"

3.2 Word Embedding Model

Two types of word embedding method are considered in this research, which are Word2Vec and BERT. Three embedding models are built by applying Word2Vec and BERT. The first embedding model (objectivity) is trained using objectivity embedding dataset (Wikipedia), second embedding model (subjectivity) is trained using subjective embedding dataset (Twitter) and the third embedding model (Combine) is trained using the combination of both objective and subjective embedding dataset. The purpose of merged embedding dataset is to check whether objective or subjective sentences would affect the result of embedding.

To train a BERT embedding model, a BERT config is needed to set the parameter. In this research, the default BERT config is used. Another pre-trained BERT model, called Bert-Base-Bahasa-Cased [19] is used as a reference to the own trained BERT model. Figure 3 shows an overview for the embedding models.

3.3 Subjectivity Classification

After the word embedding models have been trained, the subjectivity classification dataset is fed into the models to get the vector form of the sentence. This subjectivity classification contains 400k row of sentences where 200k rows are objectivity, and another 200k row are subjectivity. The embedded word vector is split for training and testing data using 70:30 ratio respectively.

The training data is utilized to train three classification models and the testing data will be used to evaluate the models trained by comparing the accuracy. The classifier used are LR, RF, and DNN. Due to the time constraint, The DNN model is only applied to word embedding using BERT and executed under the default parameter.

3.4 Sentiment Classification

Malay Twitter dataset, which contains 500k row of sentences where 250k are labelled positive, and another 250k are labelled negative. Malay Twitter dataset is fed into the

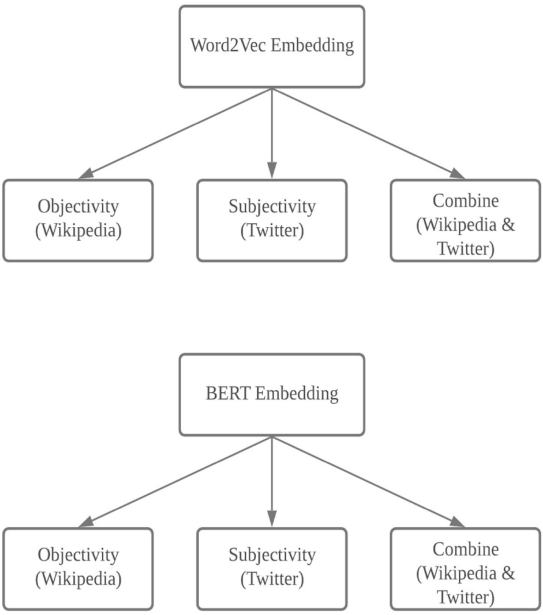

Fig. 3. Overview of the embedding model

embedding model with the highest accuracy based on previous subjectivity classification accuracy from Word2Vec and BERT.

After the sentences has been embedded, it is split into train and test data using 70:30 ratio respectively. The same classifier from subjectivity classification is used on sentiment classification. Figure 4 shows the flowchart of classification using Word2Vec as embedding method. Figure 5 shows the flowchart of classification using BERT as embedding method.

4 Results and Discussion

4.1 Subjectivity Classification

This research paper has implemented two supervised machine learning and one deep learning techniques which are LR, RF and DNN. Three embedding models for each Word2Vec and BERT have been trained. The accuracy from both the classification is benchmarked to compare Word2Vec and BERT word embedding method. This research also intends to find out if there is a significant on the type of corpus used to build a word embedding model. Table 1. Shows the accuracy result from subjectivity classification.

4.2 Sentiment Classification

Sentiment classification is carried out the pre-trained Bert-Base-Bahasa-Cased [19] and the two trained embedding models which are Word2Vec and BERT. Only the combined

Table 1. Subjectivity classification accuracy result

Embedding		Classifier		
		LR	RF	DNN
Word2Vec	Objectivity	94.53	94.32	–
	Subjectivity	96.62	96.47	–
	Combined	96.32	95.84	–
BERT	Objectivity	90.24	87.86	74.22
	Subjectivity	89.11	87.18	73.85
	Combined	89.59	87.31	74.25
Bert-Base-Bahasa-Cased [19]		99.95	99.90	96.72

Table 2. Sentiment classification accuracy result

Model	Classifier		
	LR	RF	DNN
Word2Vec (combined)	73.25	71.78	-
BERT (combined)	71.44	68.85	63.62
Bert-Base-Bahasa-Cased [19]	74.30	72.67	71.28

embedding model are used to embed the sentences in this classification, This is due to the observation from Table 1, where the single type of embedding model (Word2Vec or BERT) has less impact to the result of classification. Table 2. Shows the accuracy result of sentiment classification.

From Table 2, it can be observed that the accuracy achieved by Word2Vec on both LR, and RF are higher than DNN. The pre-train Bert-Base-Bahasa-Cased [19] achieved the highest accuracy in all three classifiers among the three embedding models. While Word2Vec achieved better result than BERT.

4.3 Discussion

From both of the experiments, it can be observed that the BERT embedding model has a lower accuracy than Word2Vec embedding model. This is due to insufficient amount of data used for training, As BERT is a model that can require a very large sample size for training [20]. Another possible factor is because BERT is originally trained in English language dataset, so more fine-tuning on the BERT config might be considered to improve the model.

Although Word2Vec may seem perform better than BERT based on the result, the pretrained model, Bert-Base-Bahasa-Cased [19] has the highest accuracy among all

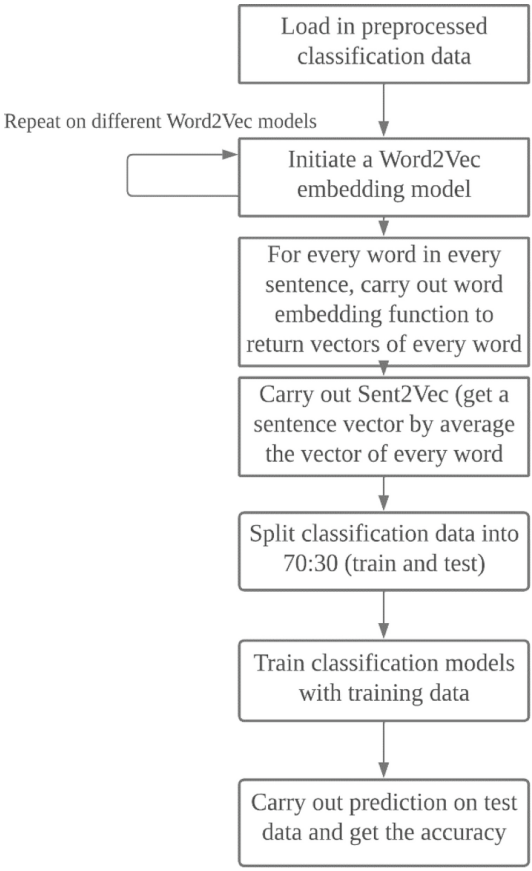

Fig. 4. Flowchart for classification using Word2Vec as embedding method

the embeddings. That concludes that a well-trained BERT embedding is better than Word2Vec. One of main reason is the sample size of data (1 million rows of data) used in this research for training the BERT embedding model is not large enough, as comparing with the pretrained model that has been trained on a very large corpus (more than 5 millions).

5 Conclusion and Future Work

In conclusion, this research paper had carried out Word2Vec and BERT word embedding model training. The accuracy result of subjectivity classification and sentiment classification are used to benchmark and compare BERT embedding method with classical Word embedding method. Based on the result, it can be found that a well-trained BERT embedding model can yield a better result for natural language processing task such as classification. Apart from that, this research also proven that objectivity, subjectivity, and the combination of both do not significantly affect the result of an embedding model.

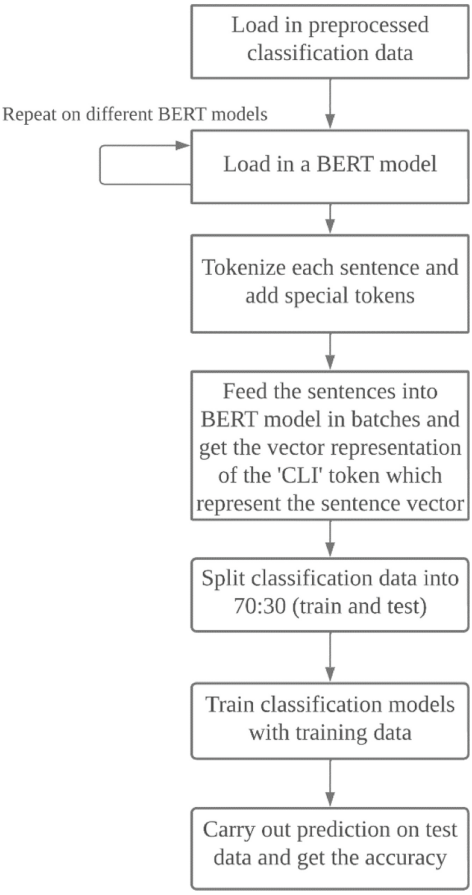

Fig. 5. Flowchart for classification using BERT as embedding method

The future work for this research is to train the BERT embedding using more resources and time with more datasets in Bahasa Melayu. Other than that, more classifiers can be applied in the classification process.

Acknowledgments. The authors would like to acknowledge staff and friends from Multimedia University for providing support in carrying out this research.

Authors' Contributions. Hu Ng, roles: Corresponding author, Conceptualization, Project Administration, Supervision, Validation, Visualization, Writing—Review & Editing.

Wing Kin Chong, roles: Data Curation, Formal Analysis, Investigation, Resources, Writing—Original Draft Preparation.

Timothy Tzen Vun Yap, roles: Conceptualization, Investigation, Administration, Supervision, Validation, Writing—Review & Editing.

Wooi King Soo, roles: Conceptualization, Validation, Visualization, Writing—Review & Editing.

Vik Tor Goh, roles: Conceptualization, Validation, Visualization, Writing—Review & Editing.
Dong Theng Cher, roles: Conceptualization, Validation, Visualization, Writing—Review & Editing.

References

1. A. Sirbu, The significance of language as a tool of communication, 05 2015.
2. Wikipedia contributors, Natural language processing — Wikipedia, the free encyclopedia," 2022, [Online; accessed 15-October-2021]. https://en.wikipedia.org/w/index.php/title=Naturallanguage processing&oldid=1081938932.
3. A. Ozgur, Supervised and unsupervised machine learning techniques for text document categorization, 2004.
4. T. Mikolov, K. Chen, G. Corrado, and J. Dean, Efficient estimation of word representations in vector space, 2013.
5. R. Collobert and J. Weston, A unified architecture for natural language processing: Deep neural networks with multitask learning, in *Proceedings of the 25th International Conference on Machine Learning*, ser. ICML '08. New York, NY, USA: Association for Computing Machinery, 2008, p. 160–167. [Online]. Available: https://doi.org/10.1145/1390156.1390177.
6. Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin, A neural probabilistic language model, *J. Mach. Learn. Res.*, vol. 3, no. null, p. 1137–1155, Mar. 2003.
7. O. Glembek, P. Matejka, L. Burget, and T. Mikolov, Advances in phonotactic language recognition. 01 2008, pp. 743–746.
8. Q. Chen, X, Guo, X., & H. Bai, Semantic-based topic detection using Markov decision processes. Neurocomputing, 242, 2017, 40–50.
9. Yin, W., Kann, K., Yu, M., & Schütze, H. Comparative study of CNN and RNN for natural language processing. 2017. arXiv preprint arXiv:1702.01923.
10. Ghosh, S., Vinyals, O., Strope, B., Roy, S., Dean, T., & Heck, L. Contextual lstm (clstm) models for large scale nlp tasks 2016. arXiv preprint arXiv:1602.06291.
11. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 6000–6010.
12. D. Rothman, Transformers for Natural Language Processing: Build innovative deep neural network architectures for NLP with Python, PyTorch, TensorFlow, BERT, RoBERTa, and more. Packt Publishing, 2021.
13. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.
14. A. Cabrera, Logistic Regression Analysis in Higher Education: An Applied Perspective, 01 1994, vol. 10, pp. 225–256.
15. T. K. Ho, The random subspace method for constructing decision forests, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 8, pp. 832–844, 1998.
16. Z. Husein, "Malay-dataset," https://github.com/huseinzol05/malay-dataset/tree/master/dumping/wikipedia, 2018.
17. Z. Husein, "Malay-dataset," https://github.com/huseinzol05/malay-dataset/tree/master/dumping/twitter, 2018.
18. Z. Husein, Malay-dataset," https://github.com/huseinzol05/malay-dataset/tree/master/sentiment/translate/twitter-sentiment, 2018.
19. Z. Husein, "Malaya-speech," https://github.com/huseinzol05/malaya-speech, 2020.

20. S. Reza, M. C. Ferreira, J. J. M. Machado & J. M. R. Tavares. A multi-head attention-based transformer model for traffic flow forecasting with a comparative analysis to recurrent neural networks. Expert Systems with Applications, 202, 2022, 117275.