Hoax analyzer for Indonesian news using RNNs with fasttext and glove embeddings

Ryan Adipradana, Bagas Pradipabista Nayoga, Ryan Suryadi, Derwin Suhartono
Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta 11480, Indonesia

Article Info	ABSTRACT
	Misinformation has become an innocuous yet potentially harmful problem ever since the development of internet. Numbers of efforts are done to prevent the consumption of misinformation, including the use of artificial intelligence (AI), mainly natural language processing (NLP). Unfortunately, most of natural language processing use English as its linguistic approach since English is a high resource language. On the contrary, Indonesia language is considered a low resource language thus the amount of effort to diminish consumption of misinformation is low compared to English-based natural language processing. This experiment is intended to compare fastText and GloVe embeddings for four deep neural networks (DNN) models: long short-term memory (LSTM), bidirectional long short-term memory (BI-LSTM), gated recurrent unit (GRU) and bidirectional gated recurrent unit (BI-GRU) in terms of metrics score when classifying news between three classes: fake, valid, and satire. The latter results show that fastText embedding is better than GloVe embedding in supervised text classification, along with BI-GRU + fastText yielding the best result.

Corresponding Author:

Derwin Suhartono
Computer Science Department, School of Computer Science
Bina Nusantara University
K.H. Syahdan 9, Kemanggisan, West Jakarta 11480, Indonesia
Email: dsuhartono@binus.edu

1. INTRODUCTION

Ever since the growth and development of internet, digital information is consumed daily regardless of their validity. The presence of social media such as Facebook, Tik Tok, and Instagram has a great impact on how digital information is created and consumed. These platforms give their users the freedom to create, access, and process information in which the chance of consuming misinformation is very likely. Misinformation itself is incorrect information with accidental or intentional purposes which has been an issue since 16th century [1]. The issue is gradually becoming more problematic and popular ever since the 2016 U.S presidential election where 73 publications in English language are deemed fake in 2016 and increased to 2210 by January 2017 [2].

Misinformation or fake news can be completely made up and manipulated in order to gain attention, can be designed to mislead readers and can be purposely false. The objective of fake news is to earn benefits in politics and finance, commonly with exaggerated or unique headlines to attract readers [3]. In Indonesia political fake news has increased by 61% between December 2018 and 17th April 2019, which was the schedule of presidential election [4]. Aside from political fake news, another example of fake news in Indonesia is a false issue of earthquake aftershocks, thus directly affects traumatized victims of the earthquake and tsunami that has just occurred [5]. One research shows that fake news and valid news can be

distinguished based on numerous aspects, most notably in the title of the news. Additionally, fake news title uses significantly fewer stop-words and nouns but more proper nouns and verb phrases. In comparison, valid news persuades the readers using backed-up arguments and citations whereas fake news gains trust through heuristics [6].

Moreover, since English is a high resource language, it can be hypothesized that most of the news are conveyed in English. For natural language processing (NLP) task to perform smoothly, the training corpus is required to be the same language as the latter input text. Therefore, as Indonesian language is considered low resource language, hoax analyzer for Indonesian language news is not as many as for English news. In this experiment, we attempted to reduce fake news consumption by sorting out the fake news from a pool of unfiltered news in Indonesian language. The latter news is sorted out based on the validity with fake news having two sub classes: hoax news, and satire news, resulting in a total of three classes: valid news, hoax news, and satire news. The sorting process is performed using a supervised text classification using deep learning approach, recurrent neural network (RNN). We use four models of RNN: long short-term memory (LSTM), gated recurrent unit (GRU), bidirectional LSTM, and bidirectional GRU. All of four models use embedding layer as the input layer, with fastText and global vectors (GloVe) as word embeddings. The objective of this experiment is to evaluate and discover new insights between different RNN models using different embeddings as input layer for low resource language, mainly Indonesian language.

2. INTRODUCTION RESEARCH METHOD
2.1. Related works
A study is presented on the analysis of language used in news media in terms of fake news detection and political fact-checking in which the researchers compared language of real news to satire, hoaxes, and propaganda to identify the linguistic characteristics of unreliable news. In the experiment, English Gigaword corpus was used as the dataset for reliable news label and seven sources were used as the dataset for satire, hoax and propaganda. The process of differentiating the characteristics between news type uses lexical sources from previous works in communication theory and stylistic analysis in computational linguistics. The next process was to run LSTM model with maximum entropy (MaxEnt) and naïve bayes classifier to train and predict the reliability and validity of Politifact dataset. The latter results showed that LSTM outperformed other models when using text as input, while MaxEnt and naïve bayes performed better when using linguistic inquiry and word count (LWIC) as a feature, allowing both to increase the performance. Same treatment was applied to LSTM yielding lower performance [7].

Another work for validity checking was the introduction of dataset, LIAR. LIAR is a dataset specifically created for fake news detection which contains 12,836 short statements that have been labelled according the truthfulness, subjects, contexts, speakers, states, parties, and prior histories. Truthfulness label are split into six parts: pants-fire, false, barely true, half-true, most-true, and true. LIAR dataset was used to evaluate the performance of hybrid convolutional neural network (CNN) model in automatic fake news detection along with support vector machine (SVM), logistic regression classifier, bidirectional LSTM and vanilla cable news network. The proposed hybrid CNN model is a combination of text and meta-data; subjects, speakers, jobs, states, parties, contexts, and histories. The results showed that text combined with speaker as meta-data only performed better in validation process, while text combined with all aforementioned meta-data performed greatly in testing process. On the contrary, vanilla CNN performed the best in testing and validation process. LIAR dataset is also viable for stance classification, argument mining, topic modelling, rumour detection and political natural language processing (NLP) research [8].

A satire news focused research was attempted in order to build a satirical news classifier. Since satirical news is relatively ambiguous, it can be distorted into both humour and criticism with unknown objective [9], thus the research target is to compare satirical news to their truthful equivalent in 12 contemporary news topics from four different domains. SVM-based algorithm was used for the research with five features: absurdity, humour, grammar, negative affect, and punctuation. The testing process of 360 news yielded 90% precision and 84% recall when classifying between satirical news and their truthful equivalent [10].

Novel method for automatic fake news detection was also proposed. The idea is to include the speaker's profile as an additional feature into an attention-based LSTM model. The speaker's profile includes: party affiliation, speaker title, location and credit history. The profile would benefit LSTM model by adding the profile as an additional input data into LSTM model. Result of this novel method outperforms state-of-the-art technique's accuracy by 14.5% when using a benchmark fake news detection dataset [11].

A framework called hierarchical discourse-level structure for fake news detection (HDSF) was also introduced for better understanding of fake news. Incorporating hierarchical discourse-level structure of fake

news and valid news is an important step to understand better about fake news. HDSF operates in data-driven and automated manner, by learning and extracting discourse-level structures of fake news and valid news. A strong point of HDSF is its ability to operate without an annotated corpus, considering that the structure between fake news and valid news is recognizable [12].

Another framework, multi-source multi-class fake news detection (MMFD) was proposed to measure spectrum of "fakeness" severity. Automated feature extraction, multi-source fusion, and automated degrees of "fakeness" detection were used to create a logical and interpretable model which can effectively classify news to different levels of "fakeness" [13]. Additionally, similar papers regarding text classification can be found but to our limitations this paper is the first to classify news in Indonesian language into three classes using LSTM and GRU with fastText and GloVe. Recent papers [14]-[17] has studied fastText word embedding, deep learning models and traditional classifiers but no regards to GloVe and satirical news.

2.2. Proposed method
2.2.1. Data preprocessing

The dataset used for this experiment can be obtained from GitHub [18]. Inside is a large corpus of news articles that have been tagged accordingly in which we retrieve 1000 data each from news annotated as reliable, fake, and satire for a total of 3000 data. Since the dataset was still in English, an online translator powered by Google Translate was used to convert the dataset to Indonesian language [19].

The dataset is then cleaned by changing all upper-case letters to lower case, removing Indonesian language stop words like (*di* and *ke*) which effectively translates to (at and to) in English. Punctuation, digits, and extra spaces were also removed in the data cleaning process. Data cleaning is an important step to reduce the memory used to store words in the vocabulary dictionary as well as reducing any potential noise in the dataset. Visualization of data pre-processing for this experiment can be seen in Figure 1. The final dataset that has gone through data pre-processing include a total of 1.036.041 words with 55.411 unique words which can be seen in Figure 2. It can be inferred that fake news tends to use longer texts, while satire news has considerably lower word count.
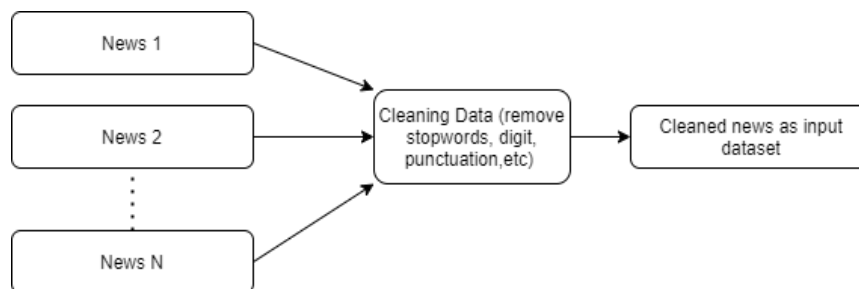

Figure 1. Cleaning data process for 3000 data which remove punctuations, Indonesian stop words, digits, and changing upper-case letters to lower case. The output would be the cleaned version of the dataset
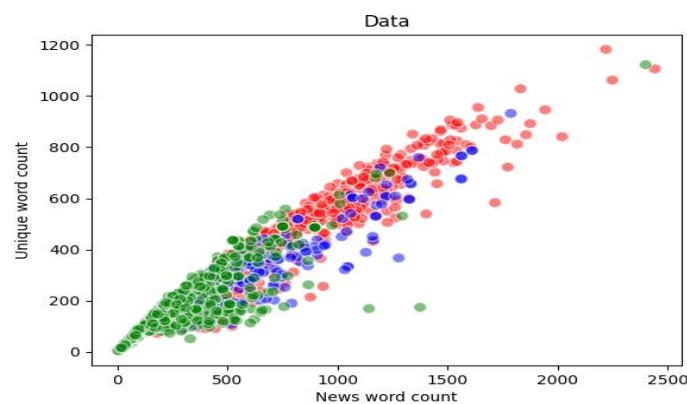

Figure 2. Scatterplot for cleaned dataset with red colour representing fake news, green colour representing satire, blue colour representing valid produced by unique word count as y axis and news word count as x axis

2.2.2. Data training process

The training process for RNN models uses keras dense layer as the classifier and the output nodes. The last node uses SoftMax for activation function since SoftMax always transforms the input values whether it is negative values, zero, and positive values. The value would be transformed into 0 and 1 which can be interpreted as probability. For model optimizer, we use Adam optimizer because of its ability to automatically adjust its learning rate. Categorical cross entropy is used as the loss function since it improves robustness for multi-label classification [20].

For the training process, 2800 data were used with random starting position. The first step is to create local corpus for fastText or GloVe embeddings and assigning unique integers for each word in dataset based on corpus values. Each weight for connected nodes is randomized, before connecting embedding layer into RNN layer. We use 100 units for each RNN layer with 10 epoch and 32 batch size. As for testing purposes, we use cross validation technique on the same dataset using 10 KFold cross validation to test the models' accuracy, precision, recall, and F1-score. All four models are trained with fastText and GloVe as the word embedding layers. The detailed visualization of the training process can be seen in Figure 3.
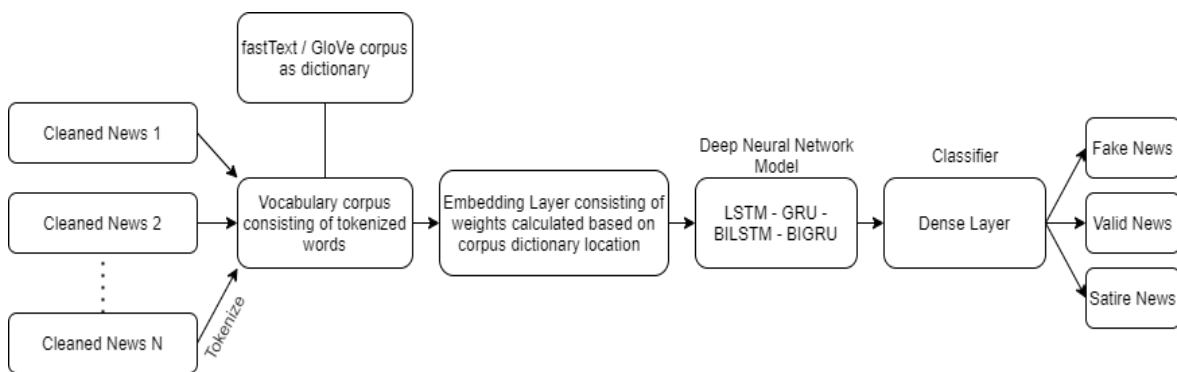

Figure 3. Illustration of deep neural network model process of classifying news into three classes, using vectored words as input for RNN model and the latter output would be between fake news, valid news or satire news

Brief explanations of the RNN models and word embedding layers used can be seen below:
a.   Long short-term memory

LSTM is a type of RNN architecture that was proposed by Hochreiter and Schmidhuber [21]. LSTM works by storing values over arbitrary time intervals which enables it to handle long-term dependency in a sequential event. The main reason of using LSTM is due to its ability to extract features from sequential input data. While originally LSTM requires input data that has timesteps, news text is suitable for LSTM since each word is recorded as one timestep. Both unidirectional and bidirectional LSTM (BI-LSTM) were used for this experiment. Each type has the same functionality with the only difference is that BI-LSTM is essentially two regular LSTM models using normal time order (from past to future) and reverse time order (from future to past) simultaneously which allows predictions to be made from both time orders.

b.   Gated recurrent unit

Unlike LSTM, GRU is an RNN model that was only recently discovered, originally proposed in 2014 [22]. GRU is also implemented alongside LSTM to compare their performances in terms of model accuracy and computational efficiency. GRU has been proven to yield higher computational efficiency which is made possible by the smaller number of gates it possesses. GRU only has two gates (update and reset) while LSTM has three (input, output, forget). For this experiment, we included both unidirectional and bidirectional (BI-GRU) versions of GRU. The logic on how data is processed by BI-GRU is the same as in BI-LSTM, where there are two GRU models with normal and reverse time order.

c.   fastText

fastText was developed by facebook's ai research (FAIR) lab and is a machine learning library used for efficient learning of word representations and sentence classification [23]. The algorithm for fastText is based on two papers released in 2016: enriching word vectors with subword information [24] and bag of tricks for efficient text classification [25]. fastText already has language support for 176 languages and have distributed pre-trained word vectors for 157 languages [26]. fastText is an extension of the word2vec model that represents each word as an n-gram of characters instead of learning vectors of words directly. fastText is

suitable for handling new, out-of-vocabulary words since fastText breaks down unknown words from corpus into n-grams which may have similarities with words inside the vocabulary corpus.

d. GloVe

GloVe was first introduced by Pennington [27] in 2014. It is an unsupervised learning algorithm used to obtain vector representations for words. In aforementioned paper, Pennington *et al.* have proven that GloVe outperforms other models like continuous bag of words (CBOW) in terms of word analogy, word similarity, and named entity tasks. GloVe learns word embeddings in a different way than word2vec. It uses a term co-occurrence matrix of size A x A, where a is the vocabulary size, in which will train the word vectors to predict co-occurrence ratios. An example is the word father will have higher cosine similarity with the word male as both words are semantically close.

3. RESULTS AND DISCUSSION

From Table 1 and Table 2, it can be seen that fastText embedding for all deep neural network (DNN) have higher metrics score compared to its counterpart GloVe which is made possible from fastText's algorithm that goes one level deeper, consisting of characters n-grams and words as the training focus instead of only words. From Table 1 and Table 2, performance score of each model can be seen with the standard deviation score. Standard deviation itself is a tool of measurement to show the amount of dispersion of the training data.

Compared to the previous experiment [16], their highest F1 macro score for bidirectional LSTM model combined with fastText was 64% while our experiment on bidirectional LSTM + fastText yields better result which is 89.234% and 1.255% standard deviation value. As for fastText experiment [17], they used fastText for text classification which produce 84% F1 score. It can be noted that our experiment using Bidirectional GRU + fastText has higher F1 score, 94.298%. From [14], their best model is 'stochastic gradient descent (SGD) modified hurbe' with 80% accuracy, 65% precision, 100% recall and 80% F1 score. Bidirectional GRU + fastText has higher overall performance, with recall score is lower. Further discussion will be separated into two parts, with one focusing on word embedding and the other focusing on recurrent neural network (RNN).

Table 1. Results from using fastText

Model Name (fastText)	Accuracy (%)	Precision (%)	Recall (%)	F1 Macro Score (%)
LSTM	81.153 (±3.168)	89.184 (±3.168)	81.023 (±3.033)	84.700 (±2.525)
BI-LSTM	87.825 (±1.455)	91.034 (±1.195)	87.786 (±1.482)	89.234 (±1.255)
GRU	89.460 (±1.694)	93.523 (±1.542)	89.439 (±1.821)	91.400 (±1.611)
BI-GRU	93.163 (±1.334)	95.710 (±1.082)	93.083 (±1.403)	94.298 (±1.245)

Table 2. Results from using GloVe

Model Name (GloVe)	Accuracy (%)	Precision (%)	Recall (%)	F1 Macro Score (%)
LSTM	81.767 (±2.418)	88.328 (±2.043)	81.725 (±2.502)	84.858 (±2.245)
BI-LSTM	86.767 (±2.066)	90.534 (±2.145)	86.766 (±2.028)	88.104 (±2.201)
GRU	85.000 (±2.399)	88.814 (±2.248)	84.921 (±2.448)	86.613 (±2.393)
BI-GRU	88.967 (±1.531)	91.805 (±1.459)	88.929 (±1.551)	89.987 (±1.457)

3.1. Word embedding layers discussion

fastText outperformed GloVe for all models except LSTM in the experiment, which is an interesting discovery. LSTM using GloVe embedding yielded higher accuracy and recall compared to LSTM that used fastText embedding. It is important to note that LSTM's and bidirectional LSTM's performance only shifts a little when using GloVe and fastText embeddings, averaging only 0.550 difference for LSTM and 0.925 for BI-LSTM in each metric. On the other hand, GRU and bidirectional GRU showed more notable differences when using fastText compared to GloVe embedding with average of 4.600 for GRU and 4.125 for BI-GRU.

Difference between fastText and GloVe lies in their approach for texts. GloVe treats each word in the corpus like an atomic entity and generates a vector for each word, where respective vectors are treated as the smallest unit to train on. On the other hand, fastText treats each word in the corpus as a combination of character n-grams and generates a vector based on the sum of each n-grams vector which notably can handle out of vocabulary (OOV) words and generate more accurate vectors of rare words since character n-grams in OOV words and rare words may still be shared with words inside the corpus. Based on this, fastText embedding have higher results compared to GloVe if the dataset has a broad spectrum.

fastText performing better than Glove in this experiment does not indicate that it will always yield better results than GloVe when used for other situations. In a paper created by Wang [28], they took six word embedding layers: Skip-gram negative sampling (SGNS), continuous bag of words (CBOW), fastText, ngram2vec, and dict2vec and conducted experiments evaluated on: word similarity, word analogy, concept categorization, outlier detection, and QVEC–a tool used for measuring intrinsic quality of word vectors. They conclude that there is no word embedding layer that is consistently better than the rest for the tasks that they were done on, which include: part-of-speech (POS) tagging, chunking, named-entity relation, sentiment analysis, and neural machine translation (NMT).

3.2. Rnn models discussion

From Table 1 and Table 2, both BI-LSTM and BI-GRU achieved better results than LSTM and GRU while disregarding the word embedding layer used. This means that the ability bidirectional models have, that is to train the dataset from both positive and negative timesteps, helped in achieving better results in all performance metrics. It is highly acceptable that bidirectional RNN is more suited to the task of supervised text classification compared to its unidirectional counterpart mainly because bidirectional RNNs can train from different standpoints.

The results from this experiment points out that GRU models are more effective compared to LSTM models in terms of metrics score. As GRU does not have a cell like LSTM, GRU yielded better metrics score when the dataset has less frequent occurrence. On the long run LSTM will yield better metrics score compared to GRU due to its model having more stability to control the flow of data as well as the presence of cell to store arbitrary data in the case of longer texts. Since the dataset used for this experiment only have few similarities and shorter texts length, GRU thrives better than LSTM.

4.   CONCLUSION

In this paper, we created neural network models for classifying fake news for the Indonesian Language using fastText and GloVe as the word embedding layer. This experiment provides conclusions is: 1) GRU has better performance compared to LSTM when the dataset has less frequent occurrence and is widely spread; 2) Both bidirectional models of LSTM and GRU yield better metrics score than their unidirectional counterparts; 3) fastText is better than GloVe in performance as fastText can handle out of vocabulary (OOV) words and rare words better than GloVe; 4) fastText and bidirectional GRU combined yielded the highest result in this experiment, mainly because the dataset is widely spread and has shorter text length. The statements above are the conclusions we reached from this experiment, which should contribute more to the natural language processing (NLP) field regarding RNN and word embedding layers. We hope that the results from our experiment can be used for future research and the study of supervised text classification and to develop a more sophisticated fake news classifier for Indonesian language. In this paper, we encountered a problem regarding the dataset. Since we used a low resource language, satirical news is harder to find so we used English news and translated them to Indonesia language. We are aware that translating English news into Indonesian language news can disrupt the actual result of the experiment, so our suggestion is to find Indonesian language news without being translated.

REFERENCES

[1]   B. Martens, L. Aguiar, E. Gomez-Herrera and F. Müller-Langer, "The digital transformation of news media and the rise," *European Commission, Joint Research Centre (JRC)*, Seville, 2018.
[2]   J. Allen, B. Howland, M. Mobius, D. Rothschild and D. J. Watts, "Evaluating the fake news problem at the scale of the information ecosystem," *Science Advances,* vol. 6, no. 14, pp. 1-6, 2020, DOI: 10.1126/sciadv.aay3539.
[3]   W. Zakharov, H. Li, and M. Fosmire, "Undergraduates' News Consumption and Perceptions of Fake News in Science," *Portal: Libraries and the Academy,* vol. 19, no. 4, pp. 653-665, 2019. DOI: 10.1353/pla.2019.0040.
[4]   K. Lamb, "Fake news spikes in Indonesia ahead of elections," *The Guardian*, 20 March 2019.
[5]   detikNews, "Kominfo Rilis 10 Hoax Paling Berdampak di 2018, Ratna Sarumpaet Nomor 1," *detikNews*, 19 December 2018.
[6]   B. D. Horne and S. Adali, "*arXiv.org*," 28 March 2017.
[7]   H. Rashkin, E. Choi, J. Y. Jang, S. Volkova and Y. Choi, "Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, 2017.
[8]   W. Y. Wang, "Liar, Liar Pants on Fire: A New Benchmark Dataset for Fake News Detection," *arXiv preprint arXiv:1705.00648*, 2017.
[9]   K. D. Landreville, "Satire as Uncertain Territory: Uncertainty expression in discussion about political satire, opinion, and news," *Humor-International Journal of Humor Research,* vol. 28, no. 4, pp. 559-582, 2015.

[10] V. Rubin, N. Conroy, Y. Chen and S. Cornwell, "Fake News or Truth? Using Satirical Cues to Detect Potentially Misleading News," in *Proceedings of the Second Workshop on Computational Approaches to Deception Detection*, San Diego, 2016, pp. 7-17,

[11] Y. Long, Q. Lu, M. Li and C-R. Huang, "Fake News Detection Through Multi-Perspective Speaker Profiles," in *Proceedings of the Eighth International Joint Conference on Natural Language Processing*, Taipei, vol. 2, 2017.

[12] H. Karimi and J. Tang, "Learning Hierarchical Discourse-level Structure for Fake News Detection," *arXiv preprint arXiv:1903.07389*, 2019.

[13] H. Karimi, P. Roy, S. Saba-Sadiya and J. Tang, "Multi-Source Multi-Class Fake News Detection," in *Proceedings of the 27th International Conference on Computational Linguistics*, Santa Fe, 2018.

[14] A. B. Prasetijo, R. R. Isnanto, D. Eridani, Y. A. A. Soetrisno, M. Arfan and A. Sofwan, "Hoax detection system on Indonesian news sites based on text classification using SVM and SGD," *2017 4th International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE)*, Semarang, Indonesia, 2017, pp. 45-49, doi: 10.1109/ICITACEE.2017.8257673.

[15] A. Nurdin, B. A. S. Aji, A. Bustamin and Z. Abidin, "Perbandingan Kinerja Word Embedding Word2vec, Glove, Dan Fasttext Pada Klasifikasi Teks," *Tekno Kompak,* vol. 18, no. 2, pp. 74-79, 2020.

[16] E. Lim, E. I. Setiawan and J. Santoso, "Stance Classification Post Kesehatan di Media Sosial Dengan FastText Embedding dan Deep Learning," *Intelligent System and Computation,* vol. I, no. 2, pp. 65-73, 2020.

[17] A. Amalia, O. S. Sitompul, E. B. Nababan and T. Mantoro, "An Efficient Text Classification Using fastText for Bahasa Indonesia Documents Classification," *2020 International Conference on Data Science, Artificial Intelligence, and Business Analytics (DATABIA)*, Medan, Indonesia, 2020, pp. 69-75, doi: 10.1109/DATABIA50434.2020.9190447.

[18] Several 27, ''Fake News Corpus,'' 2020. [Online]. Available: https://github.com/several27/FakeNewsCorpus/releases/tag/v1.0, 2020.

[19] DocTranslator, "Instantly Translate a Document from English to Indonesian," 2020.

[20] A. Rusiecki, "Trimmed categorical cross-entropy for deep learning with label noise," *Electronics Letters,* vol. 55, no. 6, pp. 319-320, 2019, DOI: 10.1049/el.2018.7980.

[21] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation,* vol. 9, no. 8, pp. 1735-1780, 1997, doi: https://doi.org/10.1162/neco.1997.9.8.1735.

[22] K. Cho, B. v. Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk and Y. Bengio, "*arXiv.org*," 3 September 2014.

[23] fastText, "*fasttext 0.9.2*," 28 April 2020.

[24] P. Bojanowski, E. Grave, A. Joulin and T. Mikolov, "*arXiv.org*," 19 June 2017.

[25] A. Joulin, E. Grave, P. Bojanowski and T. Mikolov, "*arXiv.org*," 9 August 2016.

[26] E. Grave, P. Bojanowski, P. Gupta, A. Joulin and T. Mikolov, "*arXiv.org,*" 28 March 2018

[27] J. Pennington, R. Socher and C. Manning, "GloVe: Global Vectors for Word Representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, 2014, pp. 1532-1543.

[28] B. Wang, A. Wang, F. Chen, Y. Wang and C. J. Kuo, "*arXiv.org*," 29 January 2019.