ICIC Express Letters
Volume **17**, Number **9**, September **2023**

 ISSN 1881-803X
pp. **1059–1066**

A COMPARISON OF SEVERAL EXPERIMENTAL METHODS FOR INDONESIAN AUTHORSHIP ATTRIBUTION

Karen Etania Saputra*, Riccosan and Andry Chowanda

Computer Science Department
School of Computer Science
Bina Nusantara University
Jakarta 11480, Indonesia
riccosan001@binus.ac.id; achowanda@binus.edu
*Corresponding author: karen.sapurta@binus.ac.id

Received September 2022; accepted December 2022

Abstract. *The task of determining the author's writing style in a document among a list of potential possibilities is known as Authorship Attribution (AA). Several techniques, including Natural Language Processing (NLP), Machine Learning (ML), Deep Learning (DL), and combinations of these techniques, can be used to execute AA tasks. NLP research has become highly engaged in AA over the years. However, only a few people conduct research in the Indonesian language field. As a result, this research explores and presents several comparative results of the AA task implementation on Indonesian-language documents using various methods, namely Bidirectional Long Short-Term Memory (Bi-LSTM), Bi-LSTM with Gated Recurrent Unit (GRU), Term Frequency Inverse Document Frequency (TFIDF) with Cosine Similarity, and TFIDF with Multinomial Naïve Bayes (TFIDF-MNB). Therefore, a self-collected dataset in the type of online news was produced to explore the classification of AA in order to carry out the AA task in this research. The Transformers architecture has been used in prior studies to conduct AA experiments on the DL approach, which will also be present in this comparative result method. The current worst accuracy result, as determined by TFIDF with Cosine Similarity, is 0.156. However, the best outcomes were* ***0.74*** *for the test predict accuracy and* ***0.86*** *for the Top-K Accuracy when the IndoBERT technique was combined with KFold Cross-Validation.*

Keywords: Authorship attribution, Deep learning, Natural language processing, Machine learning, Transformers, Indonesian language

1. **Introduction.** How one expresses or communicates their thoughts or linguistic style is a defining characteristic of who they are. Language style can influence readers' emotions and thoughts as well as captivate, persuade, or create a certain mood for listeners and readers [1]. Everybody has a different manner of communicating. A person's writing or conversation displays their linguistic personality when speaking or even writing. The fundamental goal of Authorship Attribution (AA) is to identify the author of a text or document by studying the language used in previously published works, such as books, essays, social media posts, and other media [2, 3]. Numerous studies describe the various methods of using AA in detail. Unfortunately, we frequently encounter situations in the real world where we are unsure who wrote the item we are reading [4]. By identifying the original author of a piece and preventing plagiarism, authorship attribution is also important for digital crimes like cyber forensics [5]. Another way to prevent hoaxes is to check to see if the article was actually written by an actual individual. According to the definition of plagiarism, it is the taking or expropriation of any words, concepts, documents, or creative works from the works of others (original authors) for economic goals or as referrals, which can lead to public trickery or hoaxes [6]. On the other hand, a hoax is

DOI: 10.24507/icicel.17.09.1059

an untrue information or news that has been mainly composed by anyone from an unreliable source in an effort to deflect attention from the facts [7]. Based on the research on literacy, Wang used three different types of techniques for AA, including Support Vector Machine (SVM), Naïve Bayes (NB), Nearest Neighbor Classifier with Global Vectors for Word Representation (GloVe) embedding, and Recurrent Neural Network (RNN) with GloVe [8]. From Wang's strategy, RNN with GloVe, which has an accuracy of 0.6, produces the best results. The problem with Wang's research is that the model is overfitting since the dataset is too small. This report presents the recommendation to experiment with various RNN modifications using GloVe as the initial value. In another experiment, Gupta et al.'s research used two different embedding types. They employed index-based word embedding for both datasets, and for the first dataset, GloVe pre-trained embedding [9]. In addition, Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) are employed techniques. The GRU model delivers the best results, but, like Wang's research, it also has dataset flaws. Both studies recommend investigating additional RNN varieties. In further research [10], examine pre-trained language models including BERT, ELMo, ULMFiT, GPT-2, and multi-headed classifier in order to perform a cross-domain AA. While ULMFiT and GPT-2 are less competitive and require more training epochs, BERT and ELMo are better able to outperform the RNN baseline. Rahgouy et al. also conducted AA tests using word embedding ensembles from N-Gram, Word2Vec, and Term Frequency Inverse Document Frequency (TFIDF) models [11], among others. Another uses ensemble approaches with SVM, such as SMO-SVM (Sequential Minimal Optimization) [12].

The dataset used in this study is the same dataset used in previous research [13], i.e., datasets that the authors independently acquired from online news portal websites. By examining previously undiscovered techniques and contrasting various Deep Learning (DL) techniques like RNN, Natural Language Processing (NLP), and Machine Learning (ML), this study seeks to advance prior research. The techniques are Bidirectional Long Short-Term Memory (BiLSTM) [14] with GloVe [15], BiLSTM with GRU [16] and GloVe, TFIDF with Cosine Similarity, and TFIDF with Multinomial Naïve Bayes (MNB) [17]. The exploration of models between the 4 experiments in the 2 groups that have been mentioned and compared with previous research [13] that employed transformers approaches like MBERT, IndoBERT, IndoBERT with KFold Cross-Validation and MBERT with KFold Cross-Validation is the contribution of this study. There are five sections in this article. The introduction in Section 1 describes the purpose, inspiration, and history of the research. The foundation is provided in Section 2 in the form of past research (i.e., the literature review and recent work). The proposed research technique and flow are described in Section 3. The findings of the experiments and their full scope are presented in Section 4. Section 5 concludes the findings and provides suggestions for future research.

2. **Literature Review.** The BERT model for the AA problem was tuned in the most recent study by Fabien and his team [18]. The Enron Email Corpus [19], the IMDb Authorship Attribution Corpus [20], and the Blog Authorship Attribution Corpus [21] are the three types of datasets used in this study. The BertAA model was shown to be 93% accurate in this analysis compared to the most recent State of the Art (SOTA). The better dataset and wide variety of the dataset are advantages of this study. The PCFG (Probabilistic Context-Free Grammar) technique was used in the research by Fuller et al. [22] to complete the authorship attribution problem. The tests by Fuller et al. focus on sentence-level detection. For performance evaluation, two sets of sample data were used: the first included ten articles/works from 10 novelists from the 19th and early 20th centuries from the Gutenberg Project, and the second included ten articles/works from contemporary suspense/mystery writers. The most accuracy was gained by Fuller et al.'s study at 87%, whereas the best accuracy was attained by the comparative model, Support

Vector Machines (SVM), at 89%. Fuller et al.'s study was carried out at the sentence level instead of concentrating on a paragraph level.

The research of Hitschler et al. [23] is the other research source, and it employs the CNN (Convolutional Neural Networks) architecture for the execution of the AA problem and achieves a model accuracy of 95% while having relatively low data variance. The information used in this study comes from single-author papers that were presented at various workshops and conferences on computational linguistics and natural language processing. Hitschler et al. used the Association for Computational Linguistics (ACL) Anthology Reference Corpus to get their data [24]. As a result of their extensive modelling capabilities, Hitschler et al.'s model may be overfitting. Its limitations are the overfitting models and data variation from Hitschler et al.'s study. However, the precision is also almost great.

By implementing Long Short-Term Memory (LSTM) and Bidirectional Long Short-Term Memory (BiLSTM) models using Chainer to extract features that correspond to the AST source code automatically, the depth first search algorithm is used in this study to contribute to authorship attribution using AST (Abstract Syntax Tree) based source code [25]. The Python dataset from Google Code Jam (GCJ) and the C++ dataset from Github are both used in this study. Utilising Stochastic Gradient Descent (SGD), the model is trained. The model has been tested using a variety of datasets, programming languages, and authors. The study's findings showed that this model effectively analyzes the structure of AST characteristics. Researchers hope to examine other source codes in the future that are larger and more diverse. The source code dataset was used in this study, which is one of its limitations, but it also produced a model that may be applied to ongoing research.

Another research by Wang [8] used a dataset containing 5,000 news stories written by 50 distinct authors (an average of 100 texts per author), and it was split 50 : 50 for training and testing. The method uses a variety of machine learning techniques, including Support Vector Machine (SVM) and Naïve Bayes (NB), with an accuracy of 12.24%, Nearest Neighbor Classifier with Global Vectors for Word Representation (GloVe) embedding for word level and the highest F1-score of 0.46, and Recurrent Neural Network (RNN) with GloVe as well, with the highest accuracy for this study of 0.6. RNN has the best outcome of all the studies because it extracts data from the training dataset at the word and sentence levels. The study's issue is that the dataset is too small, which leads to overfitting. By utilizing a larger dataset to prevent overfitting, using GloVe as initial values, and experimenting with various RNN types, the author hopes to increase the performance of RNNs in the future.

Then Gupta et al. [9] employed two datasets, the British Broadcasting Corporation (BBC) News dataset with 2225 articles and the Reuters Corpus Volume I (RCV1)/C50 dataset with 2,500 texts, 50 texts per author. For both datasets, the researchers employed index-based word embedding, and for the first dataset, GloVe pre-trained embedding (C50). Adam is chosen as the optimizer, with an epoch of 200-300 and a learning rate of 0.004. The accuracy of the C50 dataset, which is split into 90% training and 10% testing, is 66.67% with word index embedding and LSTM, 78.1% with GRU, and 61.47% with GloVe and LSTM, 69.2% with GRU. The BBC dataset, divided into the same portions as the first dataset, produced accuracy rates of 94.73% with word index embeddings and LSTM and 96.65% with GRU. Although Wang's research and this research both have dataset problems, the GRU model produces the most outstanding outcomes. In the future, researchers want to investigate several RNN architectures for author identification tasks using various word representations to enhance performance and more enormous datasets for better update weights.

3. **Methodology.** This research begins with the employment of data scraping to manually gather articles via the author profile link in order to build the necessary dataset for the AA assignment in the Indonesian language, as shown in Figure 1. The data is next manually processed in a procedure known as "Data Pre-Processing", which comprises Data Cleaning activities, including symbol removal, URL removal, and unused information removal records (such as captions and advertisement rows). We also incorporate data shuffle, label encoding, and text columns. Data shuffles are used so that the model can read multiple patterns from the available data instead of just one. The Model Training step is then further investigated using GloVe with Bidirectional Long Short-Term Memory (BiLSTM), GloVe with BiLSTM and Gated Recurrent Unit (GRU), Term Frequency Inverse Document Frequency (TFIDF) with Cosine Similarity, and TFIDF with Multinomial Naïve Bayes (MNB).

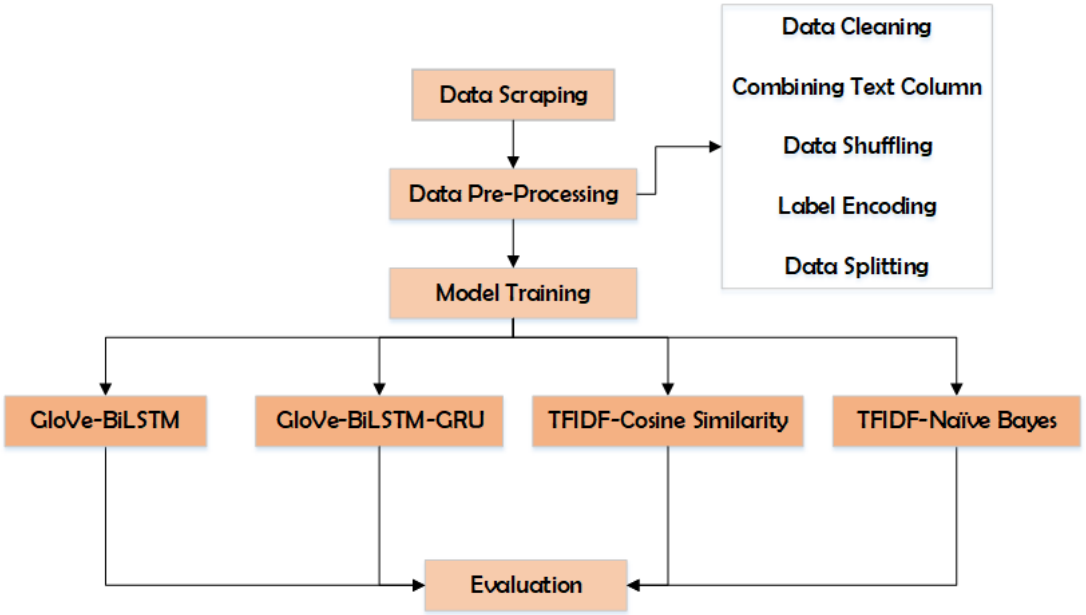

FIGURE 1. Research flow

The Evaluation step is the final phase and is where data from the model evaluation and training results are gathered. The training outcomes are first shown as a plotting graph. The model is then evaluated using test data and a classification prediction test; the results are presented as test prediction accuracy data from the classification report metric and Top-K Accuracy to assess the overall performance of the model (see Equations (1) and (2)). Cosine similarity is used explicitly in the evaluation for one of the TFIDF experiments (see Equation (3)). The classification report metric's accuracy formula is as follows, and it is calculated individually for each class.

$$\frac{\text{True Positive} + \text{True Negative}}{\text{Positive} + \text{Negative}} \tag{1}$$

The following formula (Equation (2)) is used for Top-K Accuracy.

$$\frac{\text{Top-K True Predicted Label}}{\text{Total True Label}} \tag{2}$$

The cosine similarity in TFIDF is calculated using the following formula (Equation (3)).

$$\cos(\mathbf{v}, \mathbf{w}) = \frac{\mathbf{v} \cdot \mathbf{w}}{||\mathbf{v}|| \cdot ||\mathbf{w}||} = \frac{\sum_{i=1}^{n} \mathbf{v}_i \mathbf{w}_i}{\sqrt{\sum_{i=1}^{n} \mathbf{v}_i^2} \sqrt{\sum_{i=1}^{n} \mathbf{w}_i^2}} \tag{3}$$

4. **Experiment Result.** The dataset, which is in the form of a Comma Separated Value (CSV) and consists of articles gathered from online news portals, is one of the contributions of this thesis. Data from 4,037 records, comprising 80 authors and one non-author, was used. The dataset is gathered by data scraping to manually gather articles via the author profile link in order to build the necessary dataset for the AA assignment in the Indonesian language. After the data scrapping process, the collected data proceeds through the pre-processing stage, as explained in Chapter 3 – Methodology. In the study's experiment, GloVe embedding, BiLSTM, and GRU models were used. Each model completed two separate experiments comprising different layers: a layer with BiLSTM alone (Figure 2) and a layer with BiLSTM and GRU (Figure 3). The model is run using the adam optimizer and hyperparameters, namely learning rate 3e-4, epoch 90 for BiLSTM-GRU, and epoch 100 for BiLSTM. GloVe Embedding layer (50), BiLSTM layer (512, 256, 128, 128, 64), and dense layer (81) with activation softmax are the layers that make up the BiLSTM model alone. The GloVe Embedding layer (50), BiLSTM layer (512, 256, 128, 128, 64), GRU layer (64), and dense layer (81) with activation softmax are the layers in the BiLSTM-GRU model. The BiLSTM experiment's findings are presented in Figure 2, with a training accuracy of 0.7474, training loss of 1.0756, validation accuracy of 0.2546, and validation loss with a final value of **3.0034**. A higher training accuracy of **0.8263**, a lower training loss of **0.7859**, a higher validation accuracy of **0.2603**, and a slightly higher validation loss of 3.1104 are seen with the BiLSTM-GRU technique (Figure 3), as reported in Table 1.

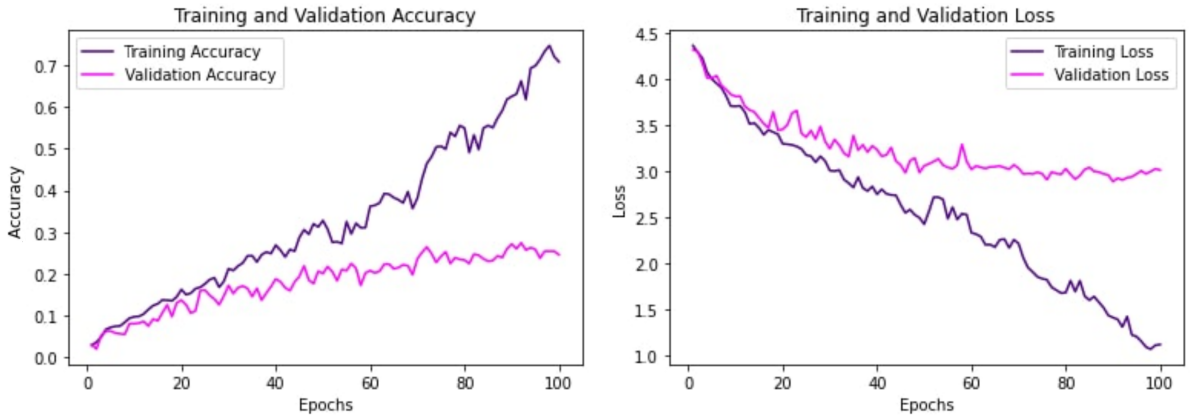

FIGURE 2. BiLSTM accuracy and loss plot

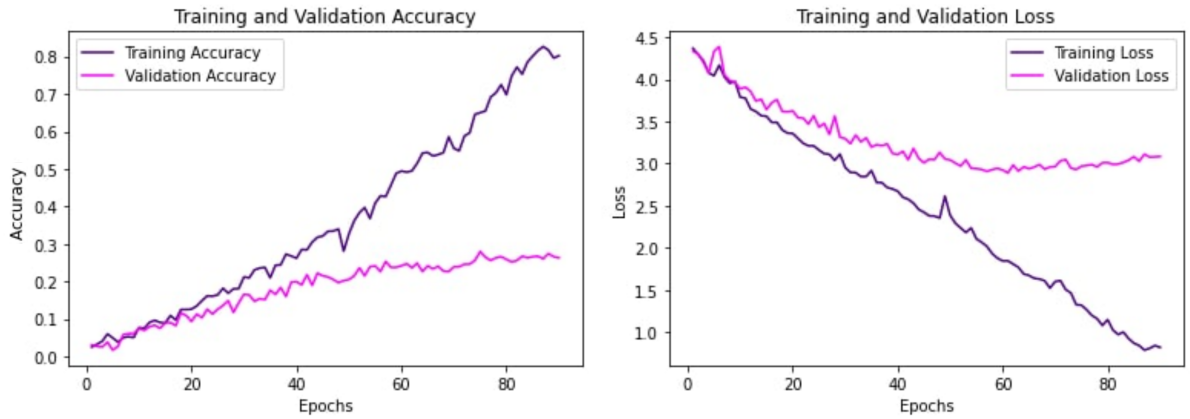

FIGURE 3. BiLSTM with GRU accuracy and loss plot

Table 1. First experiment result

Method	Training Acc.	Training Loss	Validation Acc.	Validation Loss	Test Predict Acc.	Top-K Acc.
BiLSTM + GRU with GloVe	**0.8263**	**0.7859**	**0.2603**	3.1104	**0.30**	**0.407**
BiLSTM with GloVe	0.7474	1.0756	0.2546	**3.0034**	0.25	0.395

Following the training phase, the model is evaluated with a prediction test, with the accuracy results using the classification report approach and the Top-K Accuracy measure being the same as the previous research [13]. With a Test Predict Accuracy of **0.3** and a Top-K Accuracy score of **0.407**, the BiLSTM-GRU model with GloVe embedding emerged as the top model from the evaluation phase. However, it is clear that each iteration's Validation Loss findings are quite high (Table 1). The second experiment employs NLP and machine learning, specifically TFIDF with Cosine Similarity and TFIDF with Multinomial Naïve Bayes (MNB). The experiment results indicate that TFIDF-MNB has a greater prediction accuracy of **0.45** compared to cosine, which displays 0.156, as given in Table 2. Based on these findings, it can be concluded that the NLP algorithm alone performs less well if other classifiers are not included, as NLP just interprets written text whereas Machine Learning (ML) produces predictions based on patterns discovered through experience.

Table 2. Second experiment result

Methods	Test Predict Acc.	Top-K Acc.
TFIDF with Cosine Similarity	0.156	–
TFIDF + MNB	**0.45**	0.032

The results of the M-BERT model training using the percentage split approach and KFold Cross-Validation are shown next (Table 3) from the previous experiment [13]. The KFold Cross-Validation M-BERT model generally performs better than the percentage split method. As a consequence of using KFold Cross-Validation, the training accuracy was **0.963**, the training loss was **0.206**, the validation accuracy was **0.682**, and the validation loss was slightly higher with a final value of 1.192. As demonstrated in Table 3 with the percentage split approach, the training accuracy is 0.959, the training loss is 0.235, the validation accuracy is 0.675, and the validation loss is **1.121**. IndoBERT had a greater training accuracy of **0.997** and a lower training loss of **0.062** with a worse validation accuracy of **0.738** and a higher validation loss of **0.992** using the percentage split strategy. Even though the training accuracy had a 0.006 smaller difference, the KFold Cross-Validation approach had a higher validation accuracy of **0.742** and a lower validation loss of **0.903**.

Table 3. Previous research experiment result

Method	Training Acc.	Training Loss	Validation Acc.	Validation Loss	Test Predict Acc.	Top-K Acc.
IndoBERT (8 Epoch)	**0.997**	**0.062**	0.738	0.992	0.70	0.84
IndoBERT with KFold	0.991	0.075	**0.742**	**0.903**	**0.74**	**0.86**
M-BERT (10 Epoch)	0.959	0.235	0.675	1.121	0.72	0.82
M-BERT with KFold	0.963	0.206	0.682	1.192	0.63	0.81

The best model is IndoBERT with KFold, which was tested in the two experiments mentioned above (Tables 1 and 2) as well as in earlier studies (Table 3) [13]. This model offers the best validation loss, validation accuracy, Test Predict Accuracy, and Top-K Accuracy due to the pre-trained model created particularly for the Indonesian language, while IndoBERT is superior in terms of training accuracy and loss. Although the validation loss is quite high and the training loss is quite low, the model is still overfitting despite the relatively strong training accuracy and low validation loss (Table 3). According to data analysis, more data articles are required to address this issue. The mentioned overfitting situation arises when the model fits the train data so well that it is unable to function effectively and has low specificity. This is due to the fact that the data the researchers have is insufficient when compared to the entire number of labels already in use, making the model less successful as it learns to train data by memorizing data patterns. According to the analysis of (Tables 1, 2 and 3), the best accuracy results (IndoBERT) were obtained in this study, which were 0.74 for the Test Predict Accuracy and 0.86 for the Top-K Accuracy. These results outperformed the accuracy of earlier research using BERT in the AA domain [18], which had accuracy results of 65.4% for 10 authors and 59.7% for 50 authors.

5. **Conclusion.** These researchers compared Deep Learning (DL) with Natural Language Processing (NLP) methodologies to develop a model that can categorize authors. The dataset was obtained at the paragraph level, making it possible to develop models using the IndoBERT, IndoBERT-KFold, Multilingual BERT (MBERT), MBERT-KFold, Bidirectional Long Short-Term Memory (BiLSTM), Gated Recurrent Unit (GRU), as well as Term Frequency Inverse Document Frequency (TFIDF) with Multinomial Naïve Bayes (MNB) and TFIDF with Cosine Similarity models for NLP. After completing the various stages of research – starting with data scraping, then reaching the stage of training and evaluation. According to the research analysis, the IndoBERT model employing KFold Cross-Validation is the best model with the best Test Predict Accuracy **0.74** and Top-K Accuracy **0.86**'s findings. Although the training accuracy is generally good and the training loss is relatively low, the somewhat significant validation loss can suggest that the model is overfitting. This occurs because there are not enough data compared to all the labels that are now in use; for future study, the data per label will be raised. It is still possible to expand the dataset used for research for AA assignments in order to gain larger datasets and lessen overfitting during training. Additional DL architectures can be used to create a suitable architecture probability for the AA task. DL will receive greater attention from researchers in the future because it has produced the best results so far.

REFERENCES

[1] M. Maharani, *A Sociolinguistics Analysis of Language Style in "Wild Child" Movie*, Ph.D. Thesis, Universitas Muhammadiyah Mataram, 2019.

[2] E. Stamatatos, A survey of modern authorship attribution methods, *Journal of the American Society for Information Science and Technology*, vol.60, no.3, pp.538-556, 2009.

[3] M. Koppel, J. Schler and S. Argamon, Computational methods in authorship attribution, *Journal of the American Society for Information Science and Technology*, vol.60, no.1, pp.9-26, 2009.

[4] M. Koppel, J. Schler, S. Argamon and Y. Winter, The "fundamental problem" of authorship attribution, *English Studies*, vol.93, no.3, pp.284-291, 2012.

[5] F. Iqbal, M. Debbabi and B. C. M. Fung, *Machine Learning for Authorship Attribution and Cyber Forensics*, Springer, 2020.

[6] M. Bouville, Plagiarism: Words and ideas, *Science and Engineering Ethics*, vol.14, no.3, pp.311-322, 2008.

[7] C. Juditha, Symbolic interaction in the anti-hoax virtual community to reduce the spread of hoax, *Journal of Communication and Development*, vol.19, no.1, pp.17-32, 2018.

[8] L. Z. Wang, *News Authorship Identification with Deep Learning*, https://www.semanticscholar.org/paper/News-Authorship-Identification-with-Deep-Learning-Zhou-Wang/d2af63c41d164a1f33873150758d427d8d89421b, 2017.
[9] S. T. P. Gupta, J. K. Sahoo and R. K. Roul, Authorship identification using recurrent neural networks, *Proc. of the 3rd Int. Conf. on Info. System & Data Mining*, pp.133-137, 2019.
[10] G. Barlas and E. Stamatatos, Cross-domain authorship attribution using pre-trained language models, *IFIP Int. Conf. on Artif. Intel. Apps. & Innov.*, pp.255-266, 2020.
[11] M. Rahgouy, H. B. Giglou, T. Rahgooy, M. K. Sheykhlan and E. Mohammadzadeh, Cross-domain authorship attribution: Author identification using a multi-aspect ensemble approach, *CLEF (Working Notes)*, 2019.
[12] M. Al-Sarem, F. Saeed, A. Alsaeedi, W. Boulila and T. Al-Hadhrami, Ensemble methods for instance-based Arabic language authorship attribution, *IEEE Access*, vol.8, pp.17331-17345, 2020.
[13] K. E. Saputra, Riccosan and A. Chowanda, Automatic Indonesian authorship attribution recognition using transformer, *ICIC Express Letters*, vol.17, no.5, pp.497-503, 2023.
[14] A. Graves and J. Schmidhuber, Framewise phoneme classification with bidirectional LSTM and other neural network architectures, *Neural Networks*, vol.18, nos.5-6, pp.602-610, 2005.
[15] P. Jeffrey, R. Socher and C. D. Manning, GloVe: Global vectors for word representation, *Proc. of the 2014 Conf. on Empir. Methods in NLP (EMNLP)*, pp.1532-1543, 2014.
[16] K. Cho, B. van Merriënboer, D. Bahdanau and Y. Bengio, On the properties of neural machine translation: Encoder-decoder approaches, *Proc. of the 8th Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-8)*, Doha, Qatar, pp.103-111, 2014.
[17] A. M. Kibriya, E. Frank, B. Pfahringer and G. Holmes, Multinomial Naïve Bayes for text categorization revisited, *Australasian Joint Conference on Artificial Intelligence*, pp.488-499, 2004.
[18] M. Fabien, E. Villatoro-Tello, P. Motlicek and S. Parida, BertAA: Bert fine-tuning for authorship attribution, *Proc. of the 17th International Conference on Natural Language Processing (ICON)*, pp.127-137, 2020.
[19] B. Klimt and Y. Yang, The Enron corpus: A new dataset for email classification research, *European Conference on Machine Learning*, pp.217-226, 2004.
[20] Y. Seroussi, I. Zukerman and F. Bohnert, Authorship attribution with topic models, *Computational Linguistics*, vol.40, no.2, pp.269-310, 2014.
[21] J. Schler, M. Koppel, S. Argamon and J. W. Pennebaker, Effects of age and gender on blogging, *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, vol.6, pp.199-205, 2006.
[22] S. Fuller, P. Maguire and P. Moser, A deep context grammatical model for authorship attribution, *Proc. of the 9th International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, pp.4488-4492, 2014.
[23] J. Hitschler, E. Van Den Berg and I. Rehbein, Authorship attribution with convolutional neural networks and pos-eliding, *Proc. of the Workshop on Stylistic Variation (EMNLP2017)*, pp.53-58, 2018.
[24] S. Bird, R. Dale, B. Dorr, B. Gibson, M. Joseph, M.-Y. Kan, D. Lee, B. Powley, D. Radev and Y. F. Tan, The ACL anthology reference corpus: A reference dataset for bibliographic research in computational linguistics, *Proc. of the 6th International Conference on Language Resources and Evaluation (LREC'08)*, 2008.
[25] B. Alsulami, E. Dauber, R. Harang, S. Mancoridis and R. Greenstadt, Source code authorship attribution using long short-term memory based networks, *European Symposium on Research in Computer Security*, pp.65-82, 2017.