ICIC Express Letters
Volume **17**, Number **5**, May **2023**

ICIC International ©2023 ISSN 1881-803X
pp. **497–503**

AUTOMATIC INDONESIAN AUTHORSHIP ATTRIBUTION RECOGNITION USING TRANSFORMER

Karen Etania Saputra[1,*], Riccosan[1] and Andry Chowanda[2]

[1]Computer Science Department, BINUS Graduate Program – Master of Computer Science
[2]Computer Science Department, School of Computer Science
Bina Nusantara University
Jl. K. H. Syahdan No. 9, Kemanggisan, Palmerah, Jakarta 11480, Indonesia
riccosan001@binus.ac.id; achowanda@binus.edu
*Corresponding author: karen.sapurta@binus.ac.id

Received May 2022; accepted August 2022

Abstract. *Authorship Attribution (AA) is the task of determining the writer of a document by identifying the author's writing patterns. In addition, AA can track plagiarism or detect hoaxes widely circulated publicly. Indonesia is a vast nation with a large population, enabling the propagation of hoaxes and information plagiarism. Unfortunately, there are only limited data and research in the AA area in the local language (i.e., Indonesian). Therefore, this research aims to collect a dataset for AA tasks in the Indonesian language and explore the AA classification modelling using state-of-the-art deep learning architectures. In this research, two pre-trained models (i.e., IndoBERT and Multilingual BERT or M-BERT) were implemented to accomplish the AA task by classifying authors with data collected from the Indonesian news article. The AA task done in this research was to categorize news authors using data derived from Indonesian news articles. This research found that the IndoBERT model trained using the K-fold Cross-Validation approach had the best results, with a prediction experiment accuracy value of 74% and Top-K accuracy of 86%.*

Keywords: Authorship attribution, Deep learning, Transformer, BERT, IndoBERT, Multilingual BERT

1. **Introduction.** The language style is a dimension or way of expressing or communicating one's thoughts, which characterize the individual's traits [1]. Everyone communicates in their unique way. When someone speaks or writes, their writing or conversation reflects their linguistic style. The Authorship Attribution (AA) becomes the main idea here to find out the name of the author in a particular text created by the author by analyzing the style of language in an existing work [2, 3], such as articles, social media post, and others media. Unfortunately, in the real world, we frequently encounter circumstances in which we cannot be sure that the article we are reading was written by the original author [4]. Authorship attribution has the ability to help avoid plagiarism by identifying the article's original author. Another possibility is to avoid hoaxes by determining whether or not a legitimate person produced the article. Plagiarism is defined as the stealing or expropriation of all words, ideas, documents, and creativity from the work of others (original writers) for practical purposes or as references, which can result in public fraud or hoaxes [5]. Meanwhile, a hoax is information or news that comprises things that are unclear or not facts that are made up by someone from an untrustworthy source in order to divert attention away from the truth [6]. Based on literacy findings, Hitschler et al. implemented CNN for AA with a reasonably excellent accuracy result of 78.57 percent [7]. The data for classes in Hitschler et al.'s research do not vary; therefore, a "topic" column in this research will also be used to identify the author. Both papers of Solorio et al. and

DOI: 10.24507/icicel.17.05.497

Fuller et al. used Probabilistic Context-Free Grammar (PCFG) to solve the AA problem with reasonable accuracy; however, Fuller et al.'s model has problems, while Solorio et al.'s uses a shorter text and a different text source, namely the forum [8, 9]. Based on the study, Fabien et al. added features and fine-tuned the BERT that is employed for AA [10]. Through their research, Fabien et al. obtained the State-of-the-Art (SOTA) accuracy of 93% with the BERT model. As a result of the literature findings, this research uses the transformers architecture proposed by Vaswani et al. [11] in the form of pre-trained models, namely Multilingual BERT (M-BERT) [12] and IndoBERT [13].

The implementation of AA assignments in Indonesian is the primary reason for utilizing these two models. The M-BERT model trained on data that included a variety of languages, including Indonesian texts that make up a small portion of their pre-training corpus. Multilingual models such as theirs may be effective in high-resource language contexts, but they are not as useful in other situations, such as in the Indonesian language [14]. However, due to the different quantities of pre-training data and a more precise tokenization strategy, research has shown that monolingual types of models are typically more high-performing than multilingual models [15]. This is particularly evident in the pre-trained monolingual models in multiple languages, such as IndoBERT [13] and PhoBERT for Vietnamese [16], which consistently beat their multilingual equivalents on downstream tasks. Since IndoBERT has explicitly been trained using the Indonesian language corpus, the research's contribution takes the form of performing AA assignments using basic Indonesian language on a pre-trained model. We believe that the IndoBERT base model [13] might thus be used as an alternate way to multilingual models in the case of Indonesian regional languages closely linked to the national language. Therefore, on the IndoBERT model that has been generated, we apply a Multilingual BERT base uncased for comparison. This paper is divided into five sections. Section 1 is the introduction, which explains the research's contribution, motivation and background. Section 2 provides the foundation in the form of prior research (i.e., the literature review and recent work). Section 3 describes the research's methodology and flow proposed in this research. Section 4 comprehensively illustrates the experiments that were done and the results. Finally, Section 5 concludes the findings and proposes suggestions for future research.

2. **Literature Review.**

The first research by Fuller et al. [9] applied the PCFG (Probabilistic Context-Free Grammar) method to complete the authorship attribution task. Fuller et al.'s experiments apply detection at the sentence level. Two sets of sample data were employed for performance evaluation: the first was ten articles/works from 10 19th century/early 20th century novelists from the Gutenberg Project, and the second was ten articles/works from contemporary suspense/mystery writers. The maximum accuracy was attained by Fuller et al.'s study at 87%, while the comparison model, Support Vector Machines (SVM), achieved the best accuracy at 89%. Although Fuller et al.'s study is quite accurate, it stumbles from the fact that it was performed at the sentence level as opposed to this study's focus on a paragraph. The second research reference is the research of Hitschler et al. [7] which applies the CNN (Convolutional Neural Networks) architecture for the implementation of the AA task, which achieved a model accuracy of 95% even though the variation of the data used was relatively low. This study uses data from single-author research papers published at various conferences and workshops in computational linguistics and natural language processing. The data used by Hitschler et al. was obtained from the Association for Computational Linguistics (ACL) Anthology Reference Corpus [17]. Hitschler et al. mentioned that their models are potentially overfitting due to their large modeling capacity. The overfitting models and data variation from Hitschler et al.'s study are its drawbacks. In addition, the accuracy is almost excellent.

Solorio et al. used Modality Specific Meta Features (MSMF) for the AA task in a subsequent study [8]. Solorio and his team used data from the Chronicle of Higher Education

(CHE) as the source of data for this study, intending to analyze the model for AA's assignments based on posts in the web forum. Single-subject postings were retrieved and organized into five data sets, each with a distinct number of writers. The most significant results were obtained using MSMF combined with First Level Features (FLF) of 75.47% in this study. Regarding the results obtained by Solorio et al., this research was inspired to use the transformers architecture [11] in the form of a pre-trained IndoBERT [13] model for the AA assignment in Indonesian. The latest research from Fabien and team [10] applied fine-tuning to the BERT model for the AA problem. This research uses 3 kinds of datasets, namely Enron Email Corpus [18], IMDb Authorship Attribution Corpus [19], and Blog Authorship Attribution Corpus [20]. In this study, the BertAA model was shown to be 93% accurate compared to the most recent SOTA. As a benefit, this new study is better suited for our research, and the dataset is quite diverse. Pires et al.'s research [21] indicated that M-BERT (Multilingual BERT) can generalize across languages. According to Pires et al.'s result, m-BERT can handle languages without being specifically trained for them, but it is also not completely ideal if the language is complex. As a result, this model will be utilized to implement AA assignments in Indonesian in a future study. Because AA's objective is to classify the author's name, there is a chance that the model would overfit due to inadequate data or incorrect parameters. Therefore, the Cross-Validation method is used to split the training data into two parts: train and valid. K-fold Cross-Validation is the most common [22] and fundamental [23] type of Cross-Validation. K-fold divides the data into K-equal-sized folds, with each fold undergoing a separate iteration, with one portion used for validation and the remainder for training.

3. **Methodology.** As shown in Figure 1, this research begins with the implementation of Data Scrapping to manually collect articles via the author profile link to create the needed dataset for the AA task in the Indonesian language. Next, the data is manually processed in the Data Pre-Processing step, which includes data cleaning such as symbol removal, URL removal, and un-used information removal (such as captions and advertisement rows). After the data cleaning process, we integrate text columns, data shuffling, data splitting, and label encoding. The purpose of data shuffling is so that the model does not only read one particular pattern from the available data. The data splitting process is applied in two different methods. First, the percentage split divides the data into three parts: train, valid, and test according to a predetermined percentage. The second is the addition of the K-fold Cross-Validation approach, which divides the training data into train and valid. Next, further research into the Model Training stage uses two pre-trained models of the transformers architecture: Multilingual BERT (M-BERT) Base Uncased and IndoBERT Base. The training model stage was run four times because two data

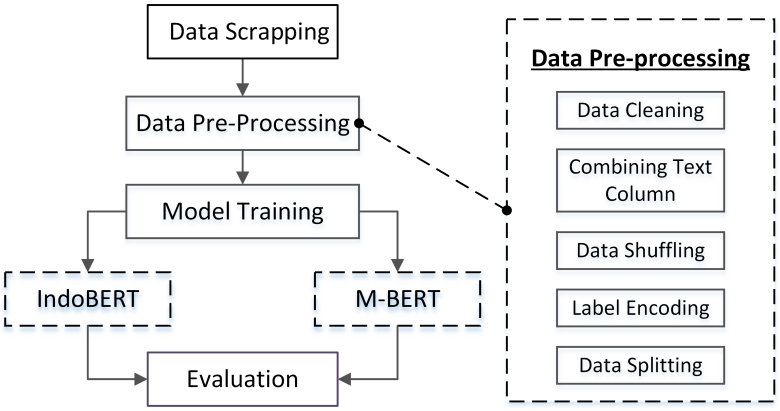

FIGURE 1. Research flow

splitting methods were applied to each model. The last stage is the Evaluation, where data from the training results and model evaluation are collected. First, the training results are displayed in a plotting graph. Then the model's evaluation is carried out through a classification prediction test using test data, the first result is in the form of test_prediction_accuracy data from the classification_report metric, and the second result is Top-K Accuracy to determine the performance of the model in general (see Equations (1) and (2)). The following is the accuracy formula in the classification report metric, and this calculation is carried out 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.

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

4. **Experiment Result.** As one of the contributions of this paper, a dataset that consists of articles obtained from Internet news portals is compiled into a dataset in the Comma Separated Value (CSV) document format. 4,037 records of data were utilized, including 80 authors and one non-author. The data is pre-processed in the form of cleaning words or sentences that are not part of the author's writing style, such as when advertising or exporting photographs to the web. Data cleansing is done manually because there is no advertising pattern in the data. After the data cleaning process, we integrate text columns, data shuffling, data splitting, and label encoding. Data splitting is carried out with two methods: percentage split and K-fold Cross-Validation. The data was divided into three sections for the percentage split: 65% train data, 25% validation data, and 10% test data. For the K-fold Cross-Validation, the data was divided into two sections: 90% train data and 10% test data, with 5 total folds. The models used in this research are IndoBERT Base and Multilingual BERT (M-BERT) Base Uncased. Each model underwent an experiment in different data splitting in this research, totaling four experiments. Hyper-parameter settings for IndoBERT and M-BERT include a learning rate of 3e-5 and a total of eight epochs for IndoBERT and ten epochs for M-BERT. The IndoBERT model was trained using the percentage split approach (see Figure 2) and K-fold Cross-Validation (see Figure 3), as shown below.

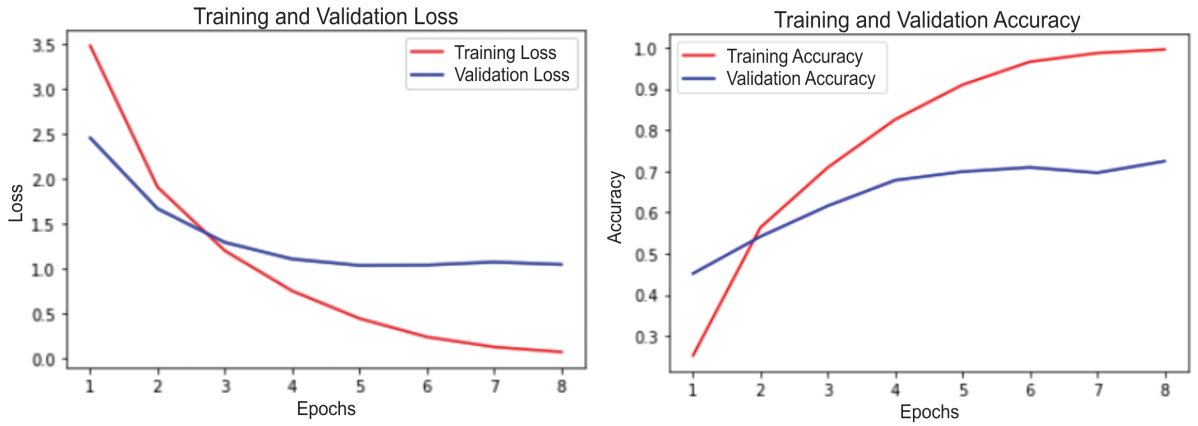

FIGURE 2. IndoBERT loss and accuracy plot

With the percentage split approach, IndoBERT had a higher training accuracy of **0.997** and a lower training loss of **0.062** with lower validation accuracy of 0.738 and a higher validation loss of 0.992. However, the K-fold Cross-Validation method had a higher validation accuracy of **0.742** and a lower validation loss of **0.903** even when the training accuracy had a 0.006 lower difference.

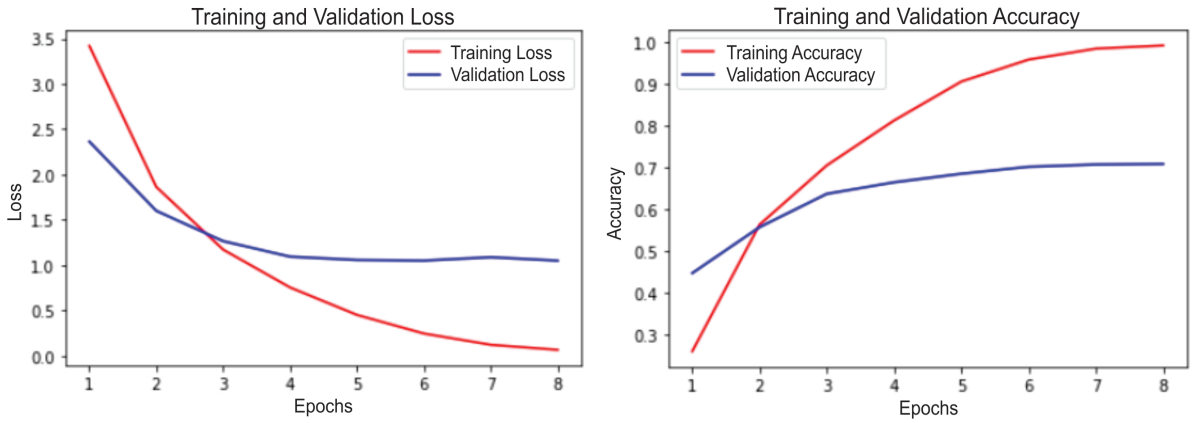

FIGURE 3. IndoBERT with K-fold loss and accuracy plot

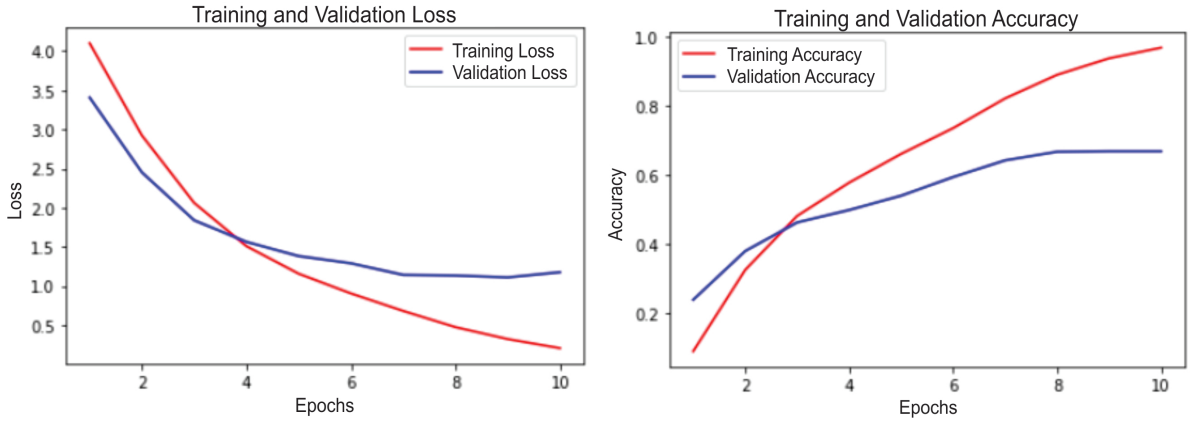

FIGURE 4. M-BERT with K-fold loss and accuracy plot

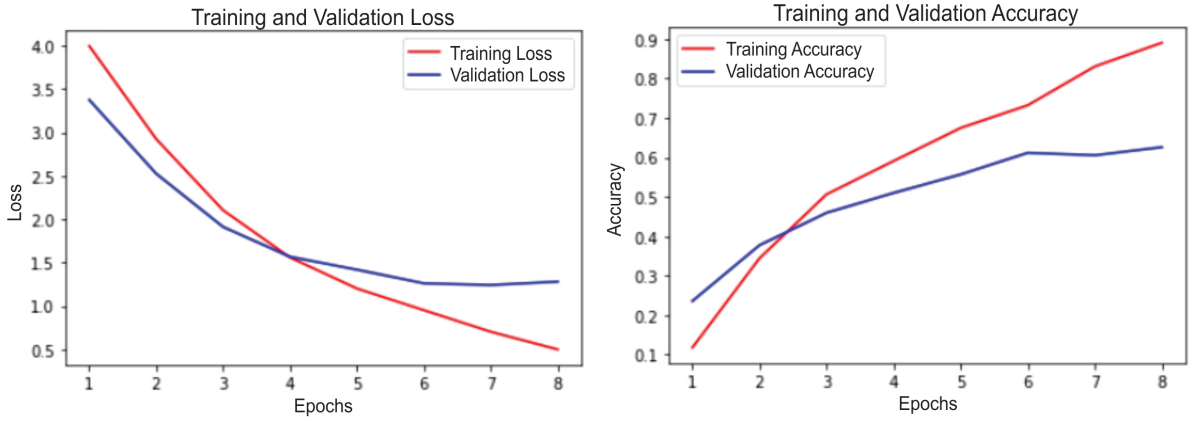

FIGURE 5. M-BERT loss and accuracy plot

Next are the results of the M-BERT model training using the percentage split method (see Figure 5) and K-fold Cross-Validation (Figure 4). The M-BERT model with K-fold Cross-Validation generally outperforms the percentage split approach. The results with K-fold Cross-Validation are training accuracy of **0.963**, training loss of **0.206**, validation accuracy of **0.682**, and slightly larger validation loss with a final value of 1.192. While with the percentage split approach, the training accuracy is 0.959, training loss of 0.235, validation accuracy of 0.675, and validation loss is **1.121**, as shown in Table 1.

Following the training phase, the model is evaluated with a prediction test, with results in prediction test accuracy derived from classification report metrics and Top-K

Table 1. Result

Method	Training accuracy	Training loss	Validation accuracy	Validation loss	Test Predict Accuracy	Top-K Accuracy
IndoBERT (8 Epoch)	**0.997**	**0.062**	0.738	0.992	0.70	0.84
IndoBERT with K-fold	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 K-fold	0.963	0.206	0.682	1.192	0.63	0.81

Accuracy metrics. From the evaluation phase, the IndoBERT model with the K-fold Cross-Validation approach emerged as the best model, with a Test Predict Accuracy of 0.74 and a Top-K Accuracy score of 0.86. However as shown in Table 1, the model has a relatively high validation loss result. It can be concluded that there is overfitting in the model trained. According to data analysis, it takes a larger quantity of article data to overcome these issues, which will be managed in the future work.

5. **Conclusion.** The IndoBERT model with the K-fold Cross-Validation approach has the most remarkable accuracy based on the experimental data. Although the training accuracy is relatively good and the training loss is quite low, the validation loss value is relatively high, indicating that the model is still experiencing overfitting. Because there is not enough data compared to the total number of labels, the number of data per label will increase in future studies. In addition, data augmentation will be used to address the existing overfitting problem. Moreover, several state-of-the-art deep learning architectures and graph deep learning architectures will be implemented to explore the AA task modeling. Data for the implementation of authorship attribution has also been obtained through this research from news articles, with 4,037 articles with a total of 80 writers and the addition of one label, non-author. Our dataset is available to the research community by contacting the corresponding author. The number of the datasets will increase over time as we are still collecting the dataset.

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] M. Bouville, Plagiarism: Words and ideas, *Science and Engineering Ethics*, vol.14, no.3, pp.311-322, 2008.
[6] C. Juditha, Symbolic interaction in the anti-hoax virtual community to reduce the spread of hoax, *Research Journal of Communication and Development*, vol.19, no.1, pp.17-32, 2018.
[7] J. Hitschler, E. V. D. Berg and I. Rehbein, Authorship attribution with convolutional neural networks and POS-eliding, *Proc. of the Workshop on Stylistic Variation (EMNLP2017)*, Copenhagen, Denmark, pp.53-28, 2017.
[8] T. Solorio, S. Pillay, S. Raghavan and M. Montes, Modality specific meta features for authorship attribution in web forum posts, *Proc. of the 5th International Joint Conference on Natural Language Processing*, pp.156-164, 2011.

[9] 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.

[10] 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.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser and I. Polosukhin, Attention is all you need, *Advances in Neural Information Processing Systems*, 30, 2017.

[12] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol.1, pp.4171-4186, 2019.

[13] B. Wilie, K. Vincentio, G. I. Winata, S. Cahyawijaya, X. Li, Z. Y. Lim, S. Soleman, R. Mahendra, P. Fung, S. Bahar and A. Purwarianti, IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding, *Proc. of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, Suzhou, China, pp.843-857, 2020.

[14] W. Wongso, H. Lucky and D. Suhartono, Pre-trained transformer-based language models for sundanese, *Journal of Big Data*, vol.9, 39, DOI: 10.1186/s40537-022-00590-7, 2022.

[15] P. Rust, J. Pfeiffer, I. Vulić, S. Ruder and I. Gurevych, How good is your tokenizer? On the monolingual performance of multilingual language models, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, vol.1, pp.3118-3135, 2021.

[16] D. Q. Nguyen and A. T. Nguyen, PhoBERT: Pre-trained language models for Vietnamese, *Findings of the Association for Computational Linguistics: EMNLP2020*, pp.1037-1042, 2020.

[17] S. Bird, R. Dale, B. J. Dorr, B. Gibson, M. T. Joseph, M.-Y. Kan, D. Lee, B. Powley, D. R. 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)*, Marrakech, Morocco, 2008.

[18] B. Klimt and Y. Yang, The Enron corpus: A new dataset for email classification research, *European Conference on Machine Learning*, pp.217-226, 2004.

[19] Y. Seroussi, I. Zukerman and F. Bohnert, Authorship attribution with topic models, *Computational Linguistics*, vol.40, no.2, pp.269-310, 2014.

[20] 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.

[21] T. Pires, E. Schlinger and D. Garrette, How multilingual is multilingual BERT?, *Proc. of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, pp.4996-5001, 2019.

[22] P. Refaeilzadeh, L. Tang and H. Liu, Cross-validation, *Encyclopedia of Database Systems*, vol.5, pp.532-538, 2009.

[23] P. Refaeilzadeh, L. Tang and H. Liu, On comparison of feature selection algorithms, *Proc. of AAAI Workshop on Evaluation Methods for Machine Learning II*, vol.3, 2007.