IJCCS (Indonesian Journal of Computing and Cybernetics Systems)
Vol.17, No.1, January 2023, pp. 23~32
ISSN (print): 1978-1520, ISSN (online): 2460-7258
DOI: https://doi.org/10.22146/ijccs.79623

Unsupervised Text Style Transfer for Authorship Obfuscation in Bahasa Indonesia

Yunita Sari*[1], Fadhlan Pasyah Al Faridzi Mail[2]
[1]Bachelor Program of Computer Science, FMIPA UGM, Yogyakarta, Indonesia
[2]Department of Computer Science and Electronics, FMIPA UGM, Yogyakarta, Indonesia
e-mail: ***[1]yunita.sari@ugm.ac.id**, [2]pasyahfadhlan@gmail.com

Abstrak

Authorship attribution adalah salah satu NLP task yang berfungsi untuk mengidentifikasi penulis berdasarkan analisis stylometri. Di sisi lain, terdapat task lain yaitu authorship obfuscation dengan tujuan untuk melindungi privacy atau anonimitas penulis sehingga tidak bisa dideteksi oleh authorship attribution sistem. Tantangan utama pada authorship obfuscation adalah bagaimana memodifikasi teks tanpa mengubah semantic dari teks itu sendiri. Pada riset ini, kami mengaplikasikan metode text style transfer untuk memodifikasi gaya penulisan dari sebuah teks, dengan tetap menjaga semantik dari teks tersebut. Kami mengimplementasikan dua unsupervised text style transfer yaitu metode dictionary-based translation dan back-translation untuk mengubah level formalitas dari teks (dari informal ke formal). Hasil eksperimen menunjukkan bahwa metode back-translation mempunyai performa yang lebih baik dari metode dictionary-based. Dari dimensi safety, penurunan performa authorship attribution untuk metode back-translation lebih besar dibandingkan penurunan performa pada metode dictionary-based. Selain itu, pada dimensi soundness dan sensibleness, bisa dilihat bahwa metode back-translation mampu memodifikasi teks tanpa mengubah semantic dari teks tersebut. Kualitas bahasa yang dihasilkan terdengar natural dan tidak menimbulkan kecurigaan bahwa teks tersebut telah dimodifikasi.

Kata kunci*— authorship obfuscation, style transfer, formality*

Abstract

Authorship attribution is an NLP task to identify the author of a text based on stylometric analysis. On the other hand, authorship obfuscation aims to protect against authorship attribution by modifying a text's style. The main challenge in authorship obfuscation is how to keep the content of the text despite the text modification. In this research, we are applying text style transfer methods for modifying the writing style while preserving the content of the input text. We implemented two unsupervised text style transfers: dictionary-based and back-translation methods to change the formality level of the text. Experiment results show that the back-translation method outperformed the dictionary-based method. The authorship attribution performance decreased to 16.15% and 23.66% on F1-score for 3 and 10 authors, respectively, using back-translation. While for the dictionary-based method, the F1-score dropped to 1.99% and 11.56% for 3 and 10 authors, respectively. Evaluation of sensibleness and soundness factors show that the back-translation method can preserve the semantics of the obfuscated texts. Moreover, the modified texts are well-formed and inconspicuous.

Keywords—*authorship obfuscation, style transfer, formality*

Received November 30th,2022; Revised January 29th, 2023; Accepted January 30th, 2023

1. INTRODUCTION

Authorship attribution can be defined as identifying the author of a text [1]. This task is part of stylometry, the study of style originally applied to handwritten texts. Current stylometry works focus more on digital documents and code [2][3]. Research in authorship attribution proposed different approaches and architectures, from traditional methods using an array of features such as function words and character n-grams [4] to more advanced techniques using deep learning [5]. Authorship attribution has been widely used in many authorship analysis applications such as forensic investigation, plagiarism detection, history, and literary science [6][7].

Authorship attribution is undeniably playing a significant role in many applications. However, Mansoorizadeh et al. [8] argue that it can threaten privacy and freedom of speech. Another stylometry task called authorship obfuscation aims to prevent the author de-anonymization by perturbing the author's writing style in a given text. Previous research works have explored distinct types of obfuscation techniques. Manual obfuscation methods were used by identifying and modifying the most revealing words and phrases from an author's writing. Moving towards full automation, several works developed a rule-based method. Moreover, heuristic-based search and machine learning-based algorithms have also been explored. The main challenge in authorship obfuscation is modifying a given text while preserving the semantics. In addition, the obfuscation method needs to consider the readability and paraphrasing quality of the obfuscated text. Another challenge, especially in machine learning-based techniques, is to provide large amounts of previously written text by authors as training data. The performance of the authorship obfuscation model is commonly evaluated in three dimensions: safety, soundness, and sensibleness. The obfuscation model is considered safe if the forensic analysis does not reveal the original author of its obfuscated text. Soundness measures whether the obfuscated text is textually entailed with the original text. Moreover, the obfuscated text is sensible when the texts are well-formed and inconspicuous [9].

Text Style Transfer (TST) is one of the tasks in NLP that aims to change the input text's style attributes while maintaining the document's semantics. Previous studies conducted in TST with different style features, including formality, politeness, gender, humour and romance, biasedness, toxicity, authorship, simplicity, and engagingness [10]. The purpose of style transfer is in line with the purpose of authorship obfuscation, which is to change the style while maintaining the content. TST approaches are generally divided based on the availability of parallel data. Sequence-to-sequence (seq2seq) models with encoder-decoder architecture are mostly adopted when a parallel corpus is provided. However, in most cases, parallel data is hard to be obtained. Thus, unsupervised approaches have become a prolific research area in TST. There are three types of techniques for non-parallel data: disentanglement, prototype editing and pseudo-parallel corpus construction. There are limited numbers of TST research works in Bahasa Indonesia. One was exploring formality style transfer using iterative forward translation [11]. Several approaches were implemented, including dictionary-based translation, phrase-based statistical (PBSMT) machine translation, neural machine translation and pretrained language modelling. Results show PBSMT outperformed other methods followed by a dictionary-based translation. This work motivates us to adopt the TST for authorship obfuscation.

In this paper, we applied prototype-editing text style transfer to obfuscate the original text. The style transfer was done by changing the formality level of the text (from informal to formal) using two different approaches: dictionary-based and back-translation methods. The obfuscated texts were evaluated in three dimensions: safety, soundness, and sensibleness. The safety evaluation is performed against the authorship attribution model. While soundness and sensibleness are measured using human evaluation.

2. METHODS

In Figure 1, it can be seen the methodological steps in our experiments. The original texts are modified using two unsupervised text style transfer methods, resulting in the obfuscated text. Evaluation will be carried out in three dimensions: safety, soundness, and sensibleness. The details of each model and the evaluation dimensions are described in the following subsections.

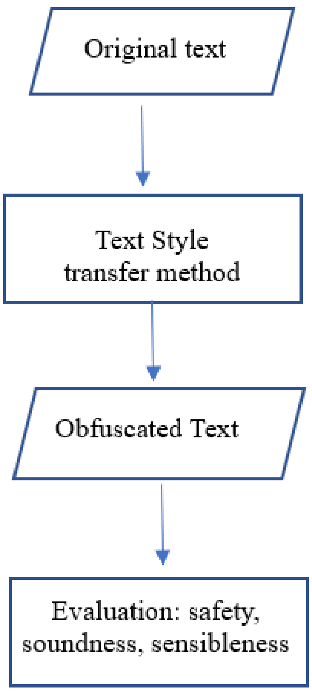

Figure 1. The proposed methodology for authorship obfuscation

2.1 Data

The data that will be processed and discussed in this paper are gathered using Tweepy, a tool for tweet scrapper. Different usernames are defined manually for the tweet-scrapping process. Only native Indonesian users were selected for this experiment. We kept any foreign language or slang words that might exist in the tweets (if it is alphanumeric), accounting that some authors preferred to write multilingually. Yet, we tried to obtain tweets from Indonesian users (judging by most of their tweet language and nationality) to have the dataset mainly consisting of Indonesian tweets.

The constructed dataset is being lightly pre-processed to preserve writing style originality in text yet formatted enough to be the as demanded. This alteration includes non-alphabetic language removal (such as Arabic), author-signature text removal (in-text name or email), and template-string removal (URLs, Quote Retweet text template, hashtags, and mentions). Emoticons, intentional mistakes in the typing, numbers, and symbols are not removed due to how closely related these contents are with the author's writing style, thus, making them the most important features.

In this paper, two different settings or slices of data are being compared to show the significance of author number regarding author attribution/prediction accuracy. The three authors' data comprised 7736 training tweets and 1935 test tweets. Each author on this data slice has 2578 documents on average in training data and 645 in testing data. Each tweet in this data slice roughly consists of 5 words length tweet, with most of them having 33 characters long. In comparison, the ten authors' data has 16953 training tweets and 4239 test tweets with 1695 documents per author on average. This data slice consists of an average seven-word tweet, roughly 48 characters long each.

2. 2 Authorship Attribution Model

First, we developed our authorship attribution model that will be used to evaluate the safety dimension of the obfuscation techniques. In this experiment, we used IndoBERTweet [12] for the word representations. IndoBERTweet is a large-scale pretrained model for Indonesian Twitter, which is the extension of the IndoBERT model with the addition of domain-specific vocabulary. IndoBERTweet was trained on 26M tweets with 409M word tokens. The training data was collected by crawling Indonesian tweets from December 2019 to December 2020, covering four topics: economy, health, education, and government. In our experiment, we used IndoBERTweet as the word representation only and passed the embedding to Support Vector Machine (SVM) classifier.

SVM is a supervised machine learning method which uses hyperplane to separate classes. We used SVM implementation from the Scikit-learn library and set the parameter C to 10, gamma to 0.01 and rbf as the kernel. The performance of the authorship attribution model is evaluated using four-standard metrics for the classification problem: accuracy, precision, recall and F1-measure.

2. 3 Authorship Obfuscation

As mentioned before, the main factor in developing the authorship obfuscation model is modifying text while preserving the semantics. The most common approach in authorship obfuscation is by changing the author's writing style. In this paper, we adopt the unsupervised text style transfer, which modifies the formality level of the text. There are two methods used: dictionary-based translation and back-translation methods.

2.3.1 Dictionary-based Translation

The idea in dictionary-based translation is to replace the informal/slang word in a text with its formal form. An Indonesian informal-formal dictionary consisting of 1043-word pairs was used for this experiment. This model translates an informal word in the text to its formal form if it appears in the dictionary. We used the dictionary in this link: https://github.com/louisowen6/NLP_bahasa_resources/blob/master/combined_slang_words.txt Table 1 shows a sample of informal words and their formal form.

Table 1. Example of informal-formal word pairs

Informal Word	Formal Word
kalo	kalau
nganggur	tidak punya pekerjaan
ortu	orang tua
rhs	rahasia
brb	segera kembali
kere	tidak punya uang
xixixixi	tertawa

Table 2 shows that some words in the translated text are modified into formal form. However, since only some slank/informal words are listed in the dictionary, some slank/informal words still need to be translated.

Table 2. Example of translated texts using dictionary-based method

Original Text	Translated text
Pernah ngga sih tiba-tiba kalian ngerasa sedih ngga jelas pengen nangis dan ngerasa hampa kayak ngga punya siapa-siapa?	pernah tidak sih tiba-tiba kalian ngerasa sedih tidak jelas pengen menangis dan ngerasa hampa kayak tidak punya siapa-siapa?
Alhamdulillah kasus Jiwasraya udah ada yg ditangkep, duitnya balik gak ya?	alhamdulillah kasus jiwasraya sudah ada yang ditangkap, duitnya kembali tidak ya?
oppp dah mulai ni pembangunannya	oppp sudah mulai ini pembangunannya
ngga gitu	tidak gitu
...setelah ngaca baru ketauan dah, ternyata hr ini saya belum bercukur...🤓	...setelah ngaca baru ketahuan dah, ternyata hari ini saya belum bercukur...🤓

2.3.2 Back-Translation (Indonesian-English-Indonesian)

In the back-translation method, we translated the original text in Indonesian into English and translated it back to Indonesian. The back-translation method is commonly used to generate synthetic data when the number of training data is limited. For this purpose, we used Googletrans, a free Python library that implemented Google Translate API. This library has features, including auto language detection, bulk translation, fast and reliability. The translated samples are described in Table 3. Compared to dictionary-based translation, the translated text from the back-translation method sounds more formal. However, based on our observations, some translated texts have different semantics than the original ones.

Table 3. Example of translated texts using back-translation method

Original Text	Translated text
Pernah ngga sih tiba-tiba kalian ngerasa sedih ngga jelas pengen nangis dan ngerasa hampa kayak ngga punya siapa-siapa?	Pernahkah Anda tiba-tiba merasa sedih, tidak ingin menangis dan merasa kosong seperti tidak memiliki siapa-siapa?
Alhamdulillah kasus Jiwasraya udah ada yg ditangkep, duitnya balik gak ya?	Alhamdulillah kasus Jiwasraya sudah ditangkap, uangnya kembali?
oppp dah mulai ni pembangunannya	oposisi telah memulai konstruksi
ngga gitu	bukan seperti itu
...setelah ngaca baru ketauan dah, ternyata hr ini saya belum bercukur...🤓	...setelah melihat-lihat, saya baru tahu, ternyata hari ini saya belum bercukur...🤓

2. 4 Evaluation

Evaluation of authorship obfuscation is carried out in three dimensions: safety, soundness, and sensibleness. The safety dimension is evaluated against the authorship attribution model in 2.2, while soundness and sensibleness are measured using human evaluation. For human evaluation, we randomly selected 200 samples of translated texts from both methods. For every 100 samples, we asked four native Indonesians to rate each translated text's soundness and sensibleness with a score from 1 to 5. Thus, there are a total of eight people involved in the human evaluation. If there is no modification in the translated text, a score of 0 is assigned. The obfuscation process is applied to test data only.

2.4.1 Safety

The safety dimension is evaluated by applying the authorship attribution model to predict the author of the obfuscated text. The obfuscation method is successful if the attribution model cannot identify the real author of the obfuscated text. It can be seen by the decreasing of authorship attribution performance in obfuscated text. The safety dimension is measured based on four different evaluation metrics: accuracy, precision, recall, and F1-score.

$$accuracy = \frac{tp + tn}{tp + tn + fp + fn} \quad (1)$$

$$precision = \frac{tp}{tp + fp} \quad (2)$$

$$recall = \frac{tp}{tp + fn} \quad (3)$$

$$F1 - score = \frac{2 \times precision \times recall}{precision + recall} \quad (4)$$

Where *tp* is true positive, *tn* is true negative, *fn* is false negative and *fp* is false positive.

2.4.2 Soundness

Suppose the obfuscated text (modified text) is the results from paraphrasing the original text. This can be seen by looking at the modified text semantics, which are still the same as the original text, even though the text has been modified. We set scoring 1 up to 5 with the following details:

- A text is given a score of 5 if the semantics of the modified text has stayed the same as the original text.
- A text is given a score of 1 if the modified text's semantics are different from the original text.
- Scores 2, 3, and 4 are given if the semantic differences are between 1 to 5.

2.4.3 Sensibleness

If the obfuscated text (modified text) still sounds/feels natural, there is proper word choice, no grammatical errors, and the content is understandable. In addition, changes in the text do not raise suspicions that the text has been modified before.

- A text is given a score of 5 if the modified text is immediately understandable, has no grammatical errors, the sentences sound natural and do not raise suspicions that the text has been modified.
- A text is given a score of 1 if the content of the modified text is difficult to understand, there are many grammatical errors, sentences sound unnatural and raise suspicions that the text has been modified.
- Scores 2, 3, and 4 are given if the grammatical quality is between 1 to 5.

3. RESULTS AND DISCUSSION

We performed two experiment settings for the authorship attribution using 3 and 10 authors. As can be seen in Table 4, the authorship attribution model achieved an F1-score of 80.97% and 64.89% for 3 and 10 authors, respectively. This performance was achieved on the original test data without any modification. To evaluate the safety dimension, we applied the same authorship attribution model on modified test data for both dictionary-based and back-translation methods. Results in Table 4 show that the back-translation method outperformed the dictionary-based method. The authorship attribution performance decreased to 16.15% and 23.66% on F1-score for 3 and 10 authors, respectively, using back-translation. While for the dictionary-based method, the F1-score dropped to 1.99% and 11.56% for 3 and 10 authors, respectively.

Table 4. Performance of authorship attribution on different translated text (safety dimension)

3 Authors				
Method	Accuracy	Precision	Recall	F1-Score
No modification	81.03	81.03	81.05	80.97
Dictionary-based	79.06	79.19	79.09	78.98
Back Translation	64.90	65.71	64.94	64.82
10 Authors				
Method	Accuracy	Precision	Recall	F1-Score
No modification	66.59	63.19	68.27	64.89
Dictionary-based	53.13	51.85	63.32	53.33
Back Translation	41.97	41.86	42.66	41.23

In addition to evaluating the safety dimension, we also evaluated soundness and sensibleness dimensions. Table 5 shows the average soundness and sensibleness scores for each method. From the Table, we can see that back translation has better soundness and sensibleness scores than the dictionary-based method. These results reflected that the semantics of the obfuscated texts can still be preserved using the back-translation method. In addition to that, the modified texts are well-formed and inconspicuous.

Table 5. average soundness and sensibleness scores for each method

Human evaluator	Dictionary-based		Back translation	
	Avg soundness score	Avg sensibleness score	Avg soundness score	Avg sensibleness score
Evaluator 1	2.4	1.91	3.52	3.94
Evaluator 2	2.23	2.18	3.55	4
Evaluator 3	2.21	2.07	3.57	3.81
Evaluator 4	2.26	2.15	3.48	4
Evaluator 5	2.01	2.04	3.26	4.05
Evaluator 6	1.73	2.49	2.79	3.71
Evaluator 7	2.63	2.54	3.03	3.89
Evaluator 8	2.61	1.55	2.45	3.45
Average	**2.26**	**2.12**	**3.21**	**3.85**

Based on our observation, there are a significant number of original texts that still need to be identified in the dictionary-based method. For these cases, evaluators assigned a 0 score for both soundness and sensibility dimensions. From 200 samples, 81 texts were unmodified in the dictionary-based method, and only 8 were unmodified in the back-translation method. These results explained why the soundness and sensibleness scores in back-translation are higher than in dictionary-based methods. Table 6 shows sample texts with zero soundness/sensibleness scores in the dictionary-based method and its translated texts using the back-translation method. The back-translation method can change the formality style of the text (from informal to formal). However, in some cases, the semantics of the original text must be preserved.

Table 6. Sample texts with zero soundness/sensibleness score in dictionary-based method

Original text	Dictionary based method	Back translation method
semoga besok-besok beda, ya	semoga besok-besok beda, ya	Saya harap besok akan berbeda, oke?
lagi di kampung?	lagi di kampung?	lagi di desa?
1 Suro sama 1 Muharam sama ya?	1 suro sama 1 muharam sama ya?	1 Suro sama dengan 1 Muharram, kan?
Takut nanti jadi penyakit hati	takut nanti jadi penyakit hati	Takut jadi penyakit jantung
Met mlm mbaa	met mlm mbaa	Bertemu mlm mbaa
bukan pernah, tapi sedang	bukan pernah, tapi sedang	tidak pernah, tapi saat ini
Aku sih gini juga	aku sih gini juga	Saya juga seperti ini
Anak terakhir juga😌	anak terakhir juga😌	Anak terakhir juga😌
paling suka sama lagu yang mana?	paling suka sama lagu yang mana?	lagu apa yang paling kamu suka?
sinyal bermasalah nih, dengan terpaksa space-nya harus berakhir □	sinyal bermasalah nih, dengan terpaksa space-nya harus berakhir □	Ada masalah dengan sinyal, ruang harus berakhir
secret forest seru. atau kalau mau yang rada ringan alias ada romance-nya touch your heart asik juga.	secret forest seru. atau kalau mau yang rada ringan alias ada romance-nya touch your heart asik juga.	hutan rahasia itu menyenangkan. atau kalau mau yang sedikit lebih ringan alias ada romancenya menyentuh hati, asik juga.

In the back-translation method, some samples were given a score of 5 for both soundness and sensibleness. Table 7 shows some examples. The translation quality in the back-translation method is better than the dictionary-based method. The grammatical errors are minimal, and the modified text sounds natural; thus, it will not raise suspicions that the text has been modified.

Table 7. Sample texts with perfect soundness/sensibleness score (score of 5) in back-translation method

Original text	Dictionary based method	Back translation method
nah jam2 segini enakan nih, gak ruwet 🤓	nah jam2 segini enakan nih, tidak ruwet 🤓	Nah, bagusnya jam segini, gak ribet
oalah sori, ternyata yg minggu jam 8 malam gaes 😁	oalah sori, ternyata yang minggu jam 8 malam teman-teman 😁	Eh maaf ternyata hari minggu jam 8 malam guys
maksudnya yg ini yg gak perlu diganti 🤓	maksudnya yang ini yang tidak perlu diganti 🤓	Maksud saya yang ini tidak perlu diganti
Ngirim foto kesini belum bisa ya...?	ngirim foto kesini belum bisa ya...?	Belum bisa kirim foto kesini...?
yang bikin kesal, buang saja dari hati dan pikiran.	yang bikin kesal, buang saja dari hati dan pikiran.	apa yang membuatmu kesal, buang saja dari hati dan pikiranmu.

4. CONCLUSIONS

Our experiments show that unsupervised text style transfer (informal to formal) is effective for authorship obfuscation. The back-translation method outperformed the dictionary-based method in all evaluated dimensions. The language quality of the modified text from the back-translation method is like the original text. In addition to that, the modification does not raise suspicion. For future works, we are interested in exploring different style features, such as politeness, gender, humour and romance, biasedness, toxicity, authorship, simplicity, and engagingness.

ACKNOWLEDGEMENT

This research was supported by the research grant from the Faculty of Mathematics and Natural Sciences, Universitas Gadjah Mada, namely Hibah Penelitian Fakultas, Dana Masyarakat Tahun 2022 Number 89/ J01.1.28/PL.06.02/2022

REFERENCES

[1] Swinson, T. and Reyna, C. (2013). Authorship Attribution Using Stopword Graphs. pages 1-9.

[2] Shrestha, P., Sierra, S., Gonzalez, F., Montes, M., Rosso, P., and Solorio, T. (2017). Convolutional neural networks for authorship attribution of short texts. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 669-674, Valencia, Spain. Association for Computational Linguistics.

[3] Caliskan-Islam, A., Harang, R., Liu, A., Narayanan, A., Voss, C., Yamaguchi, F., and Greenstadt, R. (2015). De-anonymizing programmers via code stylometry. In *Proceedings of the 24th USENIX Conference on Security Symposium*, SEC'15, pages 255{270, Berkeley, CA, USA. USENIX Association.

[4] Stamatatos, E. (2013). On the Robustness of Authorship Attribution Based on Character n-gram Features. *Journal of Law and Policy*, 21(2):421-439.

[5] Schwartz, R., Tsur, O., Rappoport, A., and Koppel, M. (2013). Authorship Attribution of Micro-Messages. In *2013 Conference on Empirical Methods in Natural Language Processing*, number October, pages 1880-1891, Seattle, USA.

[6] Burrows, J. (2002). 'Delta': A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, 17(3):267-287.

[7] Georgi Karadzhov, Tsvetomila Mihaylova, Yasen Kiprov, Georgi Georgiev, Ivan Koychev, and Preslav Nakov. 2017. The case for being average: A mediocrity approach to style masking and author obfuscation. In International Conference of the CrossLanguage Evaluation Forum for European Languages, pages 173–185. Springer.

[8] Mansoorizadeh, M., Rahgooy, T., Aminiyan, M. dan Eskandari, M., 2016, Author Obfuscation Using WordNet and Language Models, *CEUR Workshop Proceedings*, Évora.

[9] Paolo Rosso, Francisco Rangel, Martin Potthast, Efstathios Stamatatos, Michael Tschuggnall, and Benno Stein. Overview of PAN 2016–New Challenges for Authorship Analysis: Cross-genre Profiling, Clustering, Diarization, and Obfuscation. In Norbert Fuhr et al., editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction. 7th International Conference of the CLEF Initiative (CLEF 2016)*, volume 9822 of *Lecture Notes in Computer Science*, pages 518-538, September 2016. Springer

[10] Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2022. Deep Learning for Text Style Transfer: A Survey. *Computational Linguistics*, 48(1):155–205.
[11] Wibowo, Haryo & Prawiro, Tatag & Prasojo, Radityo & Mahendra, Rahmad. (2020). Semi-Supervised Low-Resource Style Transfer of Indonesian Informal to Formal Language with Iterative Forward-Translation.
[12] Fajri Koto, Jey Han Lau, and Timothy Baldwin. *IndoBERTweet: A Pretrained Language Model for Indonesian Twitter with Effective Domain-Specific Vocabulary Initialization*. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (**EMNLP 2021**), Dominican Republic (virtual).