Future Computer Science Journal (FCSJ)
Vol. 1, No. 2, December 2023: 53-57
http://asasijournal.id/index.php/fcsj
http://doi.org/10.xxxxx/fcsj.xxxx.xxx

Text Normalization on Indonesian-English Code-Mixed Twitter Text using UFAL ByT5

Rafi Dwi Rizqullah[1*], Indra Budi[1]
[1]Master of Computer Science, Faculty of Computer Science, Universitas Indonesia, Indonesia

Abstract
Social media has been grown rapidly in the global community. It also includes Twitter, which is getting increase in both users and content created. However, Twitter has character limit in one tweet which causes changes to the writing patterns of its users. Twitter users began to modify their writing from using formal words into non-formal words, one of which was using code-mixed language. For tweet analysis purposes, text normalization is required to transform non-formal words into formal ones to help analysis process. The recent state-of-the-art for Indonesian-English code-mixed Twitter text normalization is with statistical machine translation (SMT) models, however the SMT model still has weakness in word recognition. This research focuses on the Indonesian and English code-mixed Twitter text normalization using one of transformer model which is UFAL ByT5. There are two UFAL ByT5 models that were used, each of them are for Indonesian and English language. Research result shows that UFAL ByT5 model outperform SMT model on text normalization by 0.88 percent of BLEU score in difference.

Keywords:
Text normalization; Twitter; Code-mixed; Indonesian; English ; UFAL ByT5.

Article History:
Received: Dec 18, 2023
Revised: Dec 28, 2023
Accepted: Dec 28, 2023
Published: Dec 28, 2023

Corresponding Author:
Rafi Dwi Rizqullah
Master of Computer Science, Universitas Indonesia, Indonesia
Email: rafidwiriz@gmail.com

INTRODUCTION

Social media has been a place for people to interact with each other. The growing of social media can be seen from the growth of their total user [1]. Out of available social media, Twitter is one of them that is growing fast. On the second quarter of 2022, Twitter got 237.8 million in total monetized user [2].

Twitter is a social media that focuses on contents such as text called tweet. But users can post another content like pictures, videos, etc. Different from other social media, Twitter has character limit on a tweet, which is 280 characters, including links to another contents. That limitation caused users to change the form of formal words so that it's still within character limits, forming non-formal words. The characteristic found on some of Indonesian tweets are non-formal words and shortening [3], followed by code-mixed of foreign words.

Indonesia has been placed on fifth biggest social media users by 18.45 million users in total [4]. Even though, there is not much research that focuses on Indonesian code-mixed text. One of them was done by [5] to build a system that translate code-mixed Twitter text into Indonesian formal words. But that system has weakness on text normalization that only use word mapping and normalization rulesets.

Then, a research was done by [6] as an improvement from [5], one of them was to improve text normalization method by using statistical machine translation or SMT as a replacement for word mapping. The result shows that SMT with normalization rulesets could outperform method from [5]. But there is still some weakness. Based on SMT model that trained using dataset from [5], there are some words that cannot be normalized by model. Some of the non-formal word types according to [7] that cannot be normalized are phonetic change, shortening, and disemvoweling.

An improvement can be made to replace the SMT model with another solutions. The problems that exist on mentioned word types can be fixed by using model that has been trained using synthesis datasets, one of them was UFAL ByT5 model from [8]. Other than trained using synthesis datasets, UFAL ByT5 model has another advantage which is character-level tokenization.

Literature Review

Twitter is a social media platform that is based on microblogging [9]. Microblogging is a media platform [10] that let its users to share contents in small size [11]. On the launch of Twitter, it has limitation of 140 characters on one tweet, but doubled to 280 characters on 2017 [12], [13].

Code-mixed text is a text that contains two or more languages and are mixed non-uniformly. The code-mixed text phenomenon already happens on Indonesia. A research by [14] shows the Indonesian-English code-mixed phenomenon that happens on South Jakarta people in WhatsApp and Twitter. There are three code-mixed types that were defined in the research, they are intra sentential, intra lexical, and involving change of pronunciation.
Language identification is a document classification [15] task that consist of giving a document to class or category that is represented by limited sets of label, in this case is language labels [16]. Research of language identification are splits into two groups, they are language identification on whole document and identification on every words [17].

Based on KBBI or Kamus Besar Bahasa Indonesia, "normalisasi" (normalization) is an action to make thing normal again [18] while "normal" is (thing) in accordance with applicable rule or pattern. Based on those definitions, it can be said that "normalization" is a task or action that is done to change things in accordance with applicable rule or pattern, and text normalization can be defined as task to change text into normal form, in this case their canonical form.

There were researches about the type of changes that happen on canonical form of words based on some languages. In Indonesian language, there are type of changes based on [7] which are disemvoweling, shortening, space or dash removal, phonetic changes, informal affixation, compounding and acronym, reverse, loan words, and jargon. While for English language, there are also type of changes based on which are misspellings [19], phonetic substitutions, shortening, acronyms, slang, emphasis, and punctuation.

METHOD

This research was done by following steps that can be seen in Figure 1. Those steps are literature review, error analysis, method design, implementation and experiment, result analysis, and lastly conclusion.

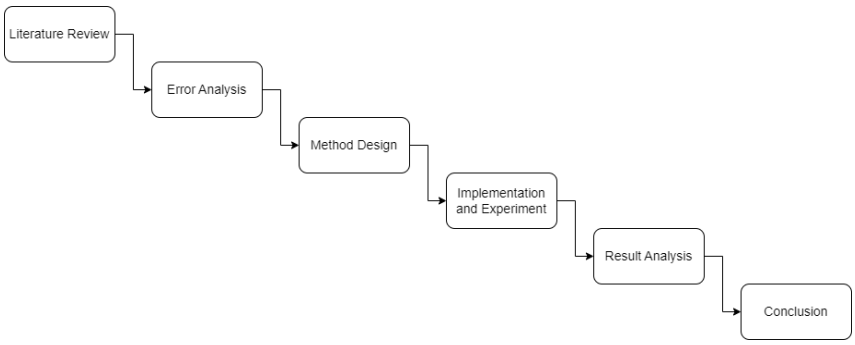

Figure 1. Flow Steps of Research

The proposed method design for this research can be seen in Figure 2. This design is similar to the design of [6] but without translation and emotion classification [20] module. There are three modules in this method, which are tokenization module, language identification module, and text normalization module.

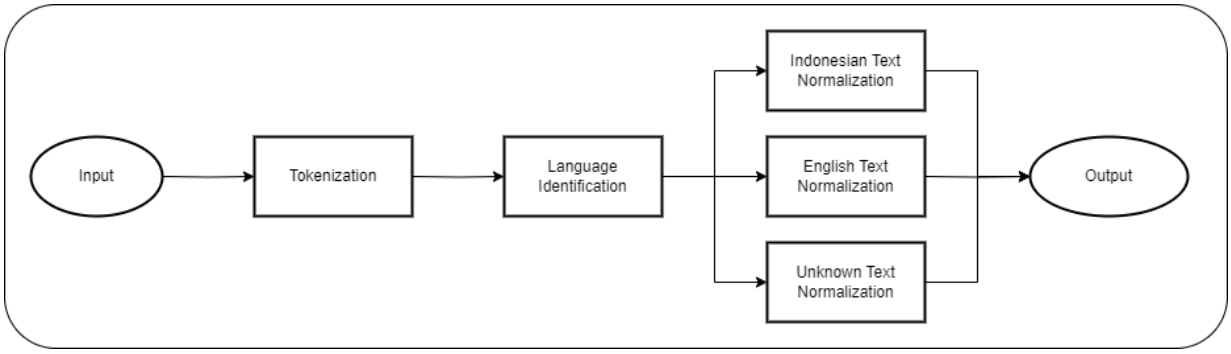

Figure 2. Design Flow of Proposed Method

The models used for the proposed method were applied to both language identification and text normalization modules. Model that was used for language identification is a pre-trained transformer model which is BERT [21]. The tool that was used for sequence labelling is MaChAmp [22] while the BERT model that was used is mBERT or BERT multilingual [23]. The model that was used for text normalization is like language identification ones which is using transformer model, but with more variant. For text normalization is using IndoBERT model [24] and UFAL ByT5 model [8]. While for pre-processing is using normalization ruleset from [5].

The dataset that was used for the method in the research is dataset from [5]. The dataset was consisting of 825 Indonesian Twitter text and containing the data such as original text, its tokens, language label of each token, and the canonical form of each token. Then, the dataset was split into three data, which are training, validation, and test data with 60:20:20 in proportion.

Experiment Scenario

There were two scenarios that were running on this research, which are language identification module scenario and end-to-end scenario. Both scenarios were done to test every module in the proposed method and to do ablation test to calculate error aggregation of whole method. For every scenario, there were two data variations. They are regular data variation, which uses token input data from tokenization module and combine it with language label data from dataset, and token + language label data variation, which uses both token data and language label data from dataset.

RESULTS AND DISCUSSION

The result for language identification module scenario with regular data variation is presented on Table 1. The result shows that mBERT model [21], [25] outperformed CRF model [5] on accuracy by the score of 91 percent, 2 percent more than CRF. Also, mBERT model outperformed on recall and F1-score with respectively 5 percent more and 3 percent more than CRF model scores.

Table 1. The Result for Language Identification Module Scenario With Regular Data Variation

Model	Accuracy	Precision	Recall	F1-score
CRF [5]	88.00 %	90.00 %	85.00 %	87.00 %
mBERT base uncased [21]	91.00 %	90.00 %	90.00 %	90.00 %

For the same scenario with other data variations can be seen on Table 2. The mBERT model [21] on token + language label data variation has outperformed the same model with regular data variation by 5 percent more on all metric scores.

Table 2. The Result for Language Identification Module Scenario With Other Data Variations

Model	Accuracy	Precision	Recall	F1-score
Token + Language Label Data Variation				
mBERT base uncased [21]	96.00 % (+5.00 %)	95.00 % (+5.00 %)	95.00 % (+5.00 %)	95.00 % (+5.00 %)

The result for end-to-end scenario with regular data variation is presented on Table 3. The result shows that the proposed method outperformed other methods on BLEU by the score of 85.03 percent. The score is 0.88 percent more than the method from [6] and 7.66 percent more than the method from [8] that also use UFAL ByT5 model.

Table 3. The Result for End-to-end Scenario With Regular Data Variation

Model	BLEU
The method from [6]	84.15 %
UFAL ByT5 from [8]	77.37 %
Proposed method	85.03 %

For the same scenario with other data variations can be seen on Table 4. The result shows the increase BLEU score on proposed method with token + language label data variation, by 6.39 percent more than regular data variation.

Table 4. The Result for End-to-end Scenario With Other Data Variations

Model	BLEU
Token + Language Label Data Variation	
Proposed method	91.42 % (+6.39 %)

On language identification module, there are some patterns found in the result of tokens that can be handled by mBERT model [21]. For Indonesian tokens, mBERT model can recognize Indonesian words while on full capital form which previously could not be handled well by CRF model [5], for example the words "DISURUH", "SALAH", "SUKA", etc. For English tokens, some patters that can be solved by mBERT model are regular English words, words with typo, and English words with Indonesian affixes such as words with "nge-" affix.

And then, there are patterns found in the result that could not be handled well by mBERT model [21]. Some of them are patterns that can't be handled both by mBERT model and CRF model [5] such as English words that got change of form, like "Pliss" which is "please" and "hellaw" which is "hello". Both recognized by models as Indonesian words. Lastly, some patterns can't be handled by mBERT model only such as Indonesian words that has repetitive letter such as "hayooooooooooooooo" that was recognized as English word or "yaa" that was recognized as unknown word.

Then, on text normalization module, some patterns were found in the result from proposed method. For Indonesian words, the patterns found on words changes based on [7]. The changes found were phonetic changes such as changes on one letter, like "malem" that has canonical form of "malam"; disemvoweling such as "mnrt" became "menurut"; and shortening such as "g" became "tidak". And for English words, the patterns found on words changes based on [26]. The changes found were misspelling such as words with Indonesian affixes like "nge-" in "ngeskip" became "skip"; and shortening such as words with "-s" suffix like "lets" became "let us".

Some factors that helped proposed method able to handle those patterns are synthesis dataset that was used to train UFAL ByT5 model able to simulate some type of word changes with probability that has been determined, and the model itself that able to normalize words that are not on training data before.

While there are patterns that can be handled by the proposed method, there are also patterns that cannot be handled by the method. For Indonesian words, the pattern found on disemvoweling words where model could not add proper vocal letters between words. For English words, it was found that there are result with repetitive words, such as "the 14th" has been normalized as "the 14th 14th".

CONCLUSION

The research has been done to propose a text normalization method using mBERT model [21] for language identification and UFAL ByT5 model [8] as text normalization. The result on end-to-end scenario shows that the proposed method outperformed the method from [6] with 85.03 percent on BLEU score, 0.88 percent.

REFERENCES

[1] M. R. Ningsih, K. A. H. Wibowo, A. U. Dullah, dan J. Jumanto, "Global recession sentiment analysis utilizing VADER and ensemble learning method with word embedding," *J. Soft Comput. Explor.*, vol. 4, no. 3, 2023, doi: https://doi.org/10.52465/joscex.v4i3.193.

[2] M. A. Rizaty, "Pengguna Aktif Twitter Global Capai 830 Juta per Kuartal II/2022," *DataIndonesia.id*, 2022. https://dataindonesia.id/internet/detail/pengguna-aktif-twitter-global-capai-830-juta-per-kuartal-ii2022.

[3] A. F. Hidayatullah, "Language tweet characteristics of Indonesian citizens," *Int. Conf. Sci. Technol.*, 2015, doi: 10.1109/TICST.2015.7369393.

[4] C. M. Annur, "Pengguna Twitter Indonesia Masuk Daftar Terbanyak di Dunia, Urutan Berapa?," *databoks*, 2022. https://databoks.katadata.co.id/datapublish/2022/03/23/pengguna-twitter-indonesia-masuk-daftar-terbanyak-di-dunia-urutan-berapa.

[5] A. M. Barik, R. Mahendra, dan M. Adriani, "Normalization of Indonesian-English Code-Mixed Twitter Data," *Assoc. Comput. Linguist.*, 2019, doi: 10.18653/v1/D19-5554.

[6] E. Yulianti, A. Kurnia, M. Adriani, dan Y. S. Duto, "Normalisation of Indonesian-English Code-Mixed Text and its Effect on Emotion Classification," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 11, 2021, doi: 10.14569/IJACSA.2021.0121177.

[7] H. A. Wibowo *et al.*, "IndoCollex: A Testbed for Morphological Transformation of Indonesian Colloquial Words," *Assoc. Comput. Linguist.*, hal. 3170–3183, 2021, doi: 10.18653/v1/2021.findings-acl.280.

[8] D. Samuel dan M. Straka, "ÚFAL at MultiLexNorm 2021: Improving Multilingual Lexical Normalization by Fine-tuning ByT5," *Comput. Lang.*, 2021.

[9] B. Singh dan D. K. Sharma, "SiteForge: Detecting and localizing forged images on microblogging platforms using deep convolutional neural network," *Comput. Ind. Eng.*, vol. 162, 2021, doi: https://doi.org/10.1016/j.cie.2021.107733.

[10] S. Dutta, A. K. Das, S. Ghosh, dan D. Samanta, "Chapter 1 - Introduction to microblogging sites," in *Data Analytics for Social Microblogging Platforms*, 2023, hal. 3–38.

[11] A. M. Kaplan dan M. Haenlein, "The early bird catches the news: Nine things you should know about micro-blogging," *Bus. Horiz.*, vol. 54, no. 2, hal. 105–113, 2011, doi: https://doi.org/10.1016/j.bushor.2010.09.004.

[12] A. B. Boot, E. T. K. Sang, K. Dijkstra, dan R. A. Zwaan, "How character limit affects language usage in tweets," *Humanit. Soc. Sci. Commun.*, 2019, doi: 10.1057/s41599-019-0280-3.

[13] J. Jumanto, M. A. Muslim, Y. Dasril, dan T. Mustaqim, "Accuracy of Malaysia Public Response to Economic Factors During the Covid-19 Pandemic Using Vader and Random," *J. Inf. Syst. Explor. Res.*, vol. 1, no. 1, hal. 49–70, 2023.

[14] J. Jimmi dan R. E. Davistasya, "Code-mixing in language style of south jakarta community indonesia," *J. English Educ. Appl. Linguist.*, vol. 8, 2019, doi: http://dx.doi.org/10.24127/pj.v8i2.2219.

[15] R. Ramezani, "A language-independent authorship attribution approach for author identification of text documents," *Expert Syst. Appl.*, vol. 180, 2021, doi: https://doi.org/10.1016/j.eswa.2021.115139.

[16] M. Zampieri, "Chapter 8 - Automatic Language Identification," *Work. with Text*, hal. 189–208, 2016, doi: https://doi.org/10.1016/B978-1-84334-749-1.00008-1.

[17] A. Kurnia, E. Yulianti, D. Chahyati, I. Budi, dan W. C. Wibowo, "Normalisasi teks code-mixed bahasa Indonesia-Inggris pada data twitter dan analisis pengaruhnya untuk klasifikasi emosi = Code-mixed text normalization on Indonesian-English language on twitter data and the analysis of its effect on emotion classification," *Univ. Indones.*

[18] "normalisasi," *KBBI Daring*, 2016. .

[19] K. H. Lai, M. Topaz, F. R. Goss, dan L. Zhou, "Automated misspelling detection and correction in clinical free-text records," *J. Biomed. Inform.*, vol. 55, hal. 188–195, 2015, doi: https://doi.org/10.1016/j.jbi.2015.04.008.

[20] S. P. Mishra, P. Warule, dan S. Deb, "Improvement of emotion classification performance using multi-resolution variational mode decomposition method," *Biomed. Signal Process. Control*, vol. 89, 2024, doi: https://doi.org/10.1016/j.bspc.2023.105708.

[21] J. Devlin, M.-W. Chang, K. Lee, dan K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv*, 2018, doi: https://doi.org/10.48550/arXiv.1810.04805.

[22] R. van der Goot, A. Üstün, A. Ramponi, I. Sharaf, dan B. Plank, "Massive Choice, Ample Tasks (MaChAmp): A Toolkit for Multi-task Learning in NLP," *Assoc. Comput. Linguist.*, hal. 176–197, 2021, doi: 10.18653/v1/2021.eacl-demos.22.

[23] M. Li, H. Zhou, J. Hou, P. Wang, dan E. Gao, "Is cross-linguistic advert flaw detection in Wikipedia feasible? A multilingual-BERT-based transfer learning approach," *Knowledge-Based Syst.*, vol. 252, 2022, doi: https://doi.org/10.1016/j.knosys.2022.109330.

[24] B. Wilie *et al.*, "IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding," *arXiv*, 2020, doi: https://doi.org/10.48550/arXiv.2009.05387 Focus to learn more.

[25] J. Radom dan J. Kocoń, "Multi-task Sequence Classification for Disjoint Tasks in Low-resource Languages," *Procedia Comput. Sci.*, vol. 192, hal. 1132–1140, 2021, doi: https://doi.org/10.1016/j.procs.2021.08.116.

[26] I. Lourentzou, K. Manghnani, dan C. Zhai, "Adapting Sequence to Sequence models for Text Normalization in Social Media," *arXiv*, 2019, doi: https://doi.org/10.48550/arXiv.1904.06100.