Journal of Applied Engineering and Technological Science
Vol 4(2) 2023: 855-863

TWITTER DATA ANALYSIS AND TEXT NORMALIZATION IN COLLECTING STANDARD WORD

Arif Ridho Lubis[1*], Mahyuddin K M Nasution[2]
Department of Computer Engineering and Informatics, Politeknik Negeri Medan, Indonesia[1]
Faculty of Computer Science and Information Technology, Universitas Sumatera Utara, Indonesia[2]
arifridho@polmed.ac.id

Received : 18 April 2023, Revised: 07 May 2023, Accepted : 08 May 2023
**Corresponding Author*

ABSTRACT
This study discusses the Twitter data analysis and text normalization in standard word collection. Twitter is one of the most important data sources in social data analysis. However, the text contained on Twitter is often unstructured, resulting in difficulties in collecting standard words. Therefore, in this research, we analyze Twitter data and normalize text to produce standard words that can be used in social data analysis. The purpose of this research is to improve the quality of data collection on standard words on social media from Twitter and facilitate the analysis of social data that is more accurate and valid. The method used is natural language processing techniques using classification algorithms and text normalization techniques. The result of this study is a set of standard words that can be used for social data analysis with a total of 11430 words, then 4075 words with structural or formal words and 7355 informal words. Informal words are corrected by trusted sources to create a corpus of formal and informal words obtained from social media tweet data @fullSenyum. The contribution to this research is that the method developed can improve the quality of social data collection from Twitter by ensuring the words used are standard and accurate and the text normalization method used in this study can be used as a reference for text normalization in other social data, thus facilitating collection. and better-quality social data analysis. This research can assist researchers or practitioners in understanding natural language processing techniques and their application in social data analysis. This research is expected to assist in collecting social data more effectively and efficiently.
Keywords: *Natural Language processing, Word, Formal, Analysis, Twitter*

1. Introduction

The development of social media has now become a trend in terms of communication between on line users (Anandhaun et al., 2018; Zeng et al., 2010), social media is an online media with a gathering of users who can easily communicate, share and participate one another (Schreck & Keim, 2012; Middleton et al., 2013; A.R. Lubis et al., 2019). Data in social media represents research and manageable challenges in natural language processing (Young et al., 2018; Liang & Dai, 2013). Twitter is a widely popular social media platform around the world, where users can post short messages known as "tweets"(Lubis, Prayudani, Lubis, et al., 2022; Lubis, Prayudani, Nugroho, et al., 2022). Due to the brevity of tweets and the fact that they sometimes do not follow proper grammar or spelling rules, text normalization is required to collect correct and consistent words(A.R. Lubis et al., 2019; Arif Ridho Lubis et al., 2020). In this context, text normalization refers to the process of converting non-standard text to standard text by correcting spelling, grammar, and other errors(Neto et al., 2020; Dirkson et al., 2019). The results of text normalization can help improve the accuracy of collecting correct and consistent words for further analysis. The data obtained from social media was still unstructured which still needed to be improved (H. Zheng et al., 2020; X. Zheng et al., 2015). Several studies had been carried out on social media data in many languages such as Indian (Tanna et al., 2020; Roshini et al., 2019; Kumar et al., 2021) Chinese (Xuanyuan et al., 2021; Liu & Chen, 2019). The conducted research focused on improving the technique of the preprocessing process and the completion of non-standard and unstructured words due to the use of words and phrases in communication in Indonesian-language social media (Chen et al., 2020; Alhaj et al., 2022), Preprocessing is a stage in the natural language preprocessing method intended for documents in the form of text (Pano & Kashef, 2020; Villavicencio et al., 2021). The goal was to prepare data or text obtained from

unstructured social media into good data and could be easily processed for further processing(Jimenez-Marquez et al., 2019;Shu et al., 2017; Iskandar & Marjuki, 2022). In the preprocessing technique there were several processes such as parsing, case folding, tokenizing, stemming, filtering/stop words, normalization (Chen et al., 2020; Baccouche et al., 2020; Sarimole & Fadillah, 2022). In the text normalization stage, it was very important to be able to help parse Indonesian language that could understand lexical meaning well, performance in processing structural and unstructured words could be improved if the preprocessing stages were carried out properly, especially normalization for unstructured words (Izonin et al., 2022; Basan et al., 2022).

Research by (Göker & Can, 2018) In this study, two approaches were carried out for the Turkish text i.e. the contextual normalization approach and the sequence-to-sequence normalization approach using a neural encoder model. Other studies (Nguyen et al., 2017) carried out normalization which could extract information related to data accurately. Based on the conclusions and objectives of previous research, there are still weaknesses in normalizing text, so this study will apply text normalization which will use Indonesian language data originating from social media Twitter, which has unstructured and non-standard words, normalization techniques used to carry out text normalization process so that data can be processed and analyzed further. This normalization stage is adjusted to the data that has been obtained from social media (Jose & Raj, 2014;bin Sazali & Idris, 2022). The nature of the data obtained from social media is that there are users who generally communicate and express expressions (Sebastian & Nugraha, 2019), for example they say "kota in sunggh sjuk setiap hari", in the tweet snippets obtained by naked eye you can see several errors like "in" which should "ini", "sunggh" that should be "sungguh", and "sejk" harus "sejuk". If the words are processed and analyzed, such as sentiment, and classification, the results obtained are not accurate, so text processing analysis is needed. The purpose of this study is to use a dictionary-based normalization approach so that it can have superior differences from previous research. This study divides the process into 3 parts, namely text normalization, statistical words or lexicon, and non-standard words that appear in tweet data. In related research, standard Indonesian is specifically discussed. This study also provides slang and formal words to understand the characteristics of the tweet data.

2. Literature Review

Many researchers had applied the model of deep learning to many cases such as (Maghfur et al., 2021) Text-to-Speech (TTS) is widely used for both academic/non-commercial and industrial/commercial purposes. However, in some cases, text normalization is added to improve TTS performance. In this study, a rule-based approach is proposed to create a normalized Indonesian text dataset that has raw text and spoken form to improve Indonesian TTS performance. This approach shows good performance for text normalization for Indonesian TTS with a Word Error Rate (WER) of 0.0805. Another study conducted by (Khan & Lee, 2021) concluded that In this research, it is proposed to develop an application called Textual Variations Handler (TVH), which is a generic application that works in a variation-independent manner to handle various types of noise in textual data originating from various social media (SM) applications to improve text analysis. The aim of this research is to introduce a hybrid normalization technique that is effective in ensuring that information obtained from noisy text data can be utilized in the desired form. This study integrates the TVH application with a deep-learning state-of-the-art (SOTA) based text analysis method to improve its performance in analyzing noisy SM text data. The simulation results show that the proposed scheme is promising in terms of precision, recall, accuracy, and F1 scores in the analysis of informal texts on social media. This research (Sebastian & Nugraha, 2019), this research normalized the Indonesian language with data having several words from data consisted of unstructured sentences and non-standard words. The normalization method was carried out to analyze data for the next process.

This research (Rahman et al., 2019) analyzed the data of Indonesian tweets consisting of unstructured text in order to complete the word processing process and clean up tweet data from unstructured text. In processing text by using case folding, filtering stages, tokenizing. Then the normalization process was carried out so that words having excess letters, abbreviation, slang and all documents were converted into standard words and if there were words without meaning, they

were deleted. Other research conducted by (Javaloy & García-Mateos, 2020) concluded that this study proposes a new method for the encoder in the encoder-decoder architecture called Causal Feature Extractor (CFE) for text normalization as the first step in a text-to-speech system (TTS). CFE is compared to other encoding methods and shows better results in terms of accuracy, number of parameters, convergence time, and the use of the attention matrix based on the attention mechanism. The proposed method is general in nature and can be applied to various input types such as text, audio and images. This research (Sebastian & Nugraha, 2019) collected data in conducting research in the field of text mining. Data collection consists of many languages which were then processed to obtain normalized data in the word processing. Abbreviated words were a problem in text mining which resulted in the system not being able to process the text optimally due to differences in the meaning of the abbreviated words. The purpose of this study was to develop and obtain a set of Indonesian abbreviations. This research applied Crowdsourcing method in developing the dataset. Based on previous research that has been described there are deficiencies that can be resolved in this study which are listed in Table 1 below:

Table 1 - Examples of the Use of Normal and Normal Words

No	Researcher	Title	Lack
1	(Maghfur et al., 2021)	Text Normalization for Indonesian Text-to-Speech (TTS) using Rule-Based Approach: A Dataset and Preliminary Study	This research is rule-based only and does not involve machine learning, so it may not be able to handle some of the more complex cases of text normalization.
2	(Khan & Lee, 2021)	Enhancement of Text Analysis Using Context-Aware Normalization of Social Media Informal Text	There is no comparison with other text normalization methods that have been used in previous studies. This study only compares the performance of the proposed method with the same method without normalization.
3	(Rahman et al., 2019)	Normalization of Unstructured Indonesian Tweet Text For Presidential Candidates Sentiment Analysis	The text normalization dataset created in this study has not been verified considering regional variations in Indonesian and may need to be expanded.
4	(Javaloy & García-Mateos, 2020)	Text normalization using encoder–decoder networks based on the causal feature extractor	This study only discusses text normalization in English. Thus, it is not yet known how effective the proposed method is when applied to other natural language processing (NLP) problems.
5	(Gunawan et al., 2019)	Normalization of abbreviation and acronym on Microtext in Bahasa Indonesia by using dictionary-based and longest common subsequence (LCS)	This research has a weakness in filtering text, so it is necessary to add data and normalize it
6	(Kusumawardani et al., 2018)	Context-sensitive normalization of social media text in bahasa Indonesia based on neural word embeddings	This research only uses 1000 dictionary data so it needs to be added so that word tokens can represent all text data

The solution offered was the use of statistical matching translation to carry out the process of normalizing Indonesian text by utilizing translation at the phrase and character level in text data. In this case, using an external corpus for the data to be used in the normalization model. Indonesian language has a colloquial structure that could be found in the tweet data. In some cases, the everyday language words contained in the tweet data were still difficult to understand. The following examples of categories and samples could be seen in Table 2. On the linguistic side, you could also use the lexicon to search and observe slang in social media. In general, a comparison of the appearance of slang words from tweet data on social media was then carried out.

Table 2 - Examples of the Use of Normal and Normal Words

No	Non-Formal	formal
1	sjuk	Sejuk
2	Dtng	datang
3	Brlri	berlari
4	Plng	pulang
5	skolah	sekolah

3. Research Methods

The quantitative research method will assist in systematically and objectively collecting data from Twitter. The obtained data will be processed using data analysis techniques such as descriptive statistics and text classification. Additionally, this research will also employ text normalization methods to correct spelling and grammar errors in the text obtained from Twitter. Text normalization methods can include the removal of punctuation marks, the use of word-breaking algorithms, and the combination of separated words. This study will use a sample of data from the Twitter account @fullsenyum, which will be taken through a random sampling process. The data will then be processed using data analysis techniques and text normalization methods. Figure 1 shows the research flowchart.

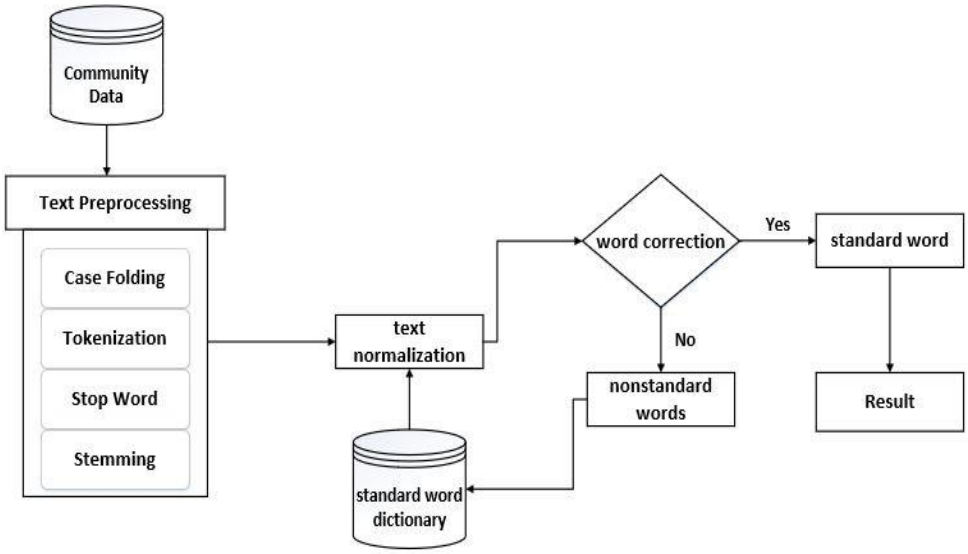

Fig. 1. Distribution of formal word Frequency

Based on Figure 1, there are two stages in collecting standardized words as follows:

a) Collecting Data

To obtain tweet data from the @fullsenyum community on the Twitter social media platform, this study used crawling techniques. The collected tweet data covers the period from 2018 to 2021, and a total of 11,430 tweets were successfully gathered. Table 3 shows the results of the crawling process.

b) Preprocessing

The preprocessing stage was needed in this research because the tweet data obtained from Twitter social media had unstructured data so that preprocessing was carried out to remove punctuation marks, emoticons, and make changes to uppercase letters to lowercase letters and eliminate words considered not important. The following was the result of preprocessing data.

Table 3 - The Preprocessing Result

No	Tweet
1	bener2 perjuangan, ngoding di hp dia.. KEREN!
2	anak perempuan milik ayahnya sampai ia menikah, tetapi anak laki-laki milik ibunya sampai ia mati
3	Udah seneng banget punya sahabat cowok. Eh, dianya malah nembak Golongan orang yg kalo udah ga respect, udah ga mau kenal lagi.
4	Mau ngumpulin orang yang makin hari makin males buka WhatsApp.
5	my mood can change from
6	take me back to the days where i sleep without over thinking.
7	"kok bisa putus sih?" "ya bisa"
8	Gimana ya kalo aku bukan kriteria yang diinginkan keluarganya

9	ni gw dah ngerjain tugas mati2an dari sd ampe kuliah awas ae gw gede ga tajir
10	hidup sudah menyebalkan, gausah -penting banget ok
10000	……..
11429	Makin kesini makin nguras mental banget ya.
11430	sebaik baiknya mood booster dan support system adalah uang, duit, cuan, dan money

c) Method

In this study, the text normalization method was used, which is the process of converting non-standardized text into standardized text, so that it is easier for computers to understand and process. There are several techniques commonly used in text normalization, such as removing non-alphanumeric characters, replacing slang words with standard words, adjusting the use of capital and small letters, and handling abbreviations and abbreviations. More sophisticated methods of text normalization use natural language processing (NLP) technologies and machine learning to recognize and correct errors in text, such as misspellings or incorrect use of words in certain contexts

4. Results and Discussions

This study built a lexicon by processing the words contained in the tweet data, totaling 11,430 data obtained from social media Twitter on the @fullsenyum account which had a public figure account. The tweet data obtained were preprocessed with several steps to improve the sentence structure of the tweet data. After the preprocessing stage, the daily Indonesian language lexicon produced contained 4075 formal words and 7355 informal words. Most of them were informal words in Indonesian. Each record had 2 columns:

- Non-formal words: words with non-formal meaning
- Formal: a formal word suited the Indonesian dictionary

Table 4 presented basic information about total words, total formal words and total informal words. There were 7355 informal words then there were 4075 formal words. Table 5 showed ten examples of informal words and then changes were made to informal words so that the quality of the data was better. Among 4075 unique formal words, 1,159 (67%) words occurred only once. Furthermore, Figure 1 showed that the distribution of words in the data to the occurrence of formal words applied Zipf law.

Table 4 - Formal and nonformal words

Total Data	Formal word	Non-Formal Word
11430	4075	7355

Table 5 - Example Of informal words and then changes were made to informal words

Non-Formal Word	URL formal word	Formal Word
yajangan	https://kbbi.web.id/jangan	jangan
temanyang	https://kbbi.web.id/teman	teman
cuplikanya	https://lambeturah.id/arti-kata-cuplikan-adalah/	cuplikan
enih	https://www.litbang.pertanian.go.id/info-aktual/3962/	benih
setidanya	https://kbbi.web.id/tidak	setidaknya
tonight	https://www.babla.co.id/bahasa-inggris-bahasa-indonesia/tonight	malam
waroeng	https://kbbi.web.id/warung	warung
dhuafa	https://dompetdhuafa.org/id/berita/detail/pengertian-dhuafa-menurut-islam	duafa
megamal	http://p2k.unkris.ac.id/id3/2-3065-2962/Mega-Mall-Batam-Center_69203_p2k-unkris.html	tempat
guestar	https://tr-ex.me/terjemahan/bahasa+inggris-bahasa+indonesia/guest+star	tamu

From the analyzed tweet data, this study found that all the words contained in the @fullsenyum community tweet data had bad structures so after preprocessing, 10221 words were obtained in the Indonesian Dictionary. This study comprehended only the types of AOA words contained in the data, this study presented 10 examples in table 6. Most of the informal words that were not listed in the Indonesian Dictionary had word excess or lack of words. Thus, it was difficult to represent the word and difficult to process

Table 6- Example Of not listed analyzed words KBBI

Non-Formal Word	Formal	Sentence advantages and disadvantages
yajangan	jangan	ya
temanyang	teman	yang
cuplikanya	cuplikan	ya
enih	benih	E
setidanya	setidaknya	k
tonight	malam	malam
waroeng	warung	e
dhuafa	duafa	duafa
megamal	tempat	megalmal
guestar	tamu	Guestar

From this study could see the distribution of words in the data shown in Figure 2, formal words tend to be more structured and had longer words than non-formal words with the average number of characters per word being 7 for the first and 6 for the last.

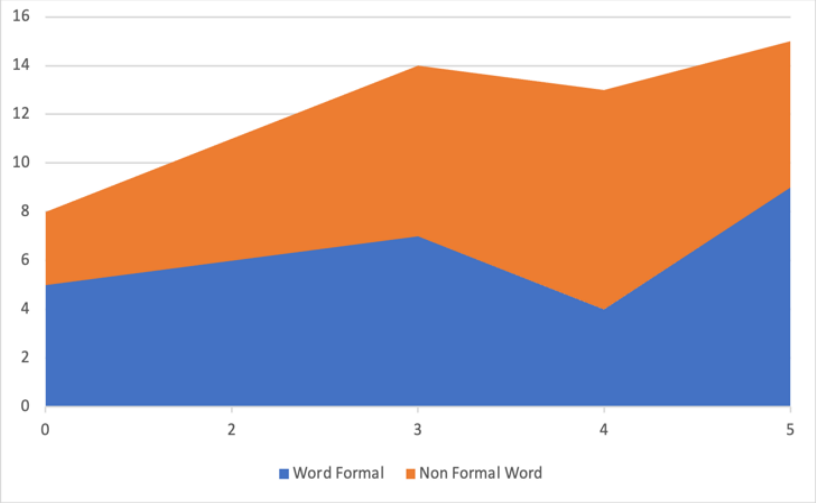

Fig. 2. The number of normalized characters for non-formal and formal words.

At this stage analyzed the frequency of words contained in the tweet data that had been preprocessed. Frequency was done to see how many repetitions of words in the data. At this stage it focused on making changes manually to words with poor structures including excess words, word deficiencies and errors in the preprocessing process. Table 7 showed examples of words with bad structure, which were then corrected manually so that they could be detected by the Indonesian dictionary.

Table 7 - Correction of Unstructured non-formal words

Non-Formal Word	URL formal word	Formal Word
yajangan	https://kbbi.web.id/jangan	jangan
temanyang	https://kbbi.web.id/teman	teman
cuplikanya	https://lambeturah.id/arti-kata-cuplikan-adalah/	cuplikan
enih	https://www.litbang.pertanian.go.id/info-aktual/3962/	benih
setidanya	https://kbbi.web.id/tidak	setidaknya
tonight	https://www.babla.co.id/bahasa-inggris-bahasa-indonesia/tonight	malam
waroeng	https://kbbi.web.id/warung	warung
dhuafa	https://dompetdhuafa.org/id/berita/detail/pengertian-dhuafa-menurut-islam	duafa
megamal	http://p2k.unkris.ac.id/id3/2-3065-2962/Mega-Mall-Batam-Center_69203_p2k-unkris.html	tempat

Based on Table 7, the results of the 11430 Tweet data were then preprocessed which resulted in data with a total of 10,221 then separated standard and non-standard sentences according to the data of the Indonesian language dictionary obtained by the author. The obtained results were 4075 standard words and non-standard words. The standard number was 4778, after a lexical analysis had been carried out that informal words had poor word structures. Thus, to get words with good structures, the author conducted a manual word identification according to the sources that had been given. Words with poor structure contained excessive words, repeated words and no space words.

5. Conclusion

In this study, the researcher presented a lexicon of formal words contained in the Twitter social media data with the normalized @fullsenyum account. The resulting corpus data was useful for natural language preprocessing processes or tasks in Indonesian. In this case, the Indonesian language corpus data were added to the data obtained from Twitter. the data were obtained available on GitHub under the MIT license. In this study there were several techniques in making formal Indonesian word corpus data as a dictionary for the stages in text normalization, and as a data collection as a model of natural language preprocessing but this corpus data could be used in developing science in the field of natural language preprocessing in detail. This research should be developed more widely to utilize tweet data with many characteristics of the sentences contained in the tweet data. Many studies had been carried out in the field of natural language preprocessing. This research was expected to improve the performance of social media analysis, especially Twitter in Indonesia This research can contribute to the development of a more standardized and structured Indonesian lexicon. In this context, the text normalization method used in this research can be the basis for collecting standard words that are often used in Indonesian based on Twitter data. Another implication is that this research can assist in the development of a more effective and accurate text mining system for analyzing social media data, especially in the Indonesian context.

References

Alhaj, Y. A., Dahou, A., Al-qaness, M. A. A., Abualigah, L., Abbasi, A. A., Almaweri, N. A. O., Elaziz, M. A., & Damaševičius, R. (2022). A novel text classification technique using improved particle swarm optimization: A case study of Arabic language. *Future Internet*, *14*(7), 194.

Anandhan, A., Shuib, L., Ismail, M. A., & Mujtaba, G. (2018). Social media recommender systems: review and open research issues. *IEEE Access*, *6*, 15608–15628.

Baccouche, A., Ahmed, S., Sierra-Sosa, D., & Elmaghraby, A. (2020). Malicious text identification: deep learning from public comments and emails. *Information*, *11*(6), 312.

Basan, E., Basan, A., Nekrasov, A., Fidge, C., Abramov, E., & Basyuk, A. (2022). A Data Normalization Technique for Detecting Cyber Attacks on UAVs. *Drones*, *6*(9), 1–21. https://doi.org/10.3390/drones6090245

bin Sazali, M. A. H., & Idris, N. B. (2022). Neural Machine Translation for Malay Text Normalization using Synthetic Dataset. *2022 10th International Conference on Information and Communication Technology (ICoICT)*, 386–390.

Chen, W., Xu, Z., Zheng, X., Yu, Q., & Luo, Y. (2020). Research on sentiment classification of online travel review text. *Applied Sciences*, *10*(15), 5275.

Dirkson, A., Verberne, S., Sarker, A., & Kraaij, W. (2019). Data-driven lexical normalization for medical social media. *Multimodal Technologies and Interaction*, *3*(3), 60.

Göker, S., & Can, B. (2018). Neural text normalization for turkish social media. *2018 3rd International Conference on Computer Science and Engineering (UBMK)*, 161–166.

Gunawan, D., Saniyah, Z., & Hizriadi, A. (2019). Normalization of abbreviation and acronym on Microtext in Bahasa Indonesia by using dictionary-based and longest common subsequence (LCS). *Procedia Computer Science*, *161*, 553–559.

Iskandar, D., & Marjuki, M. (2022). Classification of Melinjo Fruit Levels Using Skin Color Detection With RGB and HSV. *Journal of Applied Engineering and Technological Science (JAETS)*, *4*(1), 123–130. https://doi.org/10.37385/jaets.v4i1.958

Izonin, I., Tkachenko, R., Shakhovska, N., Ilchyshyn, B., & Singh, K. K. (2022). A Two-Step Data Normalization Approach for Improving Classification Accuracy in the Medical Diagnosis Domain. *Mathematics*, *10*(11), 1–18. https://doi.org/10.3390/math10111942

Javaloy, A., & García-Mateos, G. (2020). Text normalization using encoder–decoder networks based on the causal feature extractor. *Applied Sciences*, *10*(13), 4551.

Jimenez-Marquez, J. L., Gonzalez-Carrasco, I., Lopez-Cuadrado, J. L., & Ruiz-Mezcua, B. (2019). Towards a big data framework for analyzing social media content. *International Journal of Information Management*, *44*, 1–12.

Jose, G., & Raj, N. S. (2014). Noisy SMS text normalization model. *International Conference for Convergence for Technology-2014*, 1–6.

Khan, J., & Lee, S. (2021). Enhancement of Text Analysis Using Context-Aware Normalization of Social Media Informal Text. *Applied Sciences*, *11*(17), 8172.

Kumar, A., Tyagi, V., & Das, S. (2021). Deep Learning for Hate Speech Detection in social media. *2021 IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON)*, 1–4.

Kusumawardani, R. P., Priansya, S., & Atletiko, F. J. (2018). Context-sensitive normalization of social media text in bahasa Indonesia based on neural word embeddings. *Procedia Computer Science*, *144*, 105–117. https://doi.org/10.1016/j.procs.2018.10.510

Liang, P.-W., & Dai, B.-R. (2013). Opinion mining on social media data. *2013 IEEE 14th International Conference on Mobile Data Management*, *2*, 91–96.

Liu, K., & Chen, L. (2019). Medical social media text classification integrating consumer health terminology. *IEEE Access*, *7*, 78185–78193.

Lubis, A.R., Lubis, M., & Azhar, C. D. (2019). The effect of social media to the sustainability of short message service (SMS) and phone call. *Procedia Computer Science*, *161*. https://doi.org/10.1016/j.procs.2019.11.172

Lubis, Arif Ridho, Nasution, M. K. M., Sitompul, O. S., & Zamzami, E. M. (2023). A new approach to achieve the users' habitual opportunities on social media. *IAES International Journal of Artificial Intelligence*, *12*(1), 41–47. https://doi.org/10.11591/ijai.v12.i1.pp41-47

Lubis, Arif Ridho, Prayudani, S., Lubis, M., & Nugroho, O. (2022). Sentiment Analysis on Online Learning During the Covid-19 Pandemic Based on Opinions on Twitter using KNN Method. *2022 1st International Conference on Information System & Information Technology (ICISIT)*, 106–111.

Lubis, Arif Ridho, Prayudani, S., Nugroho, O., Lase, Y. Y., & Lubis, M. (2022). Comparison of Model in Predicting Customer Churn Based on Users' habits on E-Commerce. *2022 5th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, 300–305.

Lubis, Arif Ridho, Utara, U. S., Sitompul, O. S., Utara, U. S., Nasution, M. K. M., Utara, U. S., Zamzami, E. M., & Utara, U. S. (2020). *Obtaining Value From The Constraints in Finding User Habitual Words*. 8–11.

Maghfur, N. M., Ibrohim, M. O., Fahmi, J., Putera, A. S., & Riandi, O. (2021). Text Normalization for Indonesian Text-to-Speech (TTS) using Rule-Based Approach: A Dataset and Preliminary Study. *2021 4th International Conference of Computer and Informatics Engineering (IC2IE)*, 129–134.

Middleton, S. E., Middleton, L., & Modafferi, S. (2013). Real-time crisis mapping of natural disasters using social media. *IEEE Intelligent Systems*, *29*(2), 9–17.

Neto, A. F. de S., Bezerra, B. L. D., & Toselli, A. H. (2020). Towards the natural language processing as spelling correction for offline handwritten text recognition systems. *Applied Sciences*, *10*(21), 7711.

Nguyen, L. H., Salopek, A., Zhao, L., & Jin, F. (2017). A natural language normalization approach to enhance social media text reasoning. *2017 IEEE International Conference on Big Data (Big Data)*, 2019–2026.

Pano, T., & Kashef, R. (2020). A complete VADER-based sentiment analysis of bitcoin (BTC) tweets during the era of COVID-19. *Big Data and Cognitive Computing*, *4*(4), 33.

Rahman, T., Agustin, F. E. M., & Rozy, N. F. (2019). Normalization of Unstructured Indonesian Tweet Text For Presidential Candidates Sentiment Analysis. *2019 7th International Conference on Cyber and IT Service Management (CITSM)*, *7*, 1–6.

Roshini, T., Sireesha, P. V., Parasa, D., & Bano, S. (2019). Social media survey using decision tree and Naive Bayes classification. *2019 2nd International Conference on Intelligent Communication and Computational Techniques (ICCT)*, 265–270.

Sarimole, F. M., & Fadillah, M. I. (2022). Classification Of Guarantee Fruit Murability Based on HSV Image With K-Nearest Neighbor. *Journal of Applied Engineering and Technological Science (JAETS)*, *4*(1), 48–57. https://doi.org/10.37385/jaets.v4i1.929

Schreck, T., & Keim, D. (2012). Visual analysis of social media data. *Computer*, *46*(5), 68–75.

Sebastian, D., & Nugraha, K. A. (2019). Text normalization for Indonesian abbreviated word using crowdsourcing method. *2019 International Conference on Information and Communications Technology, ICOIACT 2019*, 529–532. https://doi.org/10.1109/ICOIACT46704.2019.8938463

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, *19*(1), 22–36.

Tanna, D., Dudhane, M., Sardar, A., Deshpande, K., & Deshmukh, N. (2020). Sentiment analysis on social media for emotion classification. *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*, 911–915.

Villavicencio, C., Macrohon, J. J., Inbaraj, X. A., Jeng, J.-H., & Hsieh, J.-G. (2021). Twitter sentiment analysis towards covid-19 vaccines in the Philippines using naïve bayes. *Information*, *12*(5), 204.

Xuanyuan, M., Xiao, L., & Duan, M. (2021). Sentiment classification algorithm based on multi-modal social media text information. *IEEE Access*, *9*, 33410–33418.

Young, T., Hazarika, D., Poria, S., & Cambria, E. (2018). Recent trends in deep learning based natural language processing. *Ieee Computational IntelligenCe Magazine*, *13*(3), 55–75.

Zeng, D., Chen, H., Lusch, R., & Li, S. (2010). *Social Media Analytics and Intelligence*. *DEcEMbEr*.

Zheng, H., Lin, F., Feng, X., & Chen, Y. (2020). A hybrid deep learning model with attention-based conv-LSTM networks for short-term traffic flow prediction. *IEEE Transactions on Intelligent Transportation Systems*, *22*(11), 6910–6920.

Zheng, X., Chen, W., Wang, P., Shen, D., Chen, S., Wang, X., Zhang, Q., & Yang, L. (2015). Big data for social transportation. *IEEE Transactions on Intelligent Transportation Systems*, *17*(3), 620–630.