Paper 2

by Reonal Regen

Submission date: 22-Dec-2023 09:22AM (UTC+0700)
Submission ID: 2191399129
File name: 2-DS.pdf (284.87K)
Word count: 4678
Character count: 25750

Development of Compound Non-Standard Word Dataset Using Crowdsourcing Method

1st Danny Sebastian
Informatika, Fakultas Teknologi Informasi
Universitas Kristen Duta Wacana
Yogyakarta, Indonesia
danny.sebastian@staff.ukdw.ac.id

2nd Kristian Adi Nugraha
Informatika, Fakultas Teknologi Informasi
Universitas Kristen Duta Wacana
Yogyakarta, Indonesia
adinugraha@ti.ukdw.ac.id

***Abstract* - During the COVID-19 Pandemic, many activities, such as communication and socialization are forced to be carried out through digital devices. Many communication applications have been developed to improve the users' communication experience, namely chatbots, auto-translation, sentiment analysis and many other applications. However, most Indonesian use non-standard words when communicating with each other. These non-standard words used caused the text processing software work to be nonoptimal. This research aims to create a corpus of standard and non-standard words. This corpus can then be used to normalize the non-standard word to the standard version. The crowdsourcing method was chosen to create the dataset. This research has successfully collected 371 records as final corpus data. The most common problem is the difference in respondents' perceptions in determining single or compound non-standard words. There were 5 problems found in the forming of corpus data, namely character repetition in non-standard words, various forms of loan words, typing errors, differences between respondent's answers, and expression word without standard.**

Keywords – text preprocessing, text normalization, crowdsourcing, text mining

I. Introduction

During the COVID-19 Pandemic, many activities are forced to be carried out through digital devices. The utilization of social media has become crucial in communication's needs. There are many application software developers that are innovating in order to improve communication experience. There is a software developer that attempts to develop automatic text processing applications, such as chatbot [1], auto translation [2], sentiment analysis [3], etc.

There are 3 general steps in automated text processing, namely text pre-processing, application of text processing algorithms, and analysis [4]. Text pre-processing processes raw text data so that the data is ready to be used on a text processing algorithm. Text pre-processing consists of multiple processes, such as case-folding, tokenization, punctuation, filtering, stemming, and lemmatization [5]. Case-folding is a process to change each character into lowercase. Tokenization is the process of breaking up texts into collections of words or otherwise known as tokens. Punctuation is the process of cleaning symbols and other characters that will not be used in the following stages. Filtering is the process to delete unnecessary tokens, such as stopword-removal [6]. Stemming is a method to produce the stem/root or basic form of a word [7]. Whereas lemmatization is a method to do morphological analysis of a word. Text preprocessing is important because if the text-preprocessing stage fails in preparing the data, then the output of the text processing algorithm will not be optimal [8].

There are 2 kinds of text preprocessing, which are general text preprocessing and specific text preprocessing [9]. General text preprocessing is a general process of text-preprocessing, such as case-folding, tokenization, and punctuation. Meanwhile, specific text preprocessing is a text preprocessing that focuses on the characteristics of the text data that is processed. For example, if we process text data sourced from social media, there will be emoticons [10]. However, if the text data is obtained from twitter [11], then there will be mention symbol (@), retweet symbol (RT), and hashtag symbol (#) [9] [12]. One of the important steps in specific text preprocessing is normalization. Normalization is a method to transform non-standard words into standard words [13].

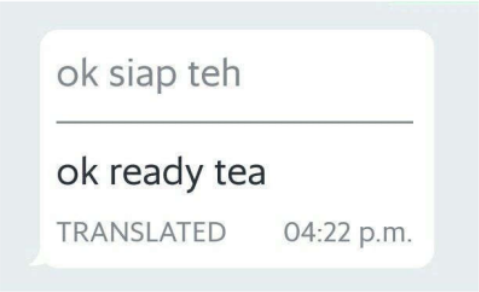

Fig. 1. Taxibike online system failed to auto translate local language [14]

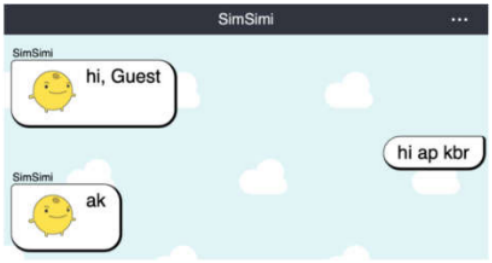

Fig. 2. Simsimi chatbot failed to response [14]

In communicating using Natural Language, there is the usage of non-standard words or "alay" words or non-standard words [14] [15]. Like on social media, humans usually communicate using Natural Language, so it contains a lot of non-standard words [16]. Non-standard words may decrease the performance of text-processing algorithms. As shown in Fig. 1, it is shown that auto translation encountered an error in performing translation. The error happens as a result of the West Java local language, which uses "teh" instead of

"bu"(ma'am). Another example is presented on Fig. 2, chatbot simsimi encountered an error when answering to the abbreviated word from the user. Another error happened as the result of the abbreviated word in Indonesian.

Currently there are 2 ways to normalize the data, byusing the stemming method and the corpus data [9] [17]. Stemming is a method to obtain the stem/root or root form of the word. For example, during the stemming process, the word "saved(disimpan)" will generate the word "save(simpan)". Stemming method does the normalization based on affix rules in Indonesian Language. Stemming method is developed to be able to stem non-standard words by adding rules [18]. Nazief and Andrianis 2 stemming method has been developed to do normalization by adding non-standard affix rules [2]. Non-standard affix rules that are added, consists of prefix-1 (n-, ny-, m-, ng-, nge-), prefix-2 (to), suffix-1 (-in, -an), suffix-2 (-san), confix-1 (n-in, in-in, late), and confix-2 (se-an). Nazief and Andrianis' method and affix rules were further developed by the Levenshtein Distance algorithm and Indonesian dictionary [19]. The disadvantage of the stemming method is there are abbreviated words and compound words that can't be handled by the stemming algorithm [20].

The normalization method uses corpus data to normalize the non-standard words, based on the data that is available in database/corpus data [21]. The disadvantage of the normalization method that uses the corpus data is that it requires the complete corpus data in order to be able to do the normalization properly. Closed corpus normally is limited to specific domains, such as on Indonesian Closed Corpus for Business (Incorbiz) [20]. Closed corpus Incorbiz was built by a combination of litterateur stemming method and human supervision. In general, incorbiz records the root word, domain, word type, word variant, and English word [22] [20]. The Incorbiz method is unable to solve the compound words and abbreviation problem. Some of the abbreviation problems can be solved, however there are too many abbreviated forms of a standard word. For example, the word "transfer" can be abbreviated into "tf", "trf", "transf", "trans", etc. This shows the disadvantage of the corpus based normalization, which requires a complete corpus.

In constructing the corpus data of the non-standard words, human supervision is needed. Humans are in charge of determining whether a word is a non-standard word and determining the standard form. However, there is so many data on social media, that it requires a method to reduce the number of words that humans have to check. In the Incorbis corpus based method, word reduction that should be processed is done by the stemming method [20] and checking the online Indonesian Dictionary (KBBI) [23]. Another way to solve the problem of large numbers of words that should be checked is to increase the number of humans by crowdsourcing [13] [23]. Crowdsourcing method is able to generate corpus non-standard words, however there are some weaknesses in the methods that have been used. The compound words fail to be processed by the crowdsourcing method due to this limitation in the system that is built. Other than compound words, the crowdsourcing method cannot generate the correct standard form of the words. Due to the crowdsourcing method, the respondents may vote the incorrect standard form.

Based on the background that has been described above, this research will make a dataset or corpus of non-standard and standard words. The system that is built must be able to collect the dataset of non-standard compound words. In order to reduce the number of words that must be processed, the stemming method is used and check the online Indonesian Dictionary (KBBI). The data that is used is text data from social media. The system will not conclude the non-standard and standard word through the voting method, instead through the administrator role.

II. Research Methodology

A. Procedures of Experiments

First, a web application is developed as the crowdsourcing tool. Then, text data is sourced from Instagram. This text data is then used as questions for the crowdsourcing process. In the web application developed, the respondents were prompted to fill out their personal data, and then answer the questions given. Every question is shown in a complete sentence so that the respondents would get a context on the what is being talked about. Respondents were asked to choose which word is a non-standard word in the given sentence and input the standard word. The web application will then collect the answers collected from the respondents and save them to the database. Every question will be asked to more than 1 respondent, and every respondent may answer more than 1 question. After the crowdsourcing process is done, the author will then compile the data to conclude every non-standard word and it's standard form. Every conclusion will then be added as a record on the final corpus data.

B. Apparatus

A website-based application was developed to collect crowdsource data from participants. The web based system was developed using the PHP programming language. Web platform was chosen because the web application will be connected to the internet so that crowdsourcing can be done easily. The crowdsourcing website application consists of 2 main pages, the biodata form page and crowdsourcing form page. Biodata form page is shown on Fig. 3. This respondent biodata form page is used to obtain the respondents' biodata information and instruction on how to fill out the response. Fig. 4 shows the crowdsourcing form page, which is used by respondents to fill out the answers from the crowdsourcing questions.

Fig. 3. Homepage and respondents' biodata form

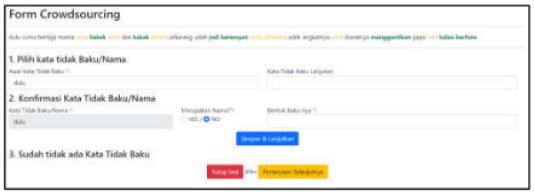

Fig. 4. Crowdsourcing form page

On the crowdsourcing form page, the respondents will be given a sentence from social media. Then, the respondents are requested to choose the non-standard words in the initial non-standard word combo box. If the non-standard word is not a compound word, then the respondent will be asked to fill out the standard word field and save the response. However if the non-standard word is a compound word, then the respondent is asked to also fill the second non-standard word field. Then respondents were also asked to choose whether it is a non-

standard word or a name by selecting a radiobox. For example on Fig.6 (b) there is the word “ber empat”, and the word “ber empat” is a compound non-standard word. Therefore the respondent needs to fill the initial non-standard word field with the word “ber” (one of the prefixes) and “empat” (four) in the second non-standard word field. After that, the respondents were asked to fill out the standard word field of “ber empat”, which is “berempat” (a group of four). Afterwards, the system will automatically show the word “berempat” instead of “ber empat” in bold and green. Whereas if the selected word is “vania", which is a name, then the system will automatically show it in orange.

C. Datasources

The data used in this research was sourced from Instagram. This social media was chosen as there are many non-standard word text data in the comment sections. Instaloader tools were used to source Instagram comments data. The Instagram accounts chosen are the top 20 Instagram accounts with the most followers in Indonesia and have been verified by Instagram. Due to the limitation of data requests to the Instagram server, 10 posts will be taken for each account. And from 10 posts, only 10 comments will be taken. Table I shows the Instagram accounts used in this research. The data itself was sourced at 30 June 2021 through StarNgage website [24]. There are 2 accounts removed from the list because those accounts deactivated their comment section. Those accounts are @pricesssyahrini and @nissa_sabyan. Those 2 accounts are replaced by the next accounts in the list, which are @ollaramlanufar and @juliaperrezz.

TABLE I. INSTAGRAM ACCOUNT CHOSEN USE IN THIS RESEARCH

Akun Instagram	Follower	Akun Instagram	Follower
@raffinagita1717	46,4M	@bramastavrl	19,8M
@laudyacynthiabella	32,9M	@jscmila	18,8M
@gisel_la	30,7M	@gadiiing	18,1M
@natashawilona12	29,9M	@aurelie.hermansyah	17,6M
@chelseaoliviaa	25,7M	@raditya_dika	16,7M
@ashanty_ash	25,3M	@bclsinclair	16,5M
@inijedar	25,2M	@dagelan	16,1M
@ivan_gunawan	23,6M	@attahalilintar	15,9M
@agnezmo	23,2M	@ollaramlanaufar	15,6M
@sandradewi88	19,9M	@juliaperrezz	15M

The data was sourced at 30 June 2021 and collected 2.026 comments data. From 2.026 comments data, filtering was done to make sure that there is a minimum of 1 non-standard word in it. The filtering process was done manually by the administrator and by the system. The manual filtering process is deleting advertisements and spam based on certain keywords. Example of spam or advertising comments could be seen at Table II.

TABLE II. EXAMPLE OF SPAM OR ADVERTISING COMMENTS

No	Spam or Advertising Comments
1	HAII kak aku mau kenalin bisnis pelajar yg modal nya cuma 50k penghsilan 50-300rb perhari💸Mau Join? Boleh bgt, YUK LGSNG DM / KLIK LINK DI BIO🤗
2	**https://vm.tiktok.com/ZSJP6RpqY/** Dan masukan id A6687201549 sama sama dapet uang 10k Terimakasih
3	I will do promotion for etsy store, shopify store, to get unlimited real traffic **https://www.fiverr.com/share/1WQ7x5** #etsypromotion #etsy #sale #cryptocurrency #promotion #marketing #store #donaldtrump #bhfyp #world #etsypromotion #etsystore #ebay #amazon #amazonstore
4	ABDUL HAKIM +918952832581 **molanaabdulhakim@gmail.com** molanaabdulhakimblackmagicspecialist.blogspot.com
5	PROMO=XL40GB:55K,INDOSAT30GB:50K,TSEL30GB:60K CEK IG 🙏

Normally there is an initial cleaning process that consists of transforming all characters into lowercase, controlling the special words (@mention, $url, #hashtag), and deleting all symbols. The list of initial cleaning processes is shown on Table III. Then, this process is being repeated for each token. In this repetition, the token was checked on the “standard word” corpus in the local database, to the “non-standard” corpus, and then to the online Indonesian Dictionary (KBBI). The preliminary data used in the “standard word” corpus came from litterateur’s basic words corpus. However, if the token is not found yet in the corpus local database, then it will be checked to the online Indonesian Dictionary (KBBI). This is to reduce the number of requests to the online Indonesian Dictionary’s (KBBI) server. At the end of the process, if there is the word in the token that is unlisted in the online Indonesian Dictionary (KBBI), then it will be recorded as a crowdsourcing question. After the filtering process was done, 728 comments data was generated and ready to be used as the non-standard data for the crowdsourcing sessions.

TABLE III. COMPLETE LIST OF THE FILTERING PROCESS

No.	Other filtering process	Process
1	Url	Change to $url
2	Emoticon	Emoticon will be removed
3	Mention	Change to @mention
4	Hashtag	Change to #hashtag
5	Phone number	Likely spam, Record will be removed
6	Comment under 3 characters	Record will be removed
7	Spam or advertising	Record will be removed
8	Not having non-standard word	Record will be removed

D. Participants

About 56 participants were selected on voluntary basis. Author confirmed that all participants can speak Indonesian. All participants were 18-30 years old and located in Yogyakarta.

III. EXPERIMENTAL RESULTS

From the test results obtained 792 records combination of standard and it's a unique non-standard word based on respondent's answer. Then, from 792 records, there were 651 single non-standard words, 78 non-standard compound words, and 63 name words. Corpus data that was generated is shown on Table IV.

TABLE IV. RESPONDEN’S VOTE AND DATA COLLECTED

No.	Type	Data Count
1	Single non-standard word	651
2	Compound non-standard word	38
3	Name word	63

Some of the non-standard and standard words produce more than 1 standard form. Some respondent’s answers are other forms of words that are still non-standard. For example, according to respondents there are 2 standard words produced from the word “kesian”, which is “kasihan” and “kasian”. Then, the administrator in this case determines the word “kasihan” (pity) as the standard form. Some of the examples of non-standard, standard, and conclusion are shown on Table V.

TABLE V. SOME EXAMPLE OF NON-STANDARD WORD AND IT’S STANDARD FORM BASED ON THE CROWDSOURCING PROCESS

Non-standard word	Answer standard word	Conclusion

adeeekk	adik, adek	adik
ber empat	berempat	berempat
bngt	banget, sekali, sangat	sangat
idulfitri	idul fitri	idul fitri
kesian	kasihan, kasian	kasihan
kk	kakak, kaka	kakak
mf	maaf	maaf
seneng	Senang, gembira	senang
tdak	tidak	tidak
yach	ya	ya

After the crowdsourcing process, the writers formed the corpus data based on the respondents' answers. The data structure consists of 4 columns, which is id, standard_word, non_standard_forms, and context. The id column is the identity for each standard_word. For one standard_word, there can be many non_standard_forms. Every combination of standard_word and non_standard_forms can have multiple contexts. As for the non_standard_forms and context columns, it is saved in JSON text format. The context column is for the sentence context where non-standard words are found. In order to complete the corpus data, the writer attempted searching in an online Indonesian Dictionary (KBBI) and adding on possible non_standard_forms. Example of the final result of the corpus data that was formed is shown on Table VI. The number of records generated after the compilation process is 371 records.

TABLE VI. EXAMPLE OF FINAL RESULT CORPUS DATA

id	standard_word	non_standard_forms	Example of context
1	anaknya	anknya, anak nya, ankny	["masyaallah seneng dech lihat bu vena ngasuh vania serasa anknya sendiri,... apa yg dilakukan bu vena tdak mengubah dunia luas namun sudah mengubah dunia bagi vania,proud of u"]
2	kasihan	kesian, kasian	["kesian si bontot tdak nyampe ciumannya"]
3	kakak	kk, kaka, kak	["happy eid mubarak kaka"]
4	tidak	tdak, tdk, ngga, gk, ga	["@mention tdk kerumah papanya?"]
5	lagi	lg, lgi	["lebarannya ga bikin tiktok? biar muncul di fyp lg"]
6	maaf lahir batin	maf lahir batin, maf lahr batin, maaf lahr batin, mf lahir dan batin	["selamat hary raya idul fitri ya kak,minal aidi wal faidin.mohon mf lahir dan batin y bunda sekeluarga."]
7	mamanya	mamahny, mamany, mama nya, mamah nya	["vania udh gede mirip bgt sm mamahny sht sllu yh klwrqa mam vena melinda"]
8	selamat	selamet, selmat, selat, met, slmt	["met idul fitri k @mention and kel .. minal aidin"]
9	sudah	udah, udh, sdh	["si hablu udh gede,, gemess,pinter photo"]
10	ya	yach, yah, yaa, yh, y	["taqobballallahu minna wa minkum,taqobbal yaa karim,mohon maaf lahir bathin yach mas bram"]

The most common problem found is the difference in perception between respondents in determining compound or single words. For example, the word "maf lahr batin" (forgiveness). There are respondents who answer as single non-standard word , "maf" as "maaf", "lahr" as "lahir", and ignore the word "batin" because it is considered standard word. However there are also respondents who consider it a non-standard compound word. Administrators have to combine the non-standard word based on answers from respondents before added to the final corpus data.

The writer found a few problems while forming the corpus data. First, there are non-standard words with recurring characters. For example, the word "adik" (younger brother/sister) is written in various forms, such as "adeeekk", "adek", "adiiikkk", etc. The writer encountered few difficulties in adding the non_standard_forms forms besides the respondents' answers. This occurred due to too many combinations of recurring words for the word "adik", moreover there are those who may write the combinations of the words "adek". This made it impossible to record every single character combination of the words that exist for non_standard_forms. Therefore, additional methods are needed to overcome this problem. Some of the examples are shown on Table VII.

TABLE VII. EXAMPLE OF REPEATED CHARACTER ON NON-STANDARD WORD

id	standard_word	non_standard_forms	context
1	adik	adek, adiiikk, adeeekk, adeekkk	["adeeekk keciiikkk kakak cubit cium ya"]
2	kecil	keciiikkk, kcl, kecl, kecik	["adeeekk keciiikkk kakak cubit cium ya"]
3	wah	Wahhhh, wuah	["wahhhh happy family! anak lelaki yang tampan, penyayang...semoga keluarga puan venna melinda sentiasa bersama-sama, sayang berpanjangan", "wuah! keluarga bahagia nih"]

The second problem is there are foreign words in which it is difficult to determine it's standard form, such as the word "Assalamualaikum" (Arabic greeting that is often used by Muslims in Indonesia too). This word is written in various forms by respondents, such as "Wassalamualaikum" and "Asalamualaikum". The writer checked it manually in the online Indonesian Dictionary (KBBI), however there was no result or proper standard form found. Therefore in the corpus formation, the word "Assalamualaikum" is decided as standard_word based on majority votes. The example is shown on Table VIII.

TABLE VIII. EXAMPLE OF FOREIGN LANGUAGE LOANWORDS

id	standard_word	non_standard_forms
1	asallamualaikum	assalamualaikum, assalamualikum, wassalamualaikum
2	masyaallah	masya allah, masyallah, massahallah
3	minal aidin wal faizin	minal aidin, minal aidzin wal faidzin, minal aidin walfaidzin, minal aidzin walfaizin, minal aidzin walfaizin, mindl adinwalfaizin

The following problem is there are several typing errors both in non-standard and standard words. For example, the word "makon" in the crowdsource question, "mamahny ka farrrl makon kesini bukan malah makin tua malah makin cantik". The typing error is between the character "o" and "i" because they are next to each other on the keyboard. The word "makon" will be considered a non-standard word from the word "makin" (more). The example of typing error could be seen at Table IX.

TABLE IX. EXAMPLE OF MISS TYPE NON STANDARD WORD

id	standard_word	non_standard_forms
1	maaf	maff
2	makin	makon
3	faidzin mohon	faidzinmohon
4	lahir	lahit
5	idul	eidul, iedul
6	selamat	selmat

There are several non-standard words that are written as only one character. For example, the word “di” (at, on) is written in abbreviated form just as “d”, where the vocal letter “i” is omitted. This also occurs on the word “ke” (to) that is abbreviated as “k”. Another example is the phrase “d IG nya” (on his/her instagram). There were respondents who classify it as compound non-standard words and those who classify it as single non-standard words. The examples are shown on Table X.

TABLE X. EXAMPLE OF DIFFERENT RESPONSE ON COMPOUND NON-STANDARD OR SINGLE NON-STANDARD WORD

id	Responden	standard_word	non_standard_forms
1	Responden 1	di	d
2		instagram	IG
3		nya	ny
4	Responden 2	di Instagram	d IG
5		ny	nya
6	Responden 3	di Instagramnya	d IG nya

The last problem is the words that express emotion, such as laughing, like the word "hahaha", "hehehe"", "wkwkwk", etc. This thing is considered as a problem as there are no standard word which represents "laughing". The result of this research is still unable to tackle the non-standard word which express emotional feelings. In this study, words like this were not added to the final data corpus.

IV. CONCLUSION

Based on the experiments’ results, this crowdsourcing method still requires an administrator to conclude the standard and non-standard form of a word in order for it fits the context of the sentence. The structure of the final data corpus formed by 4 columns, which is id, standard_word, non_standard_forms, and context. The number of record data generated is 371.

The problem most common problem is the difference in respondents' perceptions in determining single or compound non-standard words. as in the word "maf lahr batin". There were 5 problems found in the forming of corpus data, namely character repetition in non-standard words, various forms of loan words, typing errors, differences between respondent’s answers and expression word without standard. Future research is expected to add more raw text data that is used to complete the data corpus. In addition, testing needs to be done to ensure the normalization results using the data corpus that has been built.

ACKNOWLEDGMENT

This research was supported by Duta Wacana Christian University’s Institute of Research and Community Service Institute (Lembaga Penelitian dan Pengabdian pada Masyarakat), contract number 078/D.01/LPPM/2021. We also would like to thank Duta Wacana Christian University’s Faculty of Information Technology for supporting this research.

References

[1] K. A. Nugraha and D. Sebastian, "Chatbot Layanan Akademik Menggunakan K-Nearest Neighbor," *Jurnal Sains dan Informatika (JSI),* vol. 7, no. 1, pp. 11-19, 2021.

[2] K. M. Shahih and A. Purwarianti, "Utterance disfluency handling in Indonesian-English machine translation," in *2016 International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA)*, Penang, Malaysia, 2016.

[3] A. Ilmania, S. Cahyawijaya and A. Purwarianti, "Aspect Detection and Sentiment Classification Using Deep Neural Network for Indonesian Aspect-Based Sentiment Analysis," in *2018 International Conference on Asian Language Processing (IALP)*, Bandung, Indonesia, 2018.

[4] S. Vijayarani, M. J. Ilamathi and M. Nithya, "Preprocessing techniques for text mining-an overview," *International Journal of Computer Science & Communication Networks,* vol. 5, no. 1, pp. 7-16, 2015.

[5] M. Allahyari, S. Pouriyeh, M. Assefi, S. Safaei, E. D. Trippe, J. B. Gutierrez and K. Kochut, "A brief survey of text mining: Classification, clustering and extraction techniques," *arXiv preprint arXiv:1707.02919,* 2017.

[6] Z. Yao and C. Ze-wen, "Research on the construction and filter method of stop-word list in text preprocessing," in *2011 Fourth International Conference on Intelligent Computation Technology and Automation*, Shenzhen, China, 2011.

[7] A. Kurbatow, "The research of text preprocessing effect on text documents classification efficiency," in *2015 International Conference" Stability and Control Processes" in Memory of VI Zubov (SCP)*, 2015.

[8] A. I. Kadhim, Y.-N. Cheah and N. H. Ahamed, "Text document preprocessing and dimension reduction techniques for text document clustering," in *2014 4th International Conference on Artificial Intelligence with Applications in Engineering and Technology*, Kota Kinabalu, Malaysia, 2014.

[9] A. Hidayatullah and M. Ma’Arif, "Pre-processing tasks in Indonesian Twitter messages," *Journal of Physics: Conference Series,* vol. 801, no. 1, 2017.

[10] N. K. Jha, "An Approach towards Text to Emoticon conversion and vice-versa using NLTK and WordNet," in *2018 2nd International Conference on Data Science and Business Analytics (ICDSBA)*, Changsha, China, 2018.

[11] A. F. Hidayatullah, "Language tweet characteristics of Indonesian citizens," in *2015 International Conference on Science and Technology (TICST)*, 2015.

[12] K. A. Nugraha and D. Sebastian, "Pembentukan Dataset Topik Kata Bahasa Indonesia pada Twitter Menggunakan TF-IDF & Cosine Similarity," *Jurnal Teknik Informatika dan Sistem Informasi,* vol. 4, no. 3, pp. 376-386, 2018.

[13] D. Sebastian and K. A. Nugraha, "Text Normalization for Indonesian Abbreviated Word Using Crowdsourcing Method," in *2019 International Conference on Information and Communications Technology (ICOIACT)*, Yogyakarta, 2019.

[14] D. Sebastian and K. A. Nugraha, "Sistem Perbaikan Kata Tidak Baku Bahasa Indonesia Menggunakan Metode Crowdsourcing," *Jurnal Teknik Informatika dan Sistem Informasi (JuTISI),* vol. 5, no. 3, 2019.

[15] A. B. Mutiara, E. P. Wibowo and P. I. Santosa, "The Crowdsourcing Method to Normalize “Bahasa Alay”, a Case of Indonesian Corpus," in *2020 Fifth International Conference on Informatics and Computing (ICIC)*, Gorontalo, Indonesia, 2020.

[16] F. Liu, F. Weng and X. Jiang, "A broad-coverage normalization system for social media language," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2012.

[17] W. Trimastuti, "An analysis of slang words used in social media," *Jurnal Dimensi Pendidikan dan Pembelajaran,* vol. 5, no. 2, pp. 64-68, 2017.

[18] D. S. Maylawati and G. P. Saptawati, "Set of Frequent Word Item sets as Feature Representation for Text with Indonesian Slang," in *International Conference on Computing and Applied Informatics*, 2016.

[19] R. B. S. Putra and E. Utami, "Non-formal affixed word stemming in Indonesian language," in *2018 International Conference on Information and Communications Technology (ICOIACT)*, Yogyakarta, Indonesia, 2018.

[20] R. B. S. Putra, E. Utami and S. Raharjo, "Accuracy measurement on Indonesian non-formal affixed word stemming with Levenhstein," in *2019 International Conference on Information and Communications Technology (ICOIACT)*, Yogyakarta, Indonesia, 2019.

[21] Rianto, A. B. Mutiara, E. P. Wibowo and P. I. Santosa, "Improving stemming techniques for non-formal indonesian sentences using incorbiz," *ICIC Express Letters,* vol. 15, no. 1, pp. 67-74, 2021.

[22] N. Hanafiah, A. Kevin, C. Sutanto, Y. Arifin and J. Hartanto, "Text normalization algorithm on twitter in complaint category," *Procedia computer science,* vol. 116, pp. 20-26, 2017.

[23] Rianto, A. B. Mutiara, E. P. Wibowo and P. I. Santosa, "Improving the Accuracy of Text Classification using Stemming Method, A Case of

Informal Indonesian Conversation," *Journal of Big Data,* vol. 8, no. 1, pp. 1-16, 2020.

[24] StarNgage, "Top 1000 Instagram Influencers in Indonesia in 2021," StarNgage, 2021. [Online]. Available: https://starngage.com/app/id/influencer/ranking/indonesia. [Accessed 30 June 2021].

Paper 2

ORIGINALITY REPORT

4%	2%	3%	0%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	knastik.ukdw.ac.id Internet Source	1%
2	Ema Utami, Anggit Dwi Hartanto, Sumarni Adi, Rahardyan Bisma Setya Putra, Suwanto Raharjo. "Formal and Non-Formal Indonesian Word Usage Frequency in Twitter Profile Using Non-Formal Affix Rule", 2019 1st International Conference on Cybernetics and Intelligent System (ICORIS), 2019 Publication	1%
3	Ilmatus Sa'diyah, Tresna Maulana Fahrudin, Latinah, Ibnu Zahy Atha Illah, Cagiva Chaedar Bey Lima, Burhan Syarif Acarya. "Implementation Analysis of KEBI Application on Non-Standard Words Detection in Student Essay Tasks on Language Variety", 2021 IEEE 7th Information Technology International Seminar (ITIS), 2021 Publication	<1%
4	Danny Sebastian, Kristian Adi Nugraha. "Academic Customer Service Chatbot	<1%

Development using TelegramBot API", 2021 2nd International Conference on Innovative and Creative Information Technology (ICITech), 2021
Publication

5	Danny Sebastian, Restyandito -, Kristian Adi Nugraha. "Developing of Middleware and Cross Platform Chat Application", International Journal of Advanced Computer Science and Applications, 2021 Publication	<1 %
6	www.tandfonline.com Internet Source	<1 %
7	"Table of Content", 2021 7th International Conference on Electrical, Electronics and Information Engineering (ICEEIE), 2021 Publication	<1 %
8	Reza Fauzan, Muhammad Ikhwanul Atha Labib, Joshua Oktavianus Tarung Johannis, Herlinawati, Syamsudin Noor, Saifulah. "Semantic similarity of Indonesian sentences using natural language processing and cosine similarity", 2022 4th International Conference on Cybernetics and Intelligent System (ICORIS), 2022 Publication	<1 %
9	dspace.stir.ac.uk Internet Source	<1 %

10	"Artificial Intelligence and Soft Computing", Springer Science and Business Media LLC, 2018 Publication	<1 %
11	Mehdi Ben Lazreg, Morten Goodwin, Ole-Christoffer Granmo. "Combining a context aware neural network with a denoising autoencoder for measuring string similarities", Computer Speech & Language, 2020 Publication	<1 %
12	Rianto, Achmad Benny Mutiara, Eri Prasetyo Wibowo, P. Insap Santosa. "The Crowdsourcing Method to Normalize “Bahasa Alay”, a Case of Indonesian Corpus", 2020 Fifth International Conference on Informatics and Computing (ICIC), 2020 Publication	<1 %
13	Thambusamy Velmurugan, Baskaran Jayapradha. "Emotion Deduction from Social Media Text Data Using Machine Learning Algorithm", Journal of Computer and Communications, 2023 Publication	<1 %

Exclude quotes	On	Exclude matches	< 3 words
Exclude bibliography	On		