ICIC Express Letters
An International Journal of Research and Surveys

May 8, 2020

OFFICIAL RECEIPT (No. Receipt-ICICEL-2004-001)

Name: Rianto

Organization: Universitas Teknologi Yogyakarta, Indonesia

Page Charges Paid: JPY48,000–

This is to acknowledge the receipt of the Publication Paper Charges for the paper "Improving Stemming Techniques for Non-Formal Indonesian Sentences Using Incorbiz" in ICIC Express Letters – An International Journal of Research and Surveys (ISSN 1881-803X), Volume 15, Number 1 (tentative), January 2021.

ICIC-EL Editorial Office
Tokai University, Kumamoto Campus
9-1-1, Toroku, Kumamoto 862-8652, Japan
Tel: +81-96-386-2666
Fax: +81-96-381-7956
E-mail: office@icicel.org
URL: http://www.icicel.org

ICIC Express Letters
Volume **4**, Number **5**, April **2020**

 ISSN 1881-803X
pp. 1–ISII2010-000

IMPROVING STEMMING TECHNIQUES FOR NON-FORMAL INDONESIAN SENTENCES USING INCORBIZ

Rianto[1], Achmad Benny Mutiara[2], Eri Prasetyo Wibowo[2] and P. Insap Santosa[3]

[1]Faculty of Information Technology and Electrical
University Teknologi Yogyakarta
Siliwangi, Yogyakarta, Indonesia
rianto@staff.uty.ac.id

[2]Faculty of Computer Science and Information Technology
University Gunadarma
Margonda Raya No.100, Depok, Indonesia
{ amutiara; eri }@staff.gunadarma.ac.id

[3]Department of Information Technology and Electrical Engineering
Universitas Gadjah Mada
Grafika No. 2, Yogyakarta, Indonesia
insap@ugm.ac.id

Received March 2020; accepted April 2020

Abstract. *Computer technology emulates the human behaviour in presence of Artificial Intelligence (AI) and Natural Language Processing (NLP). The methods make computers act like humans to provide automatic answers according to the human's needs. However, the available tools for the Indonesian language in NLP are very limited, so the linguistic processing of Indonesian on computer is difficult. Linguistic computation requires a process, so the sentences can be analysed using certain methods. The process of data pre-processing consists of case folding, filtering, tokenizing, and stemming. Stemming is one of the potential problems. The development of NLP in Indonesia is provided by stemming techniques named "Sastrawi". The packages can be running well because of stemming on formal Indonesian sentences. On the other hand, most of the Indonesian people are using non-formal sentences in online business conversations like "sdh saya tf tlg segera proses" (I have already transferred, please proceed immediately). This research aims to improve stemming techniques for non-formal Indonesian sentences that has not been done by previous researchers. The results showed Incorbiz can find root words that have not been found by other stemming techniques.*
Keywords: Incorbiz, Indonesian languages, NLP, non-formal sentences, pre-processing, stemming

1. **Introduction.** Customer service is an important activity in business. It has not only to serve customer quickly, but also should be able to enforce customer loyalty. The e-mail addresses or telephone numbers are not suitable means for serving the customers in this area. The customers feel comfortable, when they are accompanied by customer service staff in every transaction. Their communication uses non-formal language, because it is easier for brief dialogues. In online business, the non-formal language is called *"slang"* or *"alay"* language [19, 18].

The *"alay"* language is indicated as a language used by the youth in Indonesia widely. They are often under the age of active internet users in Indonesia, ranging from 18 up to 35 years old. An example of a non-formal sentence is *"Tolongin dong issuedin tiket."* (Please issued a ticket). The word *"Tolongin"* is a non-formal word because in Indonesian grammar, there is no suffix "-in". Many *"alay"* words are abbreviations of formal words. For instance, *"transfer"* (transfer) to *"tf"*, *"gerak cepat* (fast motion) to *"gercep"*, *"mantap betul"* (great) to *"mantul"*, etc. They use these words in online business transactions.

The customer service staff and customers can use the *"alay"* language because they understand each other. The problem arises when the customer service is replaced by a computer program called chatbot. The communication between chatbot and customer service requires a special method, namely linguistic computation. Linguistic computation uses formal sentences as the knowledge for the chatbot to serve the customers. Finally, the chatbot knowledge must be upgraded to understand non-formal sentences to provide an excellent service to customers.

A technology developed for linguistic computation is Natural Language Processing (NLP). NLP has limitations in multi-language processing because NLP is prepared for English language. For English processing, NLP has high accuracy, so this technology is widely used for text analysis. Text analysis in NLP consists of text categorization, summarization, sentiment analysis, etc. NLP lacks knowledge of Indonesian language, especially in non-formal sentences, so researchers have to provide a corpus that can help the problem of stemming in Indonesian non-formal sentences.

Incorbiz (Indonesian Closed Corpus for Business) is a corpus particularly designed for online ticket reservations that contain Indonesian formal and non-formal words. Incorbiz is a solution to the problem especially in stemming section for non-formal Indonesian sentences. Stemming is a process to reduce words to their roots, for instance the word *"berlari"* (running) will be reduced to the root word *"lari"* (run). The objectives of the stemming are to minimize the number of word variants, because the words *"berlari"* and *"lari"* have basically the same meaning.

Incorbiz uses a master-detail approach to store data. Master contains words and the types of words while details contain word variants. The word *"batal"* as a master word is *"dibatalkan", "membatalkan", "dibatalin", "btl", "dicancel", "dicancelin"* in the detail. The example indicates that Incorbiz combines Indonesian words, loanwords, and abbreviations based on the root word. It makes the result of Incorbiz stemming better than other tools because of other tools will not find the root word of *"dicancel"* or *"dicancelin"*.

The development of NLP in Indonesia is marked by an Indonesian-language stemmer called *"Sastrawi"* [20]. In formal sentences, *"Sastrawi"* is very successful. For example, the word *"ditulis"* (written) will be reduced to the root word, i.e. *"tulis"* (write). *"Sastrawi"* will not perform well if applied for non-formal sentences. The sentences *"Tolong cancelin tiketnya"* (Please cancel the ticket) are containing non-formal sentences, i.e. *"cancelin"*. *"Sastrawi"* will not find the root word of *"cancelin"* even though the word *"cancelin"* derives from the root word *"batal"*. Another stemmer that supports Indonesian stemming is called spaCy [21]. However, spaCy is not different from *"Sastrawi"* in that it cannot provide stemming in Indonesian non-formal sentences.

The limitations make this research on stemming for non-formal Indonesian necessary to conduct. The solution proposed is an Incorbiz corpus containing root words and their word variants consisting of Indonesian, loanwords, and abbreviations. Incorbiz is a dynamic corpus to anticipate out of vocabulary.

2. Method and Related Work.

2.1. Method.

The Incorbiz data set is collected from conversations between customer service staff and customer on an online flight ticket reservation, by OkeTiket via WhatsApp's messenger. The conversation data from WhatsApp's Messenger are converted into text files, to prepare for the text analysis. The pre-processing steps consist of Case Folding, Filtering, Tokenizing, and Stemming. The data pre-processing is stemming and normalization. Filtering and Stemming is not done automatically by using programming language, so it is done semi-manually under human assistance.

Normalization must be completed semi-manually to identify tokenization results. Examples of tokens that are not used as corpus data include booking codes, passenger

names, flight dates, and so on. Due to two main problems the semi-manual process must be carried out. The problems are 1) the algorithm to find out if the token is a word, 2) stemming for non-formal words. This problem is related to a problem that has been previously stated, i.e. the lack of resources and NLP tools for the Indonesian language.

After data processing is completed, it is converted to a table. Incorbiz uses MySQL as a database server and will be developed using MongoDB in future. Tokenizing process produces over 25,000 tokens with non-normalised data. After normalizing and checking to the *"Kamus Besar Bahasa Indonesia" (KBBI)* [22], there are 1,009 tokens that could be identified as Indonesian words. The count of words is relatively small, so it cannot encompass all the words of the Indonesian language. Incorbiz develops a dynamic corpus approach. Although using MySQL, Incorbiz concept is a non-relational database, because non-relational database has a relatively high processing speed [17].

Incorbiz saves word variants that are used in dialogs including the *"alay"* words. The word *"batal"* has a variant of the words *"dibatalkan", "membatalkan", "pembatalan", "batalin", "dibatalin", "batalkan"*. The words *"dibatalkan", "membatalkan", "pembatalan"*, and *"batalkan"* can be resolved by stemming using *"Sastrawi"*. For the words *"batalin"* and *"dibatalin"* cannot be found their root words. Finally, the existence of Incorbiz with a dictionary approach is expected to solve the stemming problem for the *"alay"* language.

The word variants written according to root words is done semi-manually with human assistance by using knowledge from various disciplines, such as language and literature, communication, and psychology. Besides lecturers, some students from the informatics department were also involved to add and check word variants that might be used by customers in business conversation. There are 10 people that are developing Incorbiz. The researcher uses *"Kamus Besar Bahasa Indonesia"* [22] to check a word type.

2.2. **Related Work.** Stemming is an important process in text analysis to get accurate results [6]. The high accuracy level will affect the quality of predictions for sentiment analysis, questions and answers, text classification, etc. Unfortunately, NLP resources and tools for Indonesian are very limited. In addition, there are significant differences in terms of morphology between Indonesian and English [7]. Indonesian sentence has a different structure what adds complexity to its linguistic computational processing [15]. The main issue in Indonesian text processing research is the use of the *"alay"* sentences which causes errors in the stemming process.

The research of improving stemming for Indonesian was done by analysing texts from social media. This research was conducted to provide Indonesian stemming for non-formal sentences which are often called *"Slang"* or *"alay"*. It is a very important research because on social media such as Twitter, Facebook, Instagram and others, they use the *"slang"* or *"alay"* sentences. The purpose of this study consists of two factors, i.e. 1) using dictionaries and 2) accommodating existing methods by strengthening stemming. The result of this research shows 379 words, and 20 text data with an accuracy of 88.65% [1].

The other research of Indonesian stemming for the *"alay"* words was carried out by improving the Nazief and Andriani's algorithm by using Flexible Affix Classification. This research was conducted because the Nazief and Andriani's algorithm have lower accuracy when it was applied to detect the *"alay"* words. The algorithm was improved by adding non-formal affix rules comprising of 6 non-formal affix rules. The non-formal affix rules are 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). By testing of 60 words *"alay"* has resulted in an accuracy rate of 73.33% [2].

The limitations on Indonesian stemming studies that have been done previously, motivated the new research. Nazief and Andriani's algorithm and non-formal affix rules are

developed by using the Levenshtein Distance algorithm. Development is done by combining the Levenshtein Distance algorithm and the Indonesian dictionary. The results show that algorithm has increased the accuracy to 88.33% [3].

The three researches that conducted Indonesian stemming for the *"alay"* words showed good progress in increasing accuracy. The method is developed to integrate the algorithms and new rules that are defined both through programming techniques and dictionaries. But the research tried to improve the analysis of *"alay"* words by determining the prefix, insert, and suffix only. An *"alay"* word is not only formed from the prefix, insert, and suffix, but also it can be deformed. An example is *"transfer"* (transfer) to *"tf"*, (one way) to *"ow"*, *"sendiri"* (on the way) to *"otw"*, etc. By stemming methods for the words above, we cannot find the root words. This problem becomes a reason for conducting a new research on stemming methods for non-formal sentences. This is a combination of Indonesian stemming techniques and a closed corpus called Incorbiz. The stemming method in this research is not only for non-formal Indonesian, but also for formal Indonesian.

3. **Result and Discussion.**

3.1. **Result.** The result is 1) developing a closed corpus Incorbiz and 2) implementing the stemming approach by combining *"Sastrawi"* and Incorbiz. In the corpus, Incorbiz has 1,009 words that were obtained from conversations between customer service and OkeTiket customers. A word variant was obtained through analysis and normalization of 5,120 words. The collection of words in Incorbiz will dynamically increase according to their usage. The addition is done automatically by matching existing collections system, while semi-automatically it is done by humans entering the data.

The approach is using a dictionary by considering the resources and processing tools of the *"alay"* language. Besides, dynamic corpus will complete the database of formal and non-formal words. It is very important because it can be implemented for many purposes, for example in online business sector, to provide the best services to customers. The combination of formal and non-formal language data will improve the learning process of Indonesian, especially for root words and word types.

Data saving concept in Incorbiz is performed in non-relational databases by using denormalization approach. So, in searching the data, sub-queries or joins table are not needed. In the next development, Incorbiz will use NoSQL to accommodate structured, semi-structured, and unstructured data [9]. The aim is to speed up the retrieving of information. An example for storing of records in Incorbiz is presented in Table 1.

TABLE 1. Examples of Incorbiz Data

id	root word	business domain	word type	word variant	english
1	batal	general	adj	batal;batalin;btl;btlkan	cancel
2	lambat	general	adj	lambat; terlambat;lambatin	slow
3	terbang	airline	verb	terbang;penerbangan;	flight
4	cetak	airline;train	verb	cetak;cetakkan;cetakin	issue
5	kamar	hotel	noun	kamar;kamarnya	room

Incorbiz has two differences compared to another corpus. They are in the business domain and word variants field. The business domain field is developed due to plenty of specific words within a business category. The word *"kamar"* (room) is specific for hotel reservations, because airline or train ticket reservations will not use the word *"kamar"*. In general category, it is available in Incorbiz, to be implemented in general conversations. The word *"pagi"* (morning) is not only used specifically for airline, train, or hotel reservations, but also for common conversations. The purpose of business domain field is to facilitate the application in different business categories.

In writing word variants, Incorbiz uses the semicolon delimiter (;) as separator between subsequent words. The foreign words in the Incorbiz will be added in the new vocabulary according to the root words in Indonesian. For example, the word "cancel" will be added as a word variant, by the root word *"batal"*. Even foreign words are abbreviated, for example, adult to *"adt"*, one-way to *"ow"*, passenger to *"pnr"*, etc. The abbreviations are included in word variants according to the root words and their meanings in Indonesian.

3.2. **Discussion.** As a closed corpus for online businesses, Incorbiz has a word list, both formal and non-formal, including its variants of the words. Specifically, Incrobiz is more dedicated to stemming non-formal sentences. Incorbiz works by combining two approaches, namely stemming algorithms and dictionary. The Incorbiz's workflow is shown in Fig. 1.

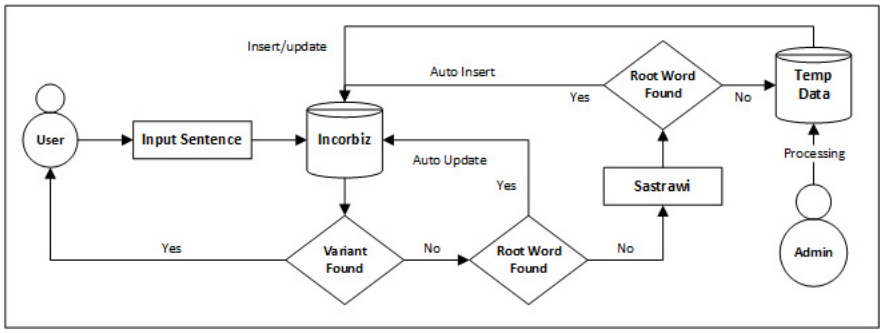

FIGURE 1. Workflow of Incorbiz

After a user enters sentences, the system will be pre-processing consist of case folding, tokenizing, stop word removing, and stemming using Incorbiz to make tokens and searching among the word variant fields to find the corresponding root word. If the system finds root words, Incorbiz would give the results to the user, but if not found, Incorbiz would proceed further. In the case of word variants not found, Incorbiz will continue to check in the root word fields using constraint "like=token". If the root word found, the token will be inserted as word variant in the root word accordingly. However, if the root word not found, the tokens will be inserted into temporary table and will be processed manualy with human asistance. For example, if the user writes *"met pagi pak, dibantu cari tiket dong"* (Good morning sir, get me a ticket please), and Fig. 2 shows the output.:

Word	Root Word	Word Type	English
met	selamat	adj	greeting
pagi	pagi	noun	morning
dibantu	bantu	verba	help
cari	cari	verba	to find
tiket	tiket	noun	ticket

FIGURE 2. The result of experiment #1 using Incorbiz

The stemming result in Fig. 2 explained that *"met pagi pak, dibantu cari tiket dong"* the pre-processing will yield the tokens including *"met"*, *"pagi"*, *"bantu"*, *"cari"*,*"tiket"*. The word *"pak"* and *"dong"* do not get processed because *"pak"* and *"dong"* are included in the stop words list. By default, the stop word list contains words such as *"di"* (in), *"ke"* (to), *"dari"* (of), *"pada"* (on), and so on [10]. The stemming result contains: *"selamat"*,

"pagi", "bantu", "cari", and *"tiket"*. The word *"met"* is a non-formal word with the root word *"selamat"* that was discovered by Incorbiz. As a comparison, the sentence is processed using *"Sastrawi"*. The results of *"Sastrawi"* is shown in Fig. 3.

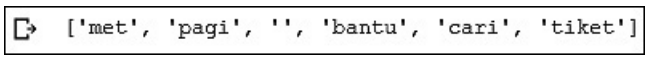

FIGURE 3. The result of experiment #1 using *"Sastrawi"*

Between Fig. 2 (Incorbiz) and Fig. 3 (*"Sastrawi"*) there are differences in finding the root words. The word *"met"*, Incorbiz translates as *"selamat"* following the Indonesian, but in *"Sastrawi"*, the word *"met"* was not found. This is an example of a difference between Incorbiz and other stemmers.

The second experiment was conducted with different non-formal sentences and it will be processed using two different stemmers, namely Incorbiz and *"Sastrawi"*. The sentence entered is *"tlg batalin tiket saya"* (please cancel my ticket). For the words *"tlg"* and *"batalin"* will not be found the root words, because Indonesian stemming is limited to formal words. The comparison results between Incorbiz and *"Sastrawi"* is shown in Fig. 4 and Fig. 5.

Word	Root Word	Word Type	English
tlg	tolong	verb	help
batalin	batal	adj	cancel
tiket	tiket	noun	ticket

FIGURE 4. The result of experiment #2 using Incorbiz

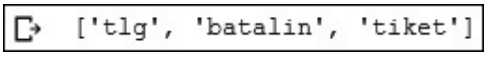

FIGURE 5. The result of experiment #2 using *Sastrawi*

The experiment results show that the stemming process with *"Sastrawi"* for both non-formal words failed, while Incorbiz could find the root words for the two non-formal words because the sentences are stored in the Incorbiz word collection. So, what if the word is not stored in the Incorbiz collection?

In Fig. 1 is the Incorbiz workflow. There are three possibilities in searching for a word. The first possibility occurs when the word variants and root words are found. The second, the word variants are not found but root words are found, and the third, when both of them are not found. The second and third possibilities require a process. The search results on the first possibility is shown in Fig. 2 and Fig. 4.

The sentence in the third experiment is *"berapa harga pertiketnya?"* (What is the ticket price?) to be processed by using Incorbiz. The stemming process only produces the words *"harga"* and *"tiket"* because *"berapa"* and "?" are included in the set of stop words so they are not processed. The results of the third experiment are shown in Fig. 6.

Experimental results in Fig. 6 show that Incorbiz runs the second possibility on the word *"pertiketnya"*, it has found the root word *"tiket"* but it does not find the word variant *"pertiketnya"*. This is indicated by the "auto-update" information after adding data with Incorbiz. It can be proven by conducting further experiment#4 using the same sentence *"Berapa harga pertiketnya?"*.

Word	Root Word	Word Type	English
harga	harga	noun	price
pertiketnya	tiket (auto update)	noun	ticket

FIGURE 6. The result of experiment #3 using Incorbiz

Word	Root Word	Word Type	English
harga	harga	noun	price
pertiketnya	tiket	noun	ticket

FIGURE 7. The result of experiment #4 using Incorbiz

The experimental results in Fig. 7 show that there is no "auto-update" information. It shows that Incorbiz proceeds according to the first possibility by finding the word variant *"pertiketnya"* from the root word *"tiket"*. The results in Fig. 7 prove that Incorbiz can do the "auto-updates" conditionally. The conditions need root words of the word variants.

Incorbiz will use the third possibility if the root word is not found. In the third possibility, Incorbiz will store words in a temporary table for further processing. In the temporary table, there are three menus options, i.e. delete, set as a root word, and set as a word variant. The delete menu is used by admin to delete data. The set as a root word menu is used to set the word as root word, while set as a word variant is used to set the word as word variant. The process requires human assistance because there is no algorithm yet to solve the third possibility. The next process is developing the algorithm to add the data automatically.

4. **Conclusion.** The research is comparing *"Sastrawi"* and Incorbiz in stemming "Indonesian" on formal and non-formal sentences. In several researches, *"Sastrawi"* is used to analyze the Indonesian text commonly [11, 12, 13, 14]. On the other side, *"Sastrawi"* is not designed for processing non-formal sentences. So, the root word cannot be found. But Sastrawi can be combined to support Incorbiz in the stemming process successfully. The future work is how to add the word collections according to the word type, word variants, and root words automatically. Finally, an algorithm is needed to add the data automatically.

Acknowledgment. This research was supported by the Ministry of Education and Culture of the Republic of Indonesia. The researcher thanks to OkeTiket Yogyakarta, Indonesia who gives suggestion and data, to Informatics and Information System students of Universitas Teknologi Yogyakarta for editing the corpus data and the Communication, Psychology, and Linguistic Study Programme of Universitas Teknologi Yogyakarta for great advising.

REFERENCES

[1] D. S. Maylawati, W. B. Zulfikar, C. Slamet, M. A. Ramdhani and Y. A. Gerhana, An Improved of Stemming Algorithm for Mining Indonesian Text with Slang on Social Media, *2018 6th International Conference on Cyber and IT Service Management (CITSM)*, Parapat, Indonesia, pp.1-6, 2018.

[2] R. B. S. Putra and E. Utami, Non-formal affixed word stemming in Indonesian language, *2018 International Conference on Information and Communications Technology (ICOIACT)*, Yogyakarta, Indonesia, pp.531-536, 2018.

[3] R. B. Setya Putra, E. Utami and S. Raharjo, Accuracy Measurement on Indonesian Non-formal Affixed Word Stemming With Levenhstein, *2019 International Conference on Information and Communications Technology (ICOIACT)*, Yogyakarta, Indonesia, pp.486-490, 2019.
[4] E. Utami, A. D. Hartanto, S. Adi, R. B. Setya Putra and S. 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)*, Bali, Indonesia, pp.173-176, 2019.
[5] N. Bhartiya, N. Jangid, S. Jannu, P. Shukla and R. Chapaneri, Artificial Neural Network Based University Chatbot System, *2019 IEEE Bombay Section Signature Conference (IBSSC)*, Mumbai, India, pp.1-6, 2019.
[6] E. Nugraheni, Indonesian Twitter Data Pre-processing for the Emotion Recognition, *2019 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, Yogyakarta, Indonesia, pp.58-63, 2019.
[7] U. Hasanah, T. Astuti, R. Wahyudi, Z. Rifai and R. A. Pambudi, An Experimental Study of Text Preprocessing Techniques for Automatic Short Answer Grading in Indonesian, *2018 3rd International Conference on Information Technology, Information System and Electrical Engineering (ICITISEE)*, Yogyakarta, Indonesia, pp.230-234, 2018.
[8] Alfina, I. Sigmawaty, D. Nurhidayati, F. Hidayanto, and A. Nizar, Utilizing Hashtags for Sentiment Analysis of Tweets in The Political Domain, *Proceedings of the 9th International Conference on Machine Learning and Computing*, Singapore, Singapore, pp.43-47, 2017.
[9] B. Jose and S. Abraham, Exploring the merits of nosql: A study based on mongodb, *2017 International Conference on Networks Advances in Computational Technologies (NetACT)*, Thiruvanthapuram, India, pp.266-271, 2017.
[10] Z. Jianqiang and G. Xiaolin, Comparison Research on Text Pre-processing Methods on Twitter Sentiment Analysis, *IEEE Access*, vol.5, no., pp.2870-2879, 2017.
[11] D. R. Kiasati Desrul and A. Romadhony, Abusive Language Detection on Indonesian Online News Comments, *2019 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, Yogyakarta, Indonesia, pp.320-325, 2019.
[12] A. A. P. Ratna, N. A. Wulandari, A. Kaltsum, I. Ibrahim and P. D. Purnamasari, Answer Categorization Method Using K-Means for Indonesian Language Automatic Short Answer Grading System Based on Latent Semantic Analysis, *2019 International Electronics Symposium (IES)*, Padang, Indonesia, pp.1-5, 2019.
[13] D. Dwimarcahyani, T. Badriyah and T. Karlita, Classification On Category Of Public Responses On Television Program Using Naive Bayes Method, *2019 International Electronics Symposium (IES)*, Surabaya, Indonesia, pp.225-231, 2019.
[14] A. A. Putri Ratna, H. Khairunissa, A. Kaltsum, I. Ibrahim and P. D. Purnamasari, Automatic Essay Grading for Bahasa Indonesia with Support Vector Machine and Latent Semantic Analysis, *2019 International Conference on Electrical Engineering and Computer Science (ICECOS)*, Batam Island, Indonesia, pp.363-367, 2019.
[15] D. Anggraini, A. B. Mutiara, T. M. Kusuma and L. Wulandari, Algorithm for Simple Sentence Identification in Bahasa Indonesia, *2018 Third International Conference on Informatics and Computing (ICIC)*, Palembang, Indonesia, pp.1-6, 2018.
[16] E. P. Wibowo, A. R. Rahman and A. Muslim, Analysis of Speech Recognition in Conversation Bot for Interactive Media Using Automated Log, *Journal of Theoretical and Applied Information Technology*, vol.97, no.4, pp.1267-1277, 2019.
[17] Hanen Abbes and Faiez Gargouri, Big Data Integration: A MongoDB Database and Modular Ontologies based Approach, *Procedia Computer Science*, vol.96, no., pp.446 - 455, 2016.
[18] A. F. Hidayatullah, Language tweet characteristics of Indonesian citizens, 2015 International Conference on Science and Technology (TICST), Pathum Thani, 2015, pp. 397-401, doi: 10.1109/TICST.2015.7369393.
[19] N. Aliyah Salsabila, Y. Ardhito Winatmoko, A. Akbar Septiandri and A. Jamal, Colloquial Indonesian Lexicon, 2018 International Conference on Asian Language Processing (IALP), Bandung, Indonesia, 2018, pp. 226-229, doi: 10.1109/IALP.2018.8629151.
[20] Sastrawi 1.0.1, https://pypi.org/project/Sastrawi/,2020 (accessed March 3, 2020).
[21] spaCy, https://spacy.io/,2020 (accessed March 7, 2020).
[22] KBBI Daring, https://kbbi.kemdikbud.go.id/,2020 (accessed February 20, 2020).