Improving the Accuracy of Text Classification using Stemming Method, A Case of Non-Formal Indonesian Conversation

Rianto Rianto (rianto@staff.uty.ac.id)
Universitas Teknologi Yogyakarta https://orcid.org/0000-0002-5058-4580
Achmad Benny Mutiara
Universitas Gunadarma
Eri Prasetyo Wibowo
Universitas Gunadarma
Paulus Insap Santosa
Universitas Gadjah Mada

Short report

Keywords: accuracy, Indonesian, stemming, text processing

Posted Date: January 11th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-41431/v3

Version of Record: A version of this preprint was published on January 29th, 2021. See the published version at https://doi.org/10.1186/s40537-021-00413-1.

SHORT REPORT

Improving the Accuracy of Text Classification using Stemming Method, A Case of Non-formal Indonesian Conversation

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

[*]Correspondence:
rianto@staff.uty.ac.id
[1]Faculty of Information Technology and Electrical Engineering, Universitas Teknologi Yogyakarta, Siliwangi St., 55285 Yogyakarta, Indonesia
Full list of author information is available at the end of the article

Abstract

Background: Stemming has long been used in data pre-processing to retrieve information by tracking affixed words back into their root. In an Indonesian setting, existing stemming methods have been observed, and the existing stemming methods are proven to result in high accuracy level. However, there are not many stemming methods for non-formal Indonesian text processing. This study introduces a new stemming method to solve problems in the non-formal Indonesian text data pre-processing. Furthermore, this study aims to improve the accuracy of text classifier models by strengthening stemming method. Using the Support Vector Machine algorithm, a text classifier model is developed, and its accuracy is checked. The experimental evaluation was done by testing 550 datasets in Indonesian using two different stemming methods.

Findings: The results show that using the proposed stemming method, the text classifier model has higher accuracy than the existing methods with a score of 0.85 and 0.73, respectively. These results indicate that the proposed stemming methods produces a classifier model with a small error rate, so it will be more accurate to predict a class of objects.

Conclusion: The existing Indonesian stemming methods are still oriented towards Indonesian formal sentences, therefore the method has limitations to be used in Indonesian non-formal sentences. This phenomenon underlies the suggestion of developing a corpus by normalizing Indonesian non-formal into formal to be used as a better stemming method. The impact of using the corpus as a stemming method is that it can improve the accuracy of the classifier model. In the future, the proposed corpus and stemming methods can be used for various purposes including text clustering, summarizing, detecting hate speech, and other text processing applications in Indonesian.

Keywords: accuracy; classification; Indonesian; stemming; text processing

Introduction

As social beings, humans always interact with one another. The interactions were carried out in verbal or non-verbal language. Language is an arbitrary sound symbol system, which is used by members of a community to cooperate, interact, and identify themselves [1]. This definition implies that language has a special character which is the identity of a country and domain of a dialog topic. In verbal communication, people use sentences in the forms of consist of words or series of words to express a complete meaning. Indonesian language is classified into two categories, namely formal and non-formal in the method of use. Indonesian formal is used in

formality situation, while Indonesian non-formal is widely used in a casual situation like in social media conversations [2]. In formal language, people use "standardized" language as prescribed by linguistic rules of the language. However, for communication, especially in casual conversations, social media, or non-formal discussions, people often tend to communicate using non-formal language [3].

In online conversation, however, the language category that is mostly used is non-formal language, namely "Bahasa Slang". "Bahasa slang" is a term used to refer to a non-formal language that is formed from abbreviations, terms, or a combination of both [4]. The simplicity of non-formal language makes people tend to use it in their communication. Communication uses non-formal language is brief, but it can be easily understood by every person. As mentioned above, non-formal language tends to deviate grammatical rules. The example of deviations is found in suffixation and abbreviated words. In non-formal Indonesian correspondence, words are shortened by omitting the vowels, for example "saya" (I) is abbreviated as "sy", "tolong" (help) abbreviated as "tlg", "bayar" (pay) abbreviated as "byr", and so on. The suffixes that are not a part of Indonesian grammar i.e. "in" make the formal language transform into non-formal language [5], for example "bayarin" (pay), "tolongin" (help), "cetakin" (print), and so on. In Indonesian grammar "in" belongs to infixes that consist of "-er-", "-el-", "-em-", and "-in-" [6]. Beside the suffixes and abbreviations, the non-formal language is also formed using loanword, for examples "cancel", "booking", "issued", and so on. The others Indonesian non-formal are the words that are deformed from the original word, for example "sendiri" (alone) into "jomblo", "santai" (relax) into "santuy", "lambat" (slow) into "lambreta", and so forth.

In linguistic computation, the non-formal language comes to problems in data pre-processing which mostly comprises tokenizing, removing, stemming, and normalizing. Among the problems in data pre-processing, however, stemming is a key factor to prepare the data to be analyzed to get an accurate result [7]. Stemming is the process to remove affixes to get the root words, for example "berlari" (running) into "lari" (run), "menulis" (writing) into "tulis" (write), "memakan" (eating) into "makan" (eat), and so on [8]. The sample words mentioned above are formal words that do not have a problem in stemming. The problem, however, is in non-formal, the words which are deviated from Indonesian standard words, for example in sentences like "sistemnya lambreta bingit" (the system is very slow). "Lambreta" and "bingit" are not formal Indonesian words, so they will raise a problem in the stemming process. Formal words in Indonesian are words that are contained in the Big Indonesian Dictionary (KBBI).

Some research on Indonesian text processing had been done by using the existing Indonesian stemmer, i.e. "Sastrawi". The research and improvements of "Sastrawi" algorithm were also done by using dynamic affixes to process non-formal Indonesian. However, the result of their research had limitations on non-formal Indonesian expressions, which were formed into abbreviated and deformed words.

This research was the first step of three-phase research concerning empathetic chatbots for airline ticket reservations. The stages of the research include 1) Indonesian closed corpus development; 2) Emotion and intent classification development; and 3) Chatbot development. The purpose of the first step was to develop an Indonesian closed corpus that can be used as a dictionary and stemming method on

formal and non-formal Indonesian. The term closed corpus shows that the developed corpus has a special purpose in a particular domain. In this study, the corpus developed should process sentences on airline ticket reservations, which will apply to chatbots as a substitute for customer service staff. This corpus is important for the stemming process, so it can produce a classifier model that has high accuracy. High accuracy will affect the prediction accuracy for the chatbot to respond to user requests. On the other hand, a corpus is an important part of language synthesis systems which will affect in synthesis quality of speech [9]. However, the Indonesian corpus for non-formal Indonesian especially for "bahasa slang" is still incomplete, so this corpus is needed.

Related Work

Indonesia is a country of about 17.000 islands, 1.300 ethnicities, and 742 languages. This condition requires a unifying means, and one of which is the national language called Bahasa (Indonesian language). In computational linguistics, however, it lacks resources for Indonesian language processing, including the area of Part-of-Speech (POS) tagger [10]. The POS tagger is used to assign a tag the class of word that can process text in NLP such as text classification, text summarization, information retrievals, and so on.[11].

Related to the lack's resources for Indonesian language processing, research in Indonesian POS (Part-of-Speech) Tagger was developed. Kwary had developed an educational corpus platform and divided into four disciplines according to user groups i.e. health, life, physics, and social science. This corpus collects about 5 million words obtained from several Indonesian university's journals, namely Airlangga, Diponegoro, Lampung, and Udayana. Data were taken by downloading 10 to 30 articles per journal. This corpus is claimed to be an Indonesian academic corpus that can be free to be accessed. Teachers and experts of Indonesian language were the target users of this corpus [12].

Research that has a correlation with the Indonesian language processing is topic summarization. Jiwanggi and Adriani had conducted research to summarize Twitter data in Indonesian. The algorithm used was "The Phrase Reinforcement". This algorithm summarized a group of tweets that discussed similar topics by using a semi-abstractive approach. The data were obtained from tweeters with recorded time about 1 - 4.5 hours and collected over 100.000 tweets. The analysis had been done by humans to assess their readability, grammaticality, and informative level. There were 37 people comprising undergraduate and graduate students to evaluate the summaries. The evaluations showed that over 60% of the total positive answers were readability, grammaticality, and informative level [13].

The lack of resources regarding Indonesian in linguistic computing makes an idea for researcher to translate Indonesian into English. The reason is that there are more libraries and packages available in English. Gunawan, Mulyono, and Budiharto had researched the arithmetic word problem in questions and answers. The approach to problem-solving was to translate Indonesian into English and then processed it with Natural Language Toolkit (NLTK). The translation process was conducted by utilizing Google Translate Application Programming Interfaces (API). Although it had a good accuracy in doing answers, which was between 80% - 100%, the speed

of the process was low due to the language-translation process up to 1.12 minutes [14].

The completeness of dictionary resources inspired research by Gunawan et al., regarding the translation of Indonesian in English. There are more data sets in English than Indonesian data sets. Unlike English, Indonesian has a more complex diversity because of many regional languages in Indonesia. Research on the classification of British and American English can be done easily because of the availability of the British National Corpus (BNC) and the Corpus of Contemporary American English (COCA) [15] [16]. Support Vector Machine (SVM) with the linear kernel is an algorithm that is often used for text processing. This is related to the advantages of SVM, which has a maximum hyperplane [17].

The application of SVM in text classification is generalized, so it is no different for both English or Indonesian, but what is different is the supporting data set. Due to the limited resources of Indonesian language data sets, especially for non-formal languages, a corpus such as BNC or COCA is needed so that it can be used for Indonesian text processing with better accuracy to develop a classifier model that can be implemented as knowledge base for a conversational agent.

The major problem in developing a corpus, however, is text normalization and verification that cannot be done automatically by using computer technology. Text-normalization was done manually by using human assistance. Further the text normalization was verified using the Indonesian online dictionary, namely "Kamus Besar Bahasa Indonesia" (KBBI).

Sebastian and Nugraha had done text normalization in non-formal Indonesian. They developed a dataset to normalize abbreviated words. Abbreviated words were one problem in text normalization. An ambiguity occurred as the key factor in it, so they cannot be processed optimally. This research used "Crowdsourcing" to develop a dataset. This method was selected because only humans can translate abbreviated words into normal forms of words. The result showed the level of accuracy was 90.85%, however, there was a problem related to abbreviated words when there had been more than one meaning. Unique keywords were needed to determine which meaning was most accurate [18].

Micro texts are a term of communication in the digital era which are defined as an expression of an idea using in a short text. It is applied in the short message service, which is limited by several characters. The limitation makes a user abbreviate the sentences or words so that it does not exceed the limit of the character number. Like the previous research, which was conducted by Sebastian and Nugraha, a research in text normalization was also done by Gunawan, Saniyah, and Hizriadi. This research used the approach of dictionary and Longest Common Subsequence (LCS). The result showed that their approach could solve the problem related to abbreviation, however, it is limited to pre-defined abbreviations and acronyms [19].

Apart from being used for expression, micro text is also used for questions and answers via short message services (SMS). The major problems, in this case, are the same questions meaning in different styles and formats that causing questions flooding. A study was conducted to solve the problem to increase the effectiveness of automatic answering machines. The approach taken in this study is to detect the ratio of the similarity between prefix and suffix in each question sentence. This study

also uses Supervised Machine Learning by implementing the Naive Bayes Theorem to improve the results. Data matching was done by calculating the probability of successive occurrences based on the FAQ training data. The results showed that the proposed model has an accuracy value of 0.863, while "Kothari" and "Shivhre" of 0.618 and 0.654, respectively [20].

The abundance of data that is automatically produced by social media such as Twitter, Facebook, Instagram, and others often confuses users in understanding information. Many reporting systems use social media to make it easier for users to submit reports related to public facilities such as transportation, accidents, incidents, and others. Flooding of data as a result of information made confused the officer to follow up on it. An approach using keywords does not solve this problem, so it takes a different, more elegant method to solve the problem. One method proposed is semantics related to the context of the problem. However, semantics still have limitations in terms of ambiguity because of the similarities in each report, so a hybrid method is proposed by combining NER, POS, RE to extract textual information. The results show that the proposed method can solve the deficiencies of previous information extraction methods [21].

The extraction of information is not only useful for taking appropriate actions in dealing with problems but also can analyze a person's personality. It is can saw the appropriateness of policies issued by a leader for their territory. The data set used in this study came from 5.3 million users that have posted over 20 million tweets during natural disasters. Based on the data set, the classifications of leadership are developed. The results of the study are in the form of useful insights that the characteristics of the leadership are influenced by hazard, domain, and user centrality. The results are important for assessing the quality of prospective leaders based on their characteristics [22].

The previous research as mentioned above shows that data pre-processing is an important factor that has an impact on data quality. In the case of text analysis, a text normalization is the part of data pre-processing that had been done by the previous researcher. The result showed that the approach using a dictionary improved the data quality effectively. Out of vocabulary, however, is the main issue that must be solved in the dictionary approach. Under the research conducted by Agarwal and Durga, the completeness of the data set in the form of an Indonesian corpus is also very important in emotional engineering to predict human characters and emotions through the sentences it conveys. This is very important for conversational agents to provide a better response.

Method

Raw data in this research were collected from the conversation between customer service staff and consumers in airline ticket reservations, namely "OkeTiket" through WhatsApp messenger. By using this method, vocabularies that widely used in airline ticket reservations will be collected. That is important because it is a rule in language learning [23]. Data were converted into Comma-Separated Values (CSV) to facilitate further processing using the Python programming language, which has high compatibility with CSV files [24, 25, 26, 27]. The step-by-step data pre-processing is shown in Figure. 1.

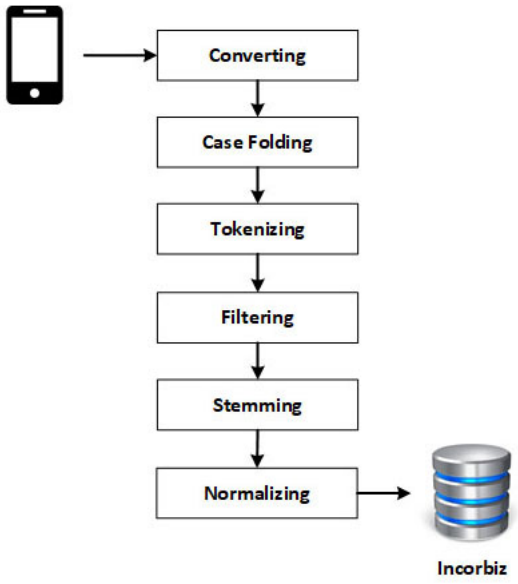

Figure 1 Step by step data preprocessing.

Officially "OkeTiket" uses the WhatsApp service to communicate with its members. The data generated from this communication is a text file that is entirely "OkeTiket's" property, so that there is no legal violation in using the data for research. The process of data acquisition was done manually by exporting data per member using a smartphone. The result of this activity are 200 files in text format with variations in the number of conversations that vary from one file to another. These files were combined in Microsoft Excel and saved as into CSV format.

Python programming is highly recommended for natural language processing because it has a library to be used, i.e. Natural Language Toolkit (NLTK). By using NLTK, pre-processing activities such as case folding, tokenizing, and stop word removal can easily be done. This is one of the reason Python programming language was used in this study.

In the stemming activity, however, researcher provided an alternative tool which was compared with the existing Indonesian stemming method. The existing stemming method which is widely used for Indonesian language processing is "Sastrawi". However, "Sastrawi" has a weakness in non-formal Indonesian language processing, so refined to produce the right data are needed. An alternative stemming method is dictionary-based, which has a collection of words in Indonesian both formal and non-formal. The method was used in stemming process is string matching by using Structure Query Language.

Following the data converting, the next process is case folding. A case folding is a process to convert letters in the text documents into uppercase or lowercase. In-text processing, however, case folding was used to convert letters into lowercase. Its processes were done by using "lower ()" as a function of NLTK. The next process was tokenizing to split sentences into tokens by using space detection. Its processes resulted in about 74.477 tokens which were verified in Indonesian rules correctly.

According to the step of data pre-processing in Fig.1, verifications were done by using a process namely "Filtering".

The filtering activity cannot be fully carried out using algorithms because it deals knowing that only humans know. For example, to detect a token is a word or not a word. The tokens consist of passenger names, email addresses, booking codes, and so on. In the case of email addresses, detecting can be done by using a regular expression with the marker "@", but for the name of the person or the booking code, algorithm and technique to identify it was not available yet. Related to this problem, a manual process by human assistance to identify the tokens are needed.

Following the activity process in filtering is stop word removal and cleared tokens from number, special characters, and punctuations. Both processes are an important factor to eliminate grammatical useless words in Indonesian like "di" (in), "pada" (on), "dari" (of), "yang" (that), and so on. Before the process of stop word removal and cleared tokens, however, the length of tokens that less than three characters was eliminated. Besides, in the grammatical word categories, many tokens are not a part of Indonesian words for example booking code, passenger name, date of flight, routes of flight, and so on. The result of filtering processes was 1.009 words which were suitable for the Indonesian dictionary.

Stemming is the next process after filtering to get root words from each token. The stemming method used in developing corpus was "Sastrawi", however, there was also verification which was done manually. In information retrieval, stemming has two main functions. The first function is to improve the ability of the information retrieval system to select the appropriate documents, and the second function reduces the size of the vocabulary by mapping variants based on root words [28]. Inserting data into the corpus table is the parallel process after the stemming process was done.

The final process in developing corpus was normalizing, which needed focus and involved three persons from linguistic discipline to normalize the corpus data. The normalizing processes were done by using several resources of Indonesian rules related to abbreviations and loanwords. The process collected 5.120 variants of words consists of abbreviations, loanwords, and affixes. The number of corpus data was small relatively, so the additions of Indonesian word collections were needed. The addition of words had been done by using source from Indonesian Natural Language Processing (NLP) of 29.605 words, so the total number of corpus data was 30.614 words. Overall, the data acquisition and pre-processing to develop a corpus is shown in Algorithm 1.

Algorithm 1: The algorithms of data processing

```
Result: Root Word
initialization;
while i to amount of tokens do
    check the length of each tokens;
    if length[i]==2 then
        word[i]=false;
        delete tokens[i];
    else
        word[i]=true;
    end
    if tokens[i] is not in stop-word then
        word[i]=true;
    else
        word[i]=false;
        delete tokens[i];
    end
    stem=stemmer.stem(tokens);
    if stem[i] in Sastrawi then
        word[i]=true;
    else
        word[i]=false;
        manually process word[i];
    end
    if word==true then
        insert word[i] into corpus data;
    else
    end
end
```

Results and Discussions

The main result in this research is dynamic corpus i.e. Indonesian Closed Corpus for Business (Incorbiz). On the other hand, the research will also test the result of stemming between "Sastrawi" and "Incorbiz" using the Support Vector Machine algorithm to know the level of accuracy, respectively. The sample data of "Incorbiz" are shown in Table 1.

Table 1 The Sample Data of Incorbiz [29]

word_id	root_word	business_domain	word_type	set_of_word	english
1	batal	airline;hotel;train	adjective	batal;dibatalkan;batalin	cancel
2	kamar	hotel	noun	kamar;kamarnya;sekamar	room
3	terbang	airline	verb	terbang;penerbangan	flight
4	cepat	general	adjective	cepat;cepetin;cpt	fast
5	agen	general	noun	agen;agennya;	agent
6	dewasa	airline;hotel	adjective	dewasa;dws;adult	adult
7	tiket	airline;hotel;train	noun	tiket;ticket;pertiket	ticket
8	sarapan	hotel	verb	sarapan;sarapannya	breakfast
9	lambat	general	adjective	lambat;lambreta	slow
10	stasiun	train	noun	stasiun;setasiun;st	station

The fields of "business_domain" in the corpus table indicate that "Incorbiz" could be used for multi-business, although, currently, "Incorbiz" is dedicated to airline ticket reservation business. On the other hand, "Incorbiz" was equipped with a variant of words that were stored in the field "set_of_word". The variant of words was provided to make it easier in the stemming process by using a matching string. To anticipate the slip of vocabulary, "Incorbiz" was designed to update data by using human assistance or automaticity. Human assistance was needed to process new data in the form of abbreviations or loanwords.

As an Indonesian stemmer, "Incorbiz" worked by using Structured Query Language (SQL) to find a root word. A root word was found by matching a string of variants of words. There are three possibilities in this method i.e. 1) variant of a word and root word were found, 2) variant of the word was not found but root word was found, and 3) were not found in both. On the second possibility, an auto-update will be done by "Incorbiz", however, on the third possibility human assistance is needed. This research also compared the result of stemming by using "Sastrawi" and "Incorbiz". The classification used Support Vector Machine by using the linear kernel to know the result of classification accuracy. The sample of sentences is shown in Table 2.

Table 2 The Sample of Sentences to be Classified

No.	Sentences	Label	English
1	mbak bantu booking ya yg flight ke dua	booking	Help me to book a second flight
2	bantu bookingin untuk flight ke jkt	booking	Help me to book a flight to jkt
3	bantu book tiket cgk-jog ya mbak	booking	Help me to book a ticket cgk-jog
4	tlg bantu bookingin tiket psw mas	booking	Help me to book an airline ticket
5	bisa bantu rebook tidak mbak?	booking	Could you please help to rebook?
6	mas issued aja	issued	Issue it please
7	data sdh benar. Tolong dicetak saja	issued	The data is correct. Please issue it
8	tolongin dong issued tiketnya	issued	Help me to issue a ticket
9	issued saja	issued	Please issue it
10	diisued saja dulu nanti saya kirimnya	issued	Issue it first, I will send later

The comparing result of stemming between "Sastrawi" and "Incorbiz" are shown in Table 3 and Table 4, respectively.

Table 3 The Result of Stemming by Using "Sastrawi"

No.	Sentences	Stemming Result
1	mbak bantu booking ya yg flight ke dua	'mbak', 'bantu', 'booking', 'yg', 'flight', 'dua'
2	bantu bookingin untuk flight ke jkt	'bantu', 'bookingin', 'flight', 'jkt'
3	bantu book tiket cgk-jog ya mbak	'bantu', 'book', 'tiket', 'cgkjog', 'mbak'
4	tlg bantu bookingin tiket psw mas	'tlg', 'bantu', 'bookingin', 'tiket', 'psw', 'mas'
5	bisa bantu rebook tidak mbak?	'bantu', 'rebook', 'mbak'
6	mas issued aja	'mas', 'issued', 'aja'
7	data sdh benar. Tolong dicetak saja	'data', 'sdh', 'benar', 'tolong', 'cetak'
8	tolongin dong issued tiketnya	'tolongin', 'dong', 'issued', 'tiket'
9	issued saja	'issued'
10	diisued saja dulu nanti saya kirimnya	'diisued', 'dulu', 'saya', 'kirim'

Table 4 The Result of Stemming by Using "Incorbiz"

No.	Sentences	Stemming Result
1	mbak bantu booking ya yg flight ke dua	'bantu', 'pesan', 'terbang'
2	bantu bookingin untuk flight ke jkt	'bantu','pesan','terbang','jakarta'
3	bantu book tiket cgk-jog ya mbak	'bantu','pesan','tiket','cgkjog'
4	tlg bantu bookingin tiket psw mas	'tolong','bantu','pesan','tiket','pesawat'
5	bisa bantu rebook tidak mbak?	'bisa','bantu','pesan','tidak'
6	mas issued aja	'cetak','saja'
7	data sdh benar. Tolong dicetak saja	'data','sudah','tolong','cetak'
8	tolongin dong issued tiketnya	'tolong','cetak','tiket'
9	issued saja	'cetak'
10	diisued saja dulu nanti saya kirimnya	'cetak','dulu','kirim'

Based on the result in Table 3 and Table 4, there are several differences in the result of stemming between "Sastrawi" and "Incorbiz". The differences are in the words which used suffix "in", abbreviation", and loanwords. The normalization which was applied in "Incorbiz" changed the loanwords into Indonesian for example booking changed into "pesan" and issued changed into "cetak". In an abbreviation,

the normalization was done by using human assistance because the only human could know the meaning of it. The result of stemming between "Sastrawi" and "Incorbiz" is a model that will be tested accurately by using the Support Vector Machine algorithm.

Incorbiz is a database of Indonesian words with the concept of a document database. A document database is a type of nonrelational database that is designed to store and query data as documents. The Incorbiz is working by searching the word in Incorbiz's collections to find a root word. An illustration of how Incorbiz works is shown in Figure 2.

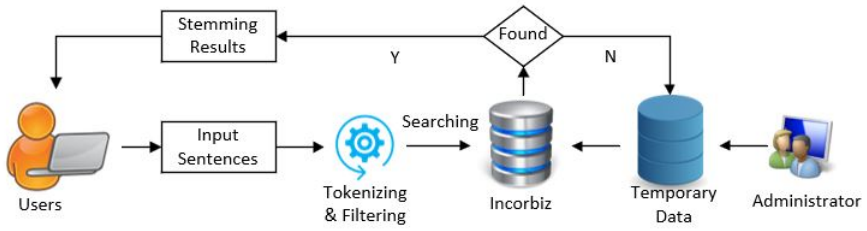

Figure 2 The Incorbiz's workflow.

The algorithms of stemming by using Incorbiz is shown in Algorithm 2.

Algorithm 2: The Incorbiz stemming algorithm

```
Result: root-word
s:sentences;
tokens:tokenizing(s);
i:0;
while tokens as word;
i=i+1;
if word[i] in stopword then
    word[i]=false;
else
    word[i]=true;
end
if word[i] then
    search in Incorbiz's documents;
    if found then
        root-word=true;
    else
        root-word=false;
        word[i] insert into temporary data
    end
else
    do none;
end
```

The stemming function in Incorbiz returns not only affixed words to the root word but also considers conformity with Indonesian rules. An example is "tolong" (help) which is abbreviated as "tlg". As a stemmer package, "Sastrawi" has not been designed to normalize abbreviations, so if encounters the word "tlg" it will not know it means "tolong". Besides, many loanwords are often used in business dialogues. The same results would be got using "Sastrawi", as it was not designed to accomplish this. These deficiencies are addressed by Incorbiz and as Incorbiz's contribution to normalizing non-formal Indonesian.

By using Python programming, this research analyzed the level of accuracy by using a classification report and confusion matrix. The number of data that were used in this research were 30 sentences using two labels i.e. book and issued. The comparison between data testing and data training was 70%:30% respectively. The kernel used in the Support Vector Machine classification was "linear" which can classify text with large features [17]. The result of accuracy level based on "Sastrawi" and "Incorbis" stemming is shown in Table 5.

Table 5 The Accuracy Level of Classifier Model by Using "Sastrawi" and "Incorbiz"

No.	Stemming Method	Accuracy Level
1	Sastrawi	0.73
2	Incorbiz	0.85

The detail of the classification report between "Sastrawi" and "Incorbis" stemming that consists of accuracy, precision, and recall are shown in Figure 2 and Figure 3, respectively.

	precision	recall	f1-score	support
0	0.65	0.73	0.69	15
1	0.80	0.89	0.84	9
2	1.00	1.00	1.00	9
3	0.46	0.46	0.46	13
4	1.00	0.95	0.97	19
5	0.71	0.71	0.71	17
6	0.81	0.94	0.87	18
7	0.67	0.44	0.53	18
8	0.29	0.33	0.31	15
9	0.80	0.80	0.80	15
10	0.94	0.88	0.91	17
accuracy			0.73	165
macro avg	0.74	0.74	0.74	165
weighted avg	0.74	0.73	0.73	165

Figure 3 The classification report based on "Sastrawi".

	precision	recall	f1-score	support
0	0.69	0.85	0.76	13
1	0.83	1.00	0.91	10
2	1.00	0.93	0.96	14
3	0.47	0.70	0.56	10
4	1.00	1.00	1.00	25
5	0.91	0.83	0.87	12
6	0.86	1.00	0.93	19
7	0.83	0.71	0.77	14
8	0.60	0.23	0.33	13
9	1.00	0.86	0.93	22
10	0.87	1.00	0.93	13
accuracy			0.85	165
macro avg	0.82	0.83	0.81	165
weighted avg	0.85	0.85	0.84	165

Figure 4 The classification report based on "Incorbiz".

The confusion matrix of the classification result based on "Sastrawi" and "Incorbis" stemming are shown in Figure 4 and Figure 5, respectively.

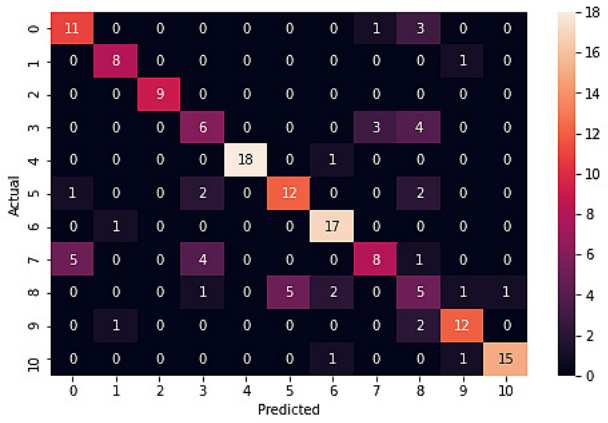

Figure 5 The confusion matrix based on "Sastrawi" .

The result in Table 5 shown the differences in the level of accuracy between stemming by using "Sastrawi" and "Incorbiz". This indicated that "Incorbiz" (0.85) was better than "Sastrawi" (0.73) in normalizing words, especially in the loanword and abbreviation. An increase in accuracy level of 0.12 obtained from difference 0.85 (Incorbiz) minus 0.73 (Sastrawi), showed that data pre-processing and normalizing had an impact on accuracy results [30]. To observe the detail of accuracy, a confusion matrix can be used to recalculate by using its components.

The components of the confusion matrix consisted of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). True Positive is actual positive that are predicted positive, while True Negative is vice versa. False Positive is actual negative that is incorrectly predicted as positive, while False Negative is vice versa [31]. The accuracy is related to the error rate of the classifier model. The accuracy is the ratio of True Positive and True Negative to the total number of accounts, while the error rate is the ratio of False Positive and False Negative to the total number of accounts. The formula of accuracy is defined as:

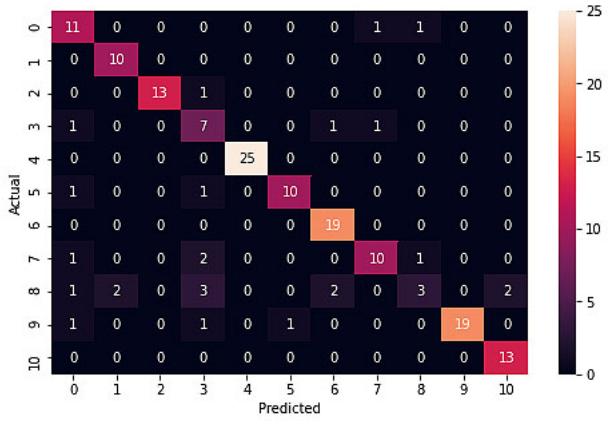

Figure 6 The confusion matrix based on "Incorbiz".

$$\text{Accuracy}=(\text{TP}+\text{TN})/(\text{TP}+\text{TN}+\text{FP}+\text{FN}) \quad (1)$$

To obtain error rate, the formula that can be used to calculate it was defines as:

$$\text{Error rate}=1-\text{Accuracy} \quad (2)$$

As mentioned, that the dataset was split into data testing and data training by using a ratio of 70%:30%. The number of rows in the dataset is 550, so the number of data testing is 550*30%=165. Based on the confusion matrix in Figure 4 that showed the result of accuracy by using "Sastrawi", indicated that it had the number of True Positive and True Negatif of 121. This indicated that it had the number of false of 165-121=44, so the error rate is 44/165 or 0.27. This is identical to the formula of error rate (2) i.e. 1- 0.73=0.27. In contrast, Figure 5 showed that stemming by using "Incorbis" increased the accuracy by indicating the number of True Positive and True Negative on 140, so the error rate is 25/165 or 0.15 or 1 - 0.85 = 0.15 (2). The small error rate showed that the level of accuracy is high to predict a class of sentences.

This research, however, still has limitations that must be resolved immediately to produce a higher quality of stemming method for Indonesian language processing. The limitations of this study includes:

1. It cannot be fully automated to correct the slip of vocabulary by updating or adding data to the corpus. The slip of vocabulary will affect in accuracy of classification because the term of frequent will be small, as a result the detected words are not the same, even though they have the same meaning.
2. The corpus which was developed as a stemming method cannot be used generally because it is being specially prepared for dialogue in airline ticket reservations.
3. Incomplete annotations for each root word. This is very important to make so that the corpus can be used for other purposes related to natural language processing.

Based on the existing limitations, further research is needed to improve the limitations. An algorithm to detect similarities in meaning, especially for "bahasa slang", must be developed immediately. If an algorithm is hard to develop, at least an automatic update technique should be developed to solve the slip of vocabulary. This

is also very important to expect the flooding of data in the big data era. Besides, it should update the searching algorithm that used in stemming process, including the environment in the database. They need a database server that has fast processing and accepting of unstructured data because the text is unstructured and high-dimensional data.

The completeness of annotations on the corpus is also urgent. It takes a fast annotation technique approach at a relatively affordable cost to complete the data, which amounts of about 30 thousand of words. Manual annotations are not recommended because their speed is slower rather than the data growth. The completed of annotations can be solved the limitations at point 2, i.e. generality to apply the corpus into multiple domain. Named Entity Recognition also needed to get a good annotation that can increase the level of accuracy.

The solving of limitation can optimize the corpus not only as stemming method but for the other benefits on languages. Emotional engineering is an example of a case that can be developed by using this corpus to classify the sentences that delivered by someone. The corpus can also analyse a hate speech by using classification to summary of sentences. Finally, the algorithms and techniques are needed to develop further "Incorbiz".

Conclusions

This research aimed to develop Indonesian dynamic closed corpus namely "Incorbiz", which can be used as an alternative stemming method to process formal and non-formal Indonesian. When it was created, the "Incorbiz's" word collections are 30.614 root words, however, the number of word variants is still limited on 5.120, because editing and normalizing is not finished yet. This research also compared two stemming technique i.e. "Sastrawi" and "Incorbiz" to process the 550-sample of dataset. The result showed that stemming by using "Incorbiz" had a level of accuracy more than "Sastrawi" on 0.85 and 0.73, respectively.

Abbreviations
Incorbiz: Indonesian Closed Corpus for Business; SVM: Support Vector Machine; NLP: Natural Language Processing; TP: True Positive; TN: True Negative; FP: False Positive; FN: False Negative.

Ethics approval and consent to participate
Not applicable

Consent for publication
Not applicable

Availability of data and materials
The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Competing interests
The authors declare that they have no competing interests.

Funding
This work was supported by the Ministry of Education and Culture Republic of Indonesian, contract number 07.1/LP/UG/III/2020 March,26 2020.

Authors' contributions
RR—writing—original draft. ABM-review & supervision, funding acquisition. EPW-review & editing. PIS-writing & editing. All authors read and approved the final manuscript.

Acknowledgements
I would like to say a special thanks to my supervisor team for their guidance, and overall insights in this field that have made this an inspiring experience for me. I would also like to thank "OkeTiket" who participated in this research. Finally, I would like to thank Universitas Teknologi Yogyakarta for supporting and facilitating this research.

Author details
[1]Faculty of Information Technology and Electrical Engineering, Universitas Teknologi Yogyakarta, Siliwangi St., 55285 Yogyakarta, Indonesia. [2]Faculty of Computer Science and Information Technology, Gunadarma University, Margonda Raya, 24105 Jakarta, Indonesia. [3]Department of Electrical Engineering and Information Technology, Universitas Gadjah Mada, Grafika 2 St., 55285 Yogyakarta, Indonesia.

References

1. Kemdikbud: Kamus Besar Bahasa Indonesia (2016). https://kbbi.kemdikbud.go.id/ Accessed 2020-03-20
2. Utami, E., Hartanto, A.D., Adi, S., Setya Putra, R.B., Raharjo, S.: Formal and Non-Formal Indonesian Word Usage Frequency in Twitter Profile Using Non-Formal Affix Rule. In: 2019 1st International Conference on Cybernetics and Intelligent System (ICORIS), vol. 1, pp. 173–176 (2019). doi:10.1109/ICORIS.2019.8874908
3. Putra, R.B.S., Utami, E.: Non-formal affixed word stemming in Indonesian language. In: 2018 International Conference on Information and Communications Technology (ICOIACT), pp. 531–536. IEEE, Yogyakarta (2018). doi:10.1109/ICOIACT.2018.8350735
4. Hidayatullah, A.F.: Language tweet characteristics of Indonesian citizens. In: 2015 International Conference on Science and Technology (TICST) (2015). doi:10.1109/TICST.2015.7369393
5. Setya Putra, R.B., Utami, E., Raharjo, S.: Accuracy Measurement on Indonesian Non-formal Affixed Word Stemming With Levenhstein. In: 2019 International Conference on Information and Communications Technology (ICOIACT), pp. 486–490 (2019). doi:10.1109/ICOIACT46704.2019.8938423
6. Waridah, E.: Pedoman Umum Ejaan Bahasa Indonesia. 5. RuangKata, Bandung, Indonesia (2019)
7. Nugraheni, E.: Indonesian twitter data pre-processing for the emotion recognition. In: 2019 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), pp. 58–63 (2019)
8. Patel, H., Patel, B.: Stemmatizer—Stemmer-based Lemmatizer for Gujarati Text. In: Rathore, V.S., Worring, M., Mishra, D.K., Joshi, A., Maheshwari, S. (eds.) Emerging Trends in Expert Applications and Security vol. 841, pp. 667–674. Springer, Singapore (2019). Series Title: Advances in Intelligent Systems and Computing
9. Kong, X., Yang, J.: Indonesian Corpus Constructing and Text Processing for Speech Synthesis. In: 2018 International Conference on Asian Language Processing (IALP), pp. 193–196. IEEE, Bandung, Indonesia (2018). doi:10.1109/IALP.2018.8629122
10. Yuwana, R.S., Suryawati, E., Pardede, H.F.: On Empirical Evaluation of Deep Architectures for Indonesian POS Tagging Problem. In: 2018 International Conference on Computer, Control, Informatics and Its Applications (IC3INA), pp. 204–208 (2018). doi:10.1109/IC3INA.2018.8629531
11. Yuwana, R.S., Yuliani, A.R., Pardede, H.F.: On part of speech tagger for Indonesian language. In: 2017 2nd International Conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE), pp. 369–372 (2017). doi:10.1109/ICITISEE.2017.8285530
12. Kwary, D.A.: A corpus platform of Indonesian academic language. SoftwareX **9**, 102–106 (2019). doi:10.1016/j.softx.2019.01.011
13. Jiwanggi, M.A., Adriani, M.: Topic Summarization of Microblog Document in Bahasa Indonesia using the Phrase Reinforcement Algorithm. Procedia Computer Science **81**, 229–236 (2016). doi:10.1016/j.procs.2016.04.054
14. Gunawan, A.A.S., Mulyono, P.R., Budiharto, W.: Indonesian Question Answering System for Solving Arithmetic Word Problems on Intelligent Humanoid Robot. Procedia Computer Science **135**, 719–726 (2018). doi:10.1016/j.procs.2018.08.213
15. Ario Utomo, M.R., Sibaroni, Y.: Text classification of british english and american english using support vector machine. In: 2019 7th International Conference on Information and Communication Technology (ICoICT), pp. 1–6 (2019)
16. Ouyang, J.: Research on english text information filtering algorithm based on svm. In: 2020 IEEE International Conference on Power, Intelligent Computing and Systems (ICPICS), pp. 1001–1004 (2020)
17. Aggarwal, C.C.: Data Classification Algorithms and Application. CRC Press, ??? (2015)
18. Sebastian, D., Nugraha, K.A.: Text Normalization for Indonesian Abbreviated Word Using Crowdsourcing Method. In: 2019 International Conference on Information and Communications Technology (ICOIACT), pp. 529–532. IEEE, Yogyakarta, Indonesia (2019). doi:10.1109/ICOIACT46704.2019.8938463
19. Gunawan, D., Saniyah, Z., Hizriadi, A.: 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 (2019). doi:10.1016/j.procs.2019.11.155
20. Agarwal, A., Gupta, B., Bhatt, G., Mittal, A.: Construction of a semi-automated model for faq retrieval via short message service. In: Proceedings of the 7th Forum for Information Retrieval Evaluation. FIRE '15, pp. 35–38. Association for Computing Machinery, New York, NY, USA (2015). doi:10.1145/2838706.2838717. https://doi.org/10.1145/2838706.2838717
21. Agarwal, A., Toshniwal, D.: Face off: Travel habits, road conditions and traffic city characteristics bared using twitter. IEEE Access **7**, 66536–66552 (2019)
22. Agarwal, A., Toshniwal, D.: Identifying leadership characteristics from social media data during natural hazards using personality traits. Scientific Reports, ??? (2020)
23. Lin, N., Fu, S., Jiang, S., Chen, C., Xiao, L., Zhu, G.: Learning Indonesian Frequently Used Vocabulary from Large-Scale News. In: 2018 International Conference on Asian Language Processing (IALP), pp. 234–239 (2018). doi:10.1109/IALP.2018.8629227
24. Nugraheni, E.: Indonesian Twitter Data Pre-processing for the Emotion Recognition. In: 2019 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), pp. 58–63 (2019). doi:10.1109/ISRITI48646.2019.9034653
25. Hasanah, U., Astuti, T., Wahyudi, R., Rifai, Z., Pambudi, R.A.: An Experimental Study of Text Preprocessing Techniques for Automatic Short Answer Grading in Indonesian. In: 2018 3rd International Conference on Information Technology, Information System and Electrical Engineering (ICITISEE), pp. 230–234 (2018). doi:10.1109/ICITISEE.2018.8720957

26. Rahman, T., Agustin, F.E.M., Rozy, N.F.: Normalization of Unstructured Indonesian Tweet Text For Presidential Candidates Sentiment Analysis. In: 2019 7th International Conference on Cyber and IT Service Management (CITSM), pp. 1–6. IEEE, Jakarta, Indonesia (2019). doi:10.1109/CITSM47753.2019.8965324. https://ieeexplore.ieee.org/document/8965324/ Accessed 2020-05-23
27. Neforawati, I., Pratama, M.O., Satyawan, W.: Indonesian Lyrics Classification using Feature Level Fusion. In: 2019 2nd International Conference of Computer and Informatics Engineering (IC2IE), pp. 6–11 (2019). doi:10.1109/IC2IE47452.2019.8940826
28. Singh, J., Gupta, V.: A novel unsupervised corpus-based stemming technique using lexicon and corpus statistics. Knowledge-Based Systems **180**, 147–162 (2019). doi:10.1016/j.knosys.2019.05.025
29. Rianto, Mutiara, A.B., Wibowo, E.P., Santosa, P.I.: Improving stemming techniques for non-formal indonesian sentences using incorbiz. ICIC Express Letter (On Press) (2021)
30. Obaid, H.S., Dheyab, S.A., Sabry, S.S.: The Impact of Data Pre-Processing Techniques and Dimensionality Reduction on the Accuracy of Machine Learning. In: 2019 9th Annual Information Technology, Electromechanical Engineering and Microelectronics Conference (IEMECON), pp. 279–283 (2019). doi:10.1109/IEMECONX.2019.8877011
31. Zeng, G.: On the confusion matrix in credit scoring and its analytical properties. Communications in Statistics - Theory and Methods **49**(9), 2080–2093 (2020). doi:10.1080/03610926.2019.1568485

Figures

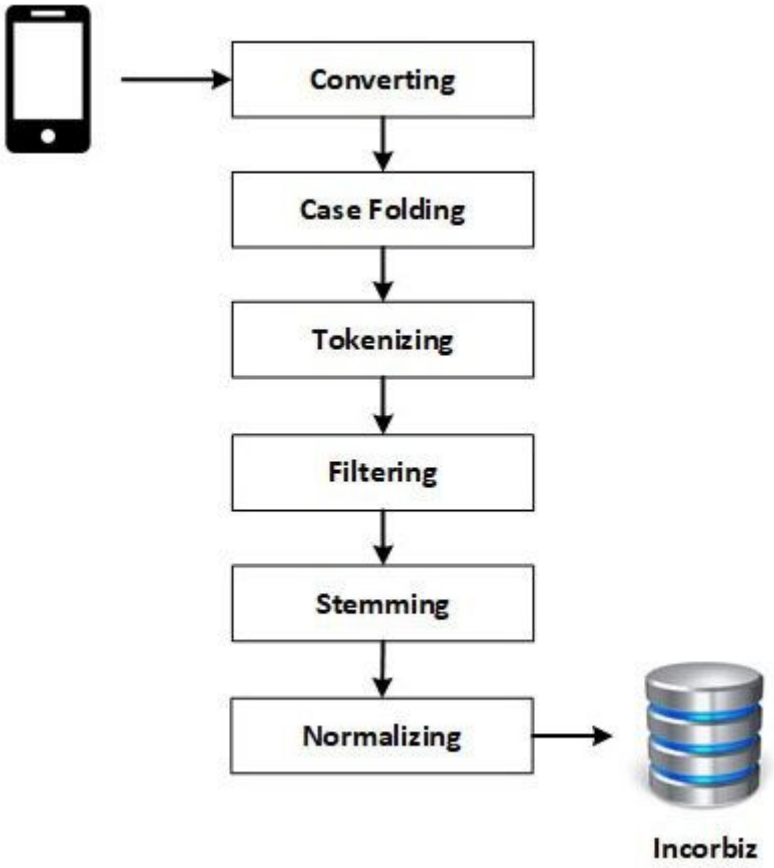

Figure 1

Step by step data preprocessing.

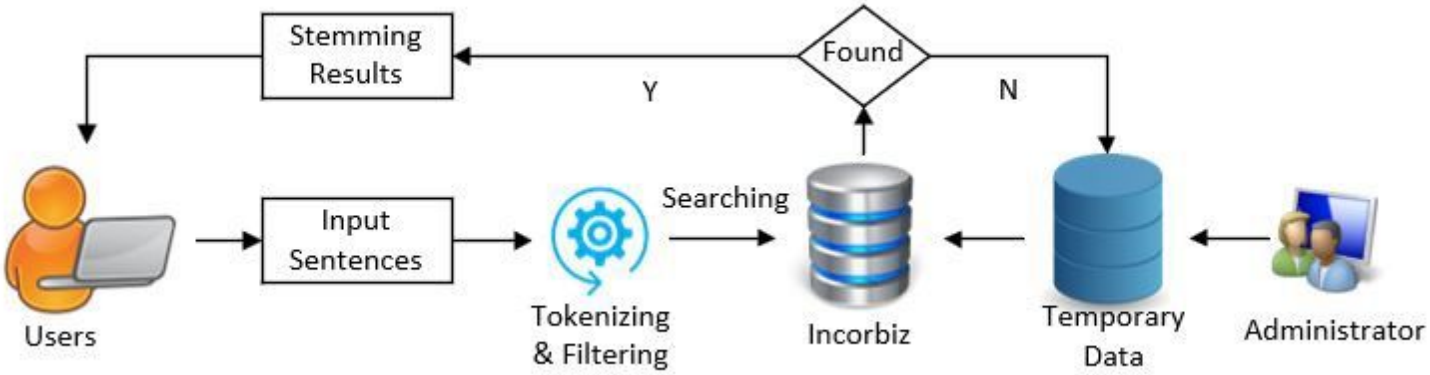

Figure 2

The Incorbiz's work flow.

	precision	recall	f1-score	support
0	0.65	0.73	0.69	15
1	0.80	0.89	0.84	9
2	1.00	1.00	1.00	9
3	0.46	0.46	0.46	13
4	1.00	0.95	0.97	19
5	0.71	0.71	0.71	17
6	0.81	0.94	0.87	18
7	0.67	0.44	0.53	18
8	0.29	0.33	0.31	15
9	0.80	0.80	0.80	15
10	0.94	0.88	0.91	17
accuracy			0.73	165
macro avg	0.74	0.74	0.74	165
weighted avg	0.74	0.73	0.73	165

Figure 3

The classication report based on "Sastrawi".

	precision	recall	f1-score	support
0	0.69	0.85	0.76	13
1	0.83	1.00	0.91	10
2	1.00	0.93	0.96	14
3	0.47	0.70	0.56	10
4	1.00	1.00	1.00	25
5	0.91	0.83	0.87	12
6	0.86	1.00	0.93	19
7	0.83	0.71	0.77	14
8	0.60	0.23	0.33	13
9	1.00	0.86	0.93	22
10	0.87	1.00	0.93	13
accuracy			0.85	165
macro avg	0.82	0.83	0.81	165
weighted avg	0.85	0.85	0.84	165

Figure 4

The classication report based on "Incorbiz".

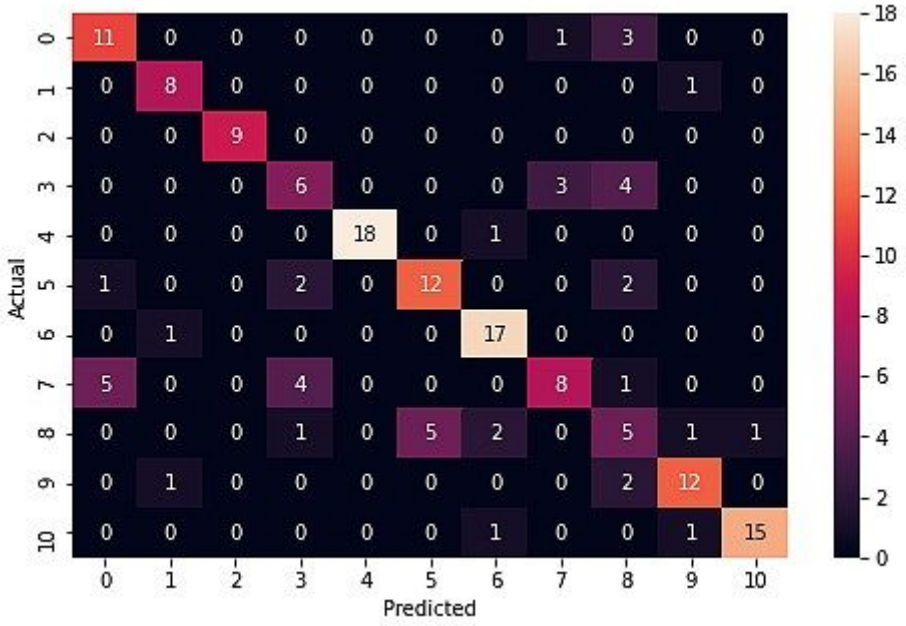

Figure 5

The confusion matrix based on "Sastrawi" .

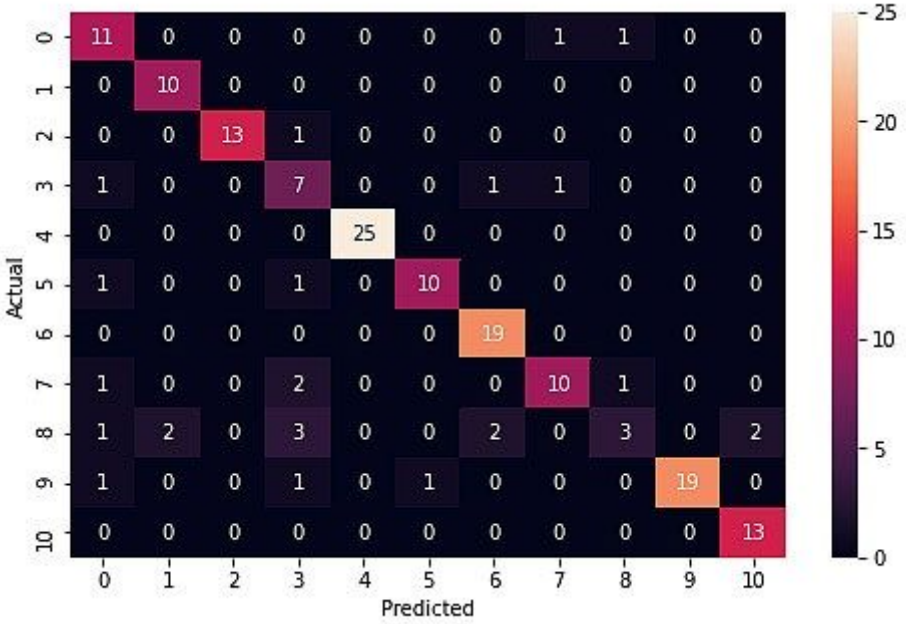

Figure 6

The confusion matrix based on "Incorbiz".