Pre-processing Tasks in Indonesian Twitter Messages

A F Hidayatullah[1] and M R Ma'arif[2]

[1]Department of Informatics, Universitas Islam Indonesia
Jl. Kaliurang KM 14,5 Sleman, Yogyakarta, Indonesia
[2]Department of Informatics Management, STMIK Jend. Achmad Yani Yogyakarta
Jl. Ringroad Barat, Banyuraden, Gamping, Yogyakarta, Indonesia

E-mail: [1]fathan@uii.ac.id, [2]rifqi@stmikayani.ac.id

Abstract. Twitter text messages are very noisy. Moreover, tweet data are unstructured and complicated enough. The focus of this work is to investigate pre-processing technique for Twitter messages in Bahasa Indonesia. The main goal of this experiment is to clean the tweet data for further analysis. Thus, the objectives of this pre-processing task is simply removing all meaningless character and left valuable words. In this research, we divide our proposed pre-processing experiments into two parts. The first part is common pre-processing task. The second part is a specific pre-processing task for tweet data. From the experimental result we can conclude that by employing a specific pre-processing task related to tweet data characteristic we obtained more valuable result. The result obtained is better in terms of less meaningful word occurrence which is not significant in number comparing to the result obtained by just running common pre-processing tasks.

1. Introduction

Twitter is a microblogging system that allows its users to post about their activities through short message called a tweet. Many tweets are posted by people around the globe. Thus, the amount of tweets is increasing steadily. From those data, people can get some useful information by classifying tweets into some categories. For example, we can classify tweets based on sentiments, contents, opinions, news, topics, etc.

However, like other social media text message, Twitter text messages are very noisy. Moreover, tweet data are unstructured and complicated enough. These are happened because there is no regulation for users to write tweet. So, they can post their tweet by ignoring grammar, spelling, etc. Therefore, Twitter messages usually contain misspelled words, abbreviations, and other bad language forms. In addition, tweets also contain symbols, link, emoticons, etc. To extract information from tweets, we need to transform those inappropriate words into standard form. According to those problems, pre-processing tasks are very important and critical in text mining [1]. Pre-processing aims to eliminate noises from the text data.

This paper addresses the issue of text pre-processing steps in Bahasa Indonesia, particularly in Twitter text messages. In addition, we will conduct two different tasks in pre-processing tasks. The first section is a common pre-processing task and the second one is a particular task for Twitter messages.

The rest of this paper is organized as follows. Section 2 presents some related works. Section 3 describes the experiment about tweet pre-processing. Section 4 explains the result and discussion of this research. Finally, section 5 presents the conclusion and future work of this research.

2. Related Works

Some previous researchers have been discussed about text pre-processing. Hemalatha et al. [2] has conducted the research about the pre-processing techniques to perform sentiment analysis in Twitter. The pre-processing steps proposed are tokenization, stemming, removing URLs, filtering, removing special characters and removing retweets. Sun et al. [3] proposed the pre-processing text on online financial text corpora. Six processing steps conducted in that research are URL and number removal; abbreviation extending; additional punctuation and lengthening words extraction and replacement before tokenization; negation identification and handling; POS tagging and removal of pronouns, prepositions and conjunctions, and punctuations; and lemmatization.

Duwairi and El-Orfali [4] presented several pre-processing and feature representation strategies in Arabic text. Some pre-processing tasks such as stemming, feature correlation, and n-gram models were experimented to investigate the effects of the accuracy in Arabic sentiment analysis. Rushdi-Saleh et al. [5] applied different pre-processing tasks on movie reviews collected from different web pages and blogs in Arabic. They proposed several pre-processing techniques including stop word elimination, stemming and n-grams generation for unigrams, bigrams and trigram.

In case of text classification, stop words removal step in pre-processing influence essentially toward the accuracy and classification performance [6] [7]. Torunoglu, et al. [1] also investigate the impact of stop words in pre-processing methods in text classification on Turkish texts. According to their experiments, the research concluded stop words only give little impact. Moreover, stemming has not significantly affect the Turkish text classification.

3. Pre-processing Experiments

In this part we divide our proposed pre-processing experiments into two parts. The first part is common pre-processing task which widely used for typical text mining job, and the second part is a specific pre-processing task for tweet data corresponding to the tweet data characteristics.

3.1. Common Pre-processing Task

3.1.1. Removing symbols, numbers, ASCII strings, and punctuations. Twitter messages usually contain symbols, numbers, and punctuations. All of these will be removed using regular expression syntax.

3.1.2. Tokenization. Tokenization task aims to divide sentence into some parts called token. The token can be formed in words, phrases or the other meaningful elements. This task is performed by using word_tokenize function from nltk.tokenize library.

3.1.3. Case folding. Case folding is the process to convert words into the same form, for instance lowercase or uppercase. In this step, we transform all words into lower case using Python string lower method.

3.1.4. Stemming. Stemming is the process to obtain the base or root of word by omitting affixes and suffixes. This research utilizes Sastrawi Python library to alleviate inflected words in Bahasa Indonesia to their base form. The algorithm for stemming in Sastrawi library is based on Nazief-Adriani algorithm.

3.1.5. Stopword removal. Stopword removal eliminates the common and frequent words which do not have the significant influence in the sentence. In this pre-processing task, we remove the stop word in the Twitter message according our stop word lists containing stop words in Bahasa Indonesia such as ‘dan’ (and), ‘atau’ (or), etc. This step is conducted by importing our stop words list from nltk.corpus library.

3.2. Specific Pre-processing Task

In typical short message text like Twitter, the messages have a wide range of quality ranging from high quality well defined text to meaningless strings. The variations raised due to typos, ad hoc abbreviations, phonetic substitutions, ungrammatical structures and emoticons. Hence the pre-processing task for tweet data cannot simply relies on common pre-processing task describes on previous sub section. On this following sub section, we have defined four characteristic of Twitter messages and our method to handle those kinds of texts.

3.2.1. Special Symbols on Twitter. Twitter has special symbols in its message such as hashtag (#), username (@username), and retweet (RT). These characters will be removed in this task. However, our method will only remove the hashtag symbol and leaving the word because the hashtag symbols usually followed by word or phrase which represent the discussed topic.

3.2.2. Emoticons Handling. Emoticons will be converted into their represented word. This work grouped seven emoticon types such as smile *(emot-senyum)*, laugh *(emot-tawa)*, love *(emot-cinta)*, sad *(emot-sedih)*, wink *(emot-kedip)*, cry *(emot-tangis)*, and stick out tongue *(emot-ejek)*. This agglomeration created based on the emoticons that found in our dataset. The snippet code of emoticon conversion can be seen in Figure 1.

```
('emot-senyum', [':-)', ':)', '(:', '(-:', '^^', '^-^', '^_^'] )    ,\
('emot-tawa',   [':-D', ':D', 'X-D', 'XD', 'xD', ] )    ,\
('emot-cinta',  ['<3', ':\*', ] )   ,\
('emot-kedip',  [';-)', ';)', ';-D', ';D', '(;', '(-;', ] ) ,\
('emot-sedih',  [':-(', ':(', '(:', '(-:', ] )  ,\
('emot-tangis', [':,(', ':\'(', ':"(', ':(('] ) ,\
('emot-ejek',   [':p', ':P', ':-p', ':-P', '=p', '=P'] )    ,\
```

Figure 1. Snippet code of emoticon conversion

3.2.3. Removing URLs. Twitter messages usually contain URLs, for example http://www.ift.tt/1QBmUt3. The URLs are removed because we only focused on the words containing in the tweets.

3.2.4. Tweet Characteristics of Indonesian People. The occurrence of non-standard words in social media text including in Twitter messages is very high. Abbreviations or shorthand and miss spellings are the most common examples of non-standard words [8]. Moreover, there are some other non-standard words that usually found such as combining letter and number, lengthening words and writing message using slang words.

Based on previous research by Hidayatullah [9], there are some unique tweet characteristics which commonly posted by Indonesian which related to the appearance of non-standard word. The first characteristic is the Indonesian sometimes show their expression by lengthening words. They repeat the letter 'e' in the word *'hore'* which means hurray in English. Indonesian people sometimes write *'horeeeee'* to show happiness in their tweet. Because of this, the lengthening words should be normalized by omitting the excess letter based on the dictionary.

The second characteristic related to a non-standard word is abbreviation or shorthand. For instance, Indonesian usually write *'g', 'gk',* or *'tdk'* to express the word *'tidak'* which means no or not. In addition, Indonesian people also often to use slang words to express the word *'tidak'* by writing *'nggak'* or *'gak'*. In addition, people also usually combining between letters and number such as *'hati2'* (be careful) that should be *'hati-hati'* to repeat the word, *'se7'* which means agree.

We presented to normalize those non-standard word into appropriate word based on the dictionary in this pre-processing step. For the lengthening word problem, we use dictionary of Indonesian standard word which contain nearly 35.000 words to obtain the standard form of those text. For the abbreviation or shorthand, we build a corpus of commonly non-standard word used by Indonesian and the corresponding standard word for replacement when such word found within incoming tweets.

3.3. Pre-processing Steps

The two previous sub sections defined the common approach of text pre-processing and specific approach to twitter text processing consecutively. Each approach has its own sub task. Figure 2 depicts the overall steps and workflow regarding the tweet pre-processing which combines the sub tasks from both approaches. In figure 2, yellow box indicating common pre-processing step, while blue box indicating specific pre-processing step.

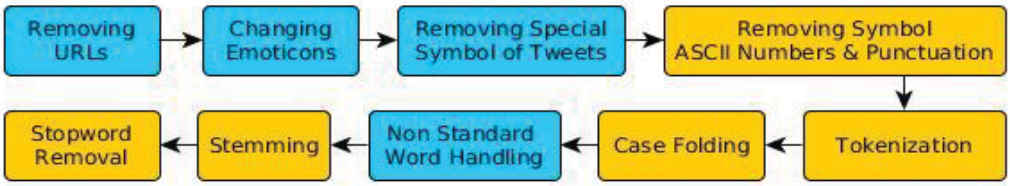

Figure 2. Pre-processing step

4. Result and Discussion

About 25,189 tweets have been obtained from Twitter using Twitter Search API v1.1. We have collected Twitter dataset by adding between our current dataset and our previous research [9] [10]. The main goal of this experiment is to clean the tweet data for further analysis. Thus, the objectives of this pre-processing task is simply removing all meaningless character and left valuable words.

Following the steps defined on sub section 3.3, we obtained the following result depicted as a word cloud in figure 3. In figure 3, we can figure out that most of words appeared on the visualization is a key term which build up the idea of the tweet. Even though there are left some of the meaningless word likes *"ya", "dr", "aja", "at",* etc., the number of such words are not significant. In word cloud, the bigger the size of the word, they appeared more frequently on the dataset.

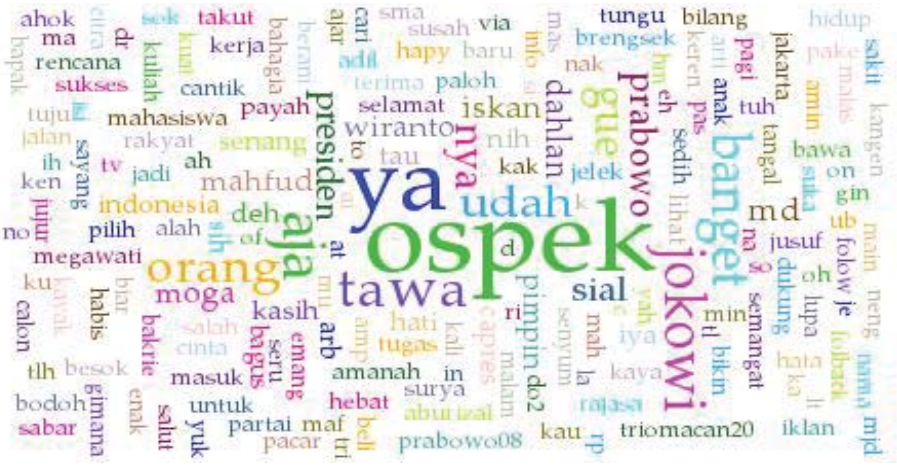

Figure 3. The result obtained by common and specific pre-processing

For a comparison purposes, we have visualized the result obtained by just employing the common pre-processing step in figure 4. In figure 4, there are exist more meaningless word in significant number. The words like "*ga*", and "*yg*" which is actually abbreviated stop word is still appeared on the dataset. Another meaningless term such as "*p*", "*d*", "*rt*", "*org*", "*udh*", and so on still also appeared in significant number.

Figure 4. The result obtained by just common pre-processing task

5. Conclusion and Future Work

In this paper, we have presented pre-processing tasks for Indonesian Twitter messages. In general, we can divide the pre-processing task into two categories. Firstly, there are common pre-processing tasks in text mining such as tokenization, case folding, removing punctuations, removing stopwords, and stemming. The second category are the specific task to handle special characteristics that frequently occurs in Twitter and social media text messages, for example removing Twitter symbols (e.g. hashtag, username, etc.), non-standard words normalization, and emoticon handling. From the experimental result, we can conclude that by employing a specific pre-processing task related to tweet data characteristic we obtained more valuable result. The result obtained is better in terms of less meaningful word occurrence which is not significant in number comparing to the result obtained by just running common pre-processing tasks.

For the future work, we can develop new algorithms to automatically recognize non-standard words in Bahasa Indonesia from social media text. In addition, we also can categorize those non-standard words into two classes such as meaningful words and meaningless words.

6. References

[1] Torunoğlu D, Çakırman E, Ganiz M C, Akyokuş S and Gürbüz M Z 2011 Analysis of Pre-processing Methods on Classification of Turkish Texts *Innovations in Intelligent Systems and Applications (INISTA), 2011 International Symposium on IEEE*

[2] Hemalatha I, Varma G P S and Govardhan A 2012 Pre-processing the informal text for efficient sentiment analysis *International Journal of Emerging Trends and Technology in Computer Science (IJETTCS)* vol 1 no 2 pp 58-61

[3] Sun F, Belatreche A, Coleman S, McGinnity T M and Li Y 2014 Pre-processing online financial text for sentiment classification: A natural language processing approach *2014 IEEE Conference on Computational Intelligence for Financial Engineering and Economics (CIFEr)* London

[4] Duwairi R and El-Orfali M 2013 A study of the effects of pre-processing strategies on sentiment analysis for Arabic text *Journal of Information Science* pp 1-14

[5] Rushdi-Saleh M, Martín-Valdivia M T, Ureña-López L A and Perea-Ortega J M 2011 OCA: Opinion corpus for Arabic *Journal of the American Society for Information Science and Technology* vol 62 no 10 pp 2045-2054

[6] Srividhya V and Anitha R 2010 Evaluating Pre-processing Techniques in Text Categorization *International Journal of Computer Science and Application* vol 47 no 11 pp 49-51

[7] Ghag K V and Shah K 2015 Comparative analysis of effect of stopwords removal on sentiment classification *International Conference on Computer, Communication and Control (IC4) 2015 IEEE* pp 1-6

[8] Baldwin T, de Marneffe M C, Han B, Kim Y B, Ritter A and Xu W 2015 Shared tasks of the 2015 workshop on noisy user-generated text: Twitter lexical normalization and named entity recognition *Proceedings of the ACL 2015 Workshop on Noisy User-generated Text*, Beijing, China

[9] Hidayatullah A F 2015 Language tweet characteristics of Indonesian citizens *International Conference on Science and Technology 2015, RMUTT*

[10] Hidayatullah A F, Ratnasari C I and Wisnugroho S 2016 Analysis of stemming influence on Indonesian Tweet sentiment analysis *TELKOMNIKA (Telecommunication Computing Electronics and Control),* vol 14 no 2