Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 116 (2017) 20–26

Procedia Computer Science

www.elsevier.com/locate/procedia

2nd International Conference on Computer Science and Computational Intelligence 2017, ICCSCI 2017, 13-14 October 2017, Bali, Indonesia

Text Normalization Algorithm on Twitter in Complaint Category

Novita Hanafiah, Alexander Kevin, Charles Sutanto, Fiona, Yulyani Arifin*, and Jaka Hartanto

Computer Science Department, School of Computer Science, Bina Nusantara University, Jl. K.H Syahdan No. 9, DKI Jakarta, 11480, Indonesia

Abstract

Many people use microblog to express complaint or criticism. However, the limitation of the length that can be written is about 160 characters and the text is in unstructured sentence. It becomes the biggest obstacle to process the information. Those unstructured sentences cause a difficulty for preprocessing in text processing tools. Therefore, normalization is needed to make the unstructured sentences can be more understandable by a machine. We proposed a normalization of Indonesian language method which adopting some ideas of normalization from other researchers and adjust to the problem of Indonesian characteristic in unstructured sentence. The experiment exploits Twitter data which use Indonesian language in complaint category. The process is divided into three stages, which are cleaning process, OOV detection and word replacement. List of Basic words and Slang dictionary are used in the OOV detection. On the other hand, Context dictionary is built to solve the ambiguity problem. The algorithm can reaches the accuracy about 90% in a complaint category.

Peer-review under responsibility of the scientific committee of the 2nd International Conference on Computer Science and Computational Intelligence 2017.

Keywords: normalization;microbolog;Indonesian language; Twitter;language processing;

1. Introduction

Technology plays important role in many aspects of life, such as in business, politic, economy, and entertainment. The information is propagated in social media, such as Twitter and Facebook. Those media are categorized as

* Corresponding author. Tel.: +62-812-9095-693.
E-mail address: yarifin@binus.edu

1877-0509
Peer-review under responsibility of the scientific committee of the 2nd International Conference on Computer Science and Computational Intelligence 2017.
10.1016/j.procs.2017.10.004

microblog, where the user can post a short sentences, pictures, as well as video link. The high popularity of microblog has been used for various purposes, for example as a media for protest, political campaigns, and comments for services or products. However, the biggest obstacle to process the information on microblog is the form of the sentences. Most of the sentences may not use the correct grammar (free style text), contains of many abbreviation, typographical errors, emoticons, and many more. Those unstructured sentences can cause a difficulty for preprocessing phase in text processing tools. Therefore, normalization is needed to make the unstructured sentences can be more understandable by a machine.

In 2001, normalization for non-standard words experimented by Richard Sproat et al. [1]. The letter sound rules in text normalizing are utilized combined with speech recognition. Taxonomy of the non-standard words is developed from four kind of distinct text. Both supervised and unsupervised methods are applied for classification and language model used to handle the disambiguation problem. Bo Han and Timothy Baldwin [2] researched in normalization of a microblog and found that most of out-of-vocabulary words (individual instances of typos, abbreviation, phonetic substitution and others that cause lexical deviation) are based on morphophonemic variations. The list of candidate canonical lexical forms generated based on the morphological and phonetic variation. Next, all candidates are ranked according to similarity measurement between the candidate word and the out-of-vocabulary (OOV) word. Twitter data is used as the data set in the experiment and the OOV word is detected utilizing a classifier which does not require any annotations. The extended research [3] is done in 2013 where some methods are experimented to measure the effectiveness of candidate selection, such as *n*-grams approach, language model based approach, noisy channel method [4], etc. The statistical model is reimplemented according to the approach of Aw et. al. [5] which has been applied for SMS normalization. They also create a method called context support (CS) to handle ambiguous words. The result shows their dictionary lookup method can achieves the best precision. Similarly, text normalization over Twitter data and SMS experimented by looking at the spelling of the words (Liu Fei, et al.) [6] which used the letter transformations approach that does not require the human supervision. Meanwhile, Zhenzhen Xue et al. [7] experiments based on the source channel model by considering four factor which are orthographic factor, phonetic factor, contextual factor, and acronym expansion.

Previous research in text normalizing of Indonesian language is included in the sentiment classification system [8]. The normalization steps consist of deletion of punctuation mark, tokenization, conversion of number to letter, reduction of repetition letter, and using corpus with Levensthein to handle the abbreviation. The same mechanism is used in a part of automatic summarization for Indonesian hashtag [9]. In other system, detection of the non-Indonesian word (OOV word) exploits dictionary lookup method and morphological analyzer, while the error correction approach measure the similarity of the OOV word and the candidates based on the value of optimum subsequences between them.

2. Scoping of Normalization

2.1. Affixes in Indonesian Language

Indonesian language has many types of affixes which each kind of affix that bound with a word can change the meaning of the word dramatically. There are four categories of affixes in Indonesian language [10], which are prefix, infix, suffix, and simulfix. In addition, there are also known as affixes absorbed from foreign languages such as *-i, -man, -wan, -wati,* and so on. The example of each category is shown in table 1. Our algorithm can scope the prefix, infix, and suffix problem, while simulfix and affixes absorbed from the foreign language cannot be detected as a correct word in the detection phase.

Table 1. Affixes in Indonesian Language

No.	Type of Affixes	Example of Affixes	Usage of Affixes in Sentence
1	Prefix	*me-, ber-, di-, ter-, pe-, per-, se-,* and *ke-*	Word: *makan* (eat) *Memakan (me-makan)* → active verb *Pemakan (pe-makan)* → the people
2	Infix	*-el-, -er-, -em-,* and *-in-*	Word: *tali* (rope)