CHAPTER 1
INTRODUCTION

1.1. Rationale

Social media is a place to find new friendships and a place to express opinions freely. One of the most popular social media today is Twitter. Twitter allows users to write and read messages commonly called tweets. Millions of tweets are written by more than 288 million active users every day[1]. Twitter is used by several political figures to campaign and increase popularity ahead of general elections.

In addition, it is common to find people using sarcasm to express their opinions. Sarcasm is a word that has the opposite meaning of what is said and it is used to mock or show resentment [2]. The problem is the difficulty is to analyze sarcasm automatically [3].

Technically sarcasm according to D. Wilson[4] is the situational disparity between the text and its context. Sarcasm requires 6-tuples consisting of speaker, listener, context, and speech, literal propositions and intended propositions[5]. Sarcasm is used as a form of negation in which the negation marker is less explicit, implying that the sarcasm is polarity-shifters[6]. Sarcasm is divided into 3 categories: sarcasm as wit, sarcasm as whimper and sarcasm as evasion. Sarcasm as wit is used with the purpose of being funny. Sarcasm as a whimper is sarcasm used to show how annoyed or angry the person is. As for sarcasm as evasion, it refers to situations when people want to avoid giving clear answers[1].

The definition in this research inspired by [2] Sarcasm is a word that has the opposite meaning of what is said that is used to mock or show resentment. The opposite meaning in this research is when positive sentiments are followed by negative sentiments.

Bouazizi and Ohtsuki [1] in their study proposed 4 feature extraction sets that are used to detect sarcasm. The first feature is the sentiment-related feature used to detect sarcasm with a positive expression in a negative context.

Furthermore, the punctuation-related feature is used to detect sarcasm with a low tone form of facial expression or intonation. Then, the syntactic and semantic features are used to detect sarcasm with ambiguous sentences in order to hide the original intent of the sentence. Finally, the pattern-relate feature is used to detect the pattern of sarcasm found in a sentence.

The method in Bouazizi and Ohtsuki's [1] research was re-implemented to detect tweets of Indonesian sarcasm. From the experiment result, 4 features extraction as in Bouazizi and Ohtsuki [1] research apparently still has some shortcomings. The first feature, which is the sentiment-related feature, is not good enough to handle sarcasm tweets that have opposite meaning, because the sentiment value can turn into positive or neutral. Second feature, Punctuation related features, Indonesian sarcasm is rarely written with quotation marks so the number of quotes is not enough to help to identify sarcasm tweets. The third one, semantic feature, the number of sarcasm written with interjection and laughing expressions is so small that they are not enough to be used to recognize sarcasm tweets. Finally lacking determining sarcasm patterns, pattern related features do not seem to fit the Indonesian sarcasm tweet implicitly expressed (satire) because sarcasm is implicitly expressed in an inconsistent pattern. Pattern related features are more suitable for sarcasm tweets that are explicitly expressed because sarcasm that is expressed in a pattern is more consistent (but the pattern is short).

Bouazizi and Ohtsuki [1] research resulted in an overall accuracy that reached 83.1% using the Random Forest for F1 scores of 81.3%. This accuracy is obtained when setting classification parameters using Number of Features: 20, Number of Trees: 100, Seeds: 20, Max. Depth: 0 (unlimited). Random Forest is an ensemble learning method for classification, regression, and other tasks. The definition used in the study is very specific, namely sarcasm is a sophisticated form of irony, the effect of the definition even though the research has a good solution but the fact is still not able to find a pattern related feature and handle sarcasm tweets that are opposite meaning, so this study proposes sarcasm detection by analyzing the text and context before and/ or after to find significant meaning from existing text data.

Based on the aforementioned previous work. This research proposes paragraph2vec and LSTM. The advantages of the paragraph2vec model can capture meaningful syntactic & semantic regularities in language and produce relationship-specific vectors. LSTM is able to recognize the required context in tweets through the input gate, forget gate, and output gate. Both combinations are very rarely used in sarcasm detection.

Therefore, the problem raised in this research is how to detect sarcasm satire and how to improve the accuracy of sarcasm detection.

1.2. Statement of the Problem and Research Question

Based on the result of previous works, the system focuses on how to detect satire sarcasm and how to improve the accuracy of sarcasm detection. Then this research is intended to find the answers of these following question :

1.Can Paragraph2vec combined with LSTM be used for sarcasm detection?

2. Will context detection produce a better sarcasm detection for Indonesian text over the bouazizi and ohtsuki ?

1.3. Objective and Hypothesis

1.3.1. Objective

The objective of this research is to detect satire sarcasm by making a system that can recognize the context. The context in this study is related words/phrases in a Twitter text. The words/phrases are represented by vector semantics namely Paragraph2vec or Doc2vec. Vector semantics learn representations of the meaning of words, directly from their distributions in texts, the meaning of word/phrases embedded in vector space. The vector used as features used to create a model for classification using LSTM.

1.3.2 Hypothesis

Sarcasm detection by context should be better than by patterns as in Bouazizi and Ohtsuki (2016). Context-based sarcasm detection method uses vector Paragraph2vec which learns representations of word meanings directly from their distributions in a sentence. Hence it is able to establish the required context by recognizing the opposite meaning containing a positive sentiment which is followed by negative sentiment. LSTM is capable of learning long term dependencies which are able to recognize the required contexts. Therefore, sarcasm detection using context could be a better approach and may improve accuracy.

1.4. Assumption

Sarcasm should be written explicitly with positive sentiment followed by negative sentiment. All information context whether a tweet is sarcasm or not should exist in a tweet. Sarcasm form oriented with satire. The author assumes the system can determine whether sarcasm or not-sarcasm with taking into consideration the information context in a tweet.

1.5. Scope and Delimitation

1.5.1 Scope

The scopes of this research are:

1. This research focuses on satire sarcasm
2. Input is twitter text with maxlen: 30-50 words with maximum 280 character in Indonesia and English

1.5.2 Delimitation

This study will only concern sarcasm in Twitter texts. The respondents in this research are people without any background on the tweets or the users who posted them.

1.6. Importance of the Study

This research shows the ability of the composition of both paragraph vectors and LSTM to the understanding of sarcasm in twitter text with the purpose of sarcasm detection. Furthermore, with the composition of both paragraph vectors and LSTM, they become the model which is built to extract detailed context satire sarcasm, written explicitly with positive sentiment followed by negative sentiment. It means that it can be used for determining sentiment in a specific issue that requires identifying sentiment patterns in a short text. Hence the composition of both paragraph vectors and LSTM have a great influence on the meaning and overall related context of a document.