Code-mixed sentiment analysis of Indonesian language and Javanese language using Lexicon based approach

C Tho¹, Y Heryadi¹, L Lukas² and A Wibowo¹

¹ Department of Computer Science, Binus Graduate Program, Bina Nusantara University, Jakarta, Indonesia
² Cognitive Engineering Research Group, Faculty of Engineering, Universitas Katolik Indonesia Atma Jaya, Jakarta, Indonesia

*cuikho@binus.edu

Abstract. Nowadays mixing one language with another language either in spoken or written communication has become a common practice for bilingual speakers in daily conversation as well as in social media. Lexicon based approach is one of the approaches in extracting the sentiment analysis. This study is aimed to compare two lexicon models which are SentiNetWord and VADER in extracting the polarity of the code-mixed sentences in Indonesian language and Javanese language. 3,963 tweets were gathered from two accounts that provide code-mixed tweets. Pre-processing such as removing duplicates, translating to English, filter special characters, transform lower case and filter stop words were conducted on the tweets. Positive and negative word score from lexicon model was then calculated using simple mathematical formula in order to classify the polarity. By comparing with the manual labelling, the result showed that SentiNetWord perform better than VADER in negative sentiments. However, both of the lexicon model did not perform well in neutral and positive sentiments. On overall performance, VADER showed better performance than SentiNetWord. This study showed that the reason for the misclassified was that most of Indonesian language and Javanese language consist of words that were considered as positive in both Lexicon model.

1. Introduction
Social media such as Twitter, Instagram, or Facebook have become popular communication channels in the past decade for many individuals to express their feelings, thoughts and ideas. There is almost no restriction on social media which makes anyone can post almost anything in social media covering all aspects and topics of life from daily life, politics, health, education, entertainment, consumer goods, services and many more. Those expressions come in many forms such as video, audio, text, or emoticon. Therefore, extracting and interpreting those expressions can give many advantages for companies, political elites, celebrities, and government to learn and predict the trends and review of their products or brand. An automated process for extracting, interpreting and categorizing the expressions into negative, positive and neutral is called sentiment analysis [1].

However, extracting and interpreting the sentiments bring many challenges. The first challenge would be identifying the language in which the sentiment is expressed in the post [2]. Such a challenge would be due to the same word in two languages having possibly different meanings. One of the examples is the word “jangan”, which means “vegetable dishes” in Javanese, but “don’t (imperative)”
in Indonesian. Another example is the word “gerah” referring to “being sick” in Javanese, but “hot or stifling” in Indonesian. The second challenge would be the variation in spelling, abbreviation and unstructured grammar [2]. Non-standard abbreviation is hard to understand, such as tdlj, standing for \textit{hati hati di jalan} (in English: take care) used by many young Indonesians. The third challenge would be limited sentiment analysis resources such as dataset or sentiment lexicon for languages other than English [3].

The multilingual society exists in social media and posting using more than one language has become common in this era. There are two styles of language combination namely: code-mixed and code-switching. Code-mixed refers to the use of more than one language in a sentence [4] (e.g. “butuh opo-opo kabari ya” which literally means “let me know if you need anything “. The terms butuh, kabari, ya are originally from Indonesian while opo-opo is in Javanese). Instead, code-switching refers to using two languages but with one language become the main or host language [5].

Three major approaches in classifying sentiment analysis are machine learning approaches that include supervised and unsupervised approach, lexicon-based approach and hybrid approach [1]. Machine learning approaches are more effective and accurate for a large dataset. Lexicon based approach will need human involvement in processing the data. Dictionary based approach and corpus-based approach are two type of approaches in lexicon-based approach. Dictionary-based approach is focused on collating the data manually and then compose it into a dictionary that have synonyms and antonym [6]. Corpus based approach focused on specific domain, the statistical and semantic method fall in this approach. The last approach is combination of machine learning approach and lexicon-based approach is called hybrid approach.

Lexicon-based approach depends heavily on the lexical resources and human involvement. Lexical resources are a set of opinionated synonym that have been categorized according to their polarity [7]. Every lexical resource has its own disadvantage and advantages. In Jurek et al. [8] stated that lexicon based approach is easy to be understood and modified by human. Five main lexical resources that have been widely used are SentiWordNet, WordNet-Affect, MPQA, and SenticNet [9]. An additional lexicon resource that consists of sensitive synset (set of synonyms) is Liu lexicon. Synset in Liu lexicon is divided into two sets: positive sensitive and negative sensitive [7]. Comparative studies on lexicon resources [10] on studies also However, all of those lexical resources are English-based, which means that for a sentiment analysis to be performed on a document written in a language other than English, a translation into English should be carried out.

Vu and Te [7] proposed a method for analysing the sentiments that consist of three steps including lexicon loading, text pre-processing and sentiment scoring. Liu lexicon and SentiWordNet are two lexicon resources that are used in that paper. Liu lexicon defines the positive and negative sensitive term, whereas SentiWordNet is used to define the score of positive and negative. Four text pre-processing steps are applied to dataset Amazon, IMDb and Yelp. Text pre-processing steps consist of text extraction, text cleaning, stemming and negation marking. Negation marking is applied by putting a suffix between negators and a clause punctuation mark. Sentiment scoring is derived from combining Liu scoring method, SentiWordNet scoring method and word count SentiWordNet scoring method. The conclusion is that by combining two lexical resources, pre-processing method and scoring method shows higher performance than only using one lexical resourcing es and scoring method. The proposed method by Vu and Te [7] shows highest performance of 84.8% compared to only LIU method 77.3%, only SentiWordNet method 80.9% and only word count 0f 49.9%.

Lexicon-based approach with splitting the tweet into several micro-phrases is adopted in [9]. Sentiment score is calculated based on sum of each score of micro-phrases. Basic, Normalized, Emphasized and Emphasized-Normalized are implementation to calculate the sentiment score of each term in micro-phrases. Two datasets are used SemEval-Test and STS-Test and combined with four lexicon resources: SentiWordNet, WordNet-Affect, MPQA, and SenticNet. The best result comes from combining SentiWordNet with emphasis – normalization for both datasets. Pamungkas [11] conducted study in sentiment analysis for Bahasa Indonesia with a simple mathematic formula that compare positive and negative score to classify the polarity of the sentiments. The dataset was taken
from review on Google Playstore and Apple Appstore and translated into English using Bing translator.

Lexicon-based approach with expanding the dictionaries in Greek was conducted in Liapakis et al. [12]. Five dictionaries were developed including Dictionary of Quality, Dictionary of Service, Dictionary of Price, Dictionary of Quantity and Dictionary of Image. Special lexicon based in finance to predict the stock movement performed well in Turner et al. [13]. Polarity such as negative, positive and neutral can be completed by adding one category such as non-sense [14]. Cross-lingual of Malay and English conducted in Zabha et al. [15], Malay lexicon was developed based on wordnet lexicon. Zabha et al. [15] proposed on new lexicon that was mixture of English and Malay that consist only positive and negative sentiment.

In Vu and Le [7], Musto et al. [9], Turneret al. [13], Wunderlich and Memmert [14], researches focused mainly on mono-lingual language, which is English, and [11] in Bahasa Indonesia. This paper is aimed to compare lexicon-model on more than one language or known as code-mixed sentiment analysis. Lexicon based approach was used in this paper and emphasized on dictionary-based approach by using two models SentiNetWord and VADER. The results of the lexicon-based approach were compared with the machine learning approach using linear regression. The arrangement of this paper will be related works in section 2, proposed method in section 3, result and discussion in section 4, and conclusion in section 5.

2. Methods
The proposed method for this experiment is presented in Figure 1. The first step consisted in gathering or scrapping the data from Twitter by using the Twitter API. The scrapping was based on twitter accounts with code-mixed tweets. A total of 3,963 tweets were gathered from @yowessory and @Kulinobareng_ from October 18, 2020 to October 22nd, 2020.

Figure 1. Proposed research method.
Pre-processing data aimed to remove word duplicates, re-tweet (RT), special characters, stop words and transform to lower case. The dataset was then translated to English by using Google Translate. This was done because the tweets were in code-mixed and there was need to compare the result from SentiWordNet, Vader with manually labelled tweets. At the phase of sentiment extraction, two lexicon resources were used. Classifying the sentiments’ polarity was based on the following scoring matrix shown in Table 1.

Table 1. Scoring matrix model 1.

| No | Requirement                      | Final Polarity |
|----|----------------------------------|----------------|
| 1  | Negativity == 0 && Positivity == 0 | Neutral        |
| 2  | Negativity == Positivity          | Neutral        |
| 3  | Negativity > Positivity           | Negative       |
| 4  | Negativity < Positivity           | Positive       |

Negativity and Positivity score depended on the model used such as SentiWordNet or VADER (Valence Aware Dictionary for sEntiment Reasoning). Sentiwordnet model will be using SentiWordNet 3.0 dictionary, while Vader will use Vader Lexicon [16]. Scoring matrix was conducted by adding one operator in RapidMiner which generated attributes with function description as follows:

```
if((Negativity==Positivity)||(Negativity==0)&&(Positivity==0),"Neutral",if(Negativity>Positivity,"Negative","Positive"))
```

The results of the scoring matrix in table 1 were saved on attribute polarity, and the results of the sentiment analysis were saved to file csv.

3. Results and discussion

Removing duplicates tweets resulted in 678 tweets that were further processed for the sentiment analysis. Sentiment model for this experiment used two lexicon models: SentiNetWord and VADER. Simple matrix polarity score was used to classify the polarity of the sentiments.

Table 2 shows result of each lexicon models compared to the manual labelling. Manual labelling was performed by native speaker of Javanese language and then translated to English by using the Google Translate. Comparing the Vader and SentiNetWord to manual labelling, both of them perform well in defining the negative tweets. However, neutral and positive don’t have the same result.

Table 2. Polarity result.

| Sentiments  | Manual Labelling | VADER | SentiNetWord |
|-------------|------------------|-------|--------------|
| Negative    | 172              | 135   | 191          |
| Neutral     | 430              | 288   | 150          |
| Positive    | 76               | 255   | 337          |

The most significant positive from VADER method obtained a score of 2.307 for positivity and 0.102 for negativity. the most significant negative is -12.564 for negativity. The most positive and negative tweets and their calculation are described in Table 3.
Table 3. Most Significant Positive and Negative from VADER model.

| Polarity | Tweets | String Score | Score |
|----------|--------|--------------|-------|
| Positive | Code-mixed: ora good-looking, ning penayang, gemati, selalu ono pas susah lan seneng. | good (0.49), compassionate (0.56), caring (0.56), hard (-0.10), happy (0.69) | Positivity: 2.307, Negativity: 0.102 |
| Negative | Code-mixed: G biasa mesoh rek. Tapi Iki kok nggilani | Curse (-0.64), disgusting (-0.62) | Positivity: 0.00, Negativity: 12.564 |

Table 4. Most Significant Positive and Negative from SentiNetWord model.

| Polarity | Tweets | String Score | Score |
|----------|--------|--------------|-------|
| Positive | Code-mixed: Nah nek ngene kan podo penake mas | Well (1.00), like (0.33) | Positivity: 14.103, Negativity: 0.808 |
| Negative | Code-mixed: wedhi rak dibales, wedhi yen dee ora ngrespon, wedhi yen gawe dee risih. Ujung-ujunge mung | afraid (-0.28), not (-0.09), afraid (-0.28), respond (0.05), afraid (-0.28), uncomfortable (-0.19), only (-0.03) | Positivity: 0.26, Negativity: 13.524 |

VADER and SentiNetWord model perform well with negative sentiments. However, SentiNetWord perform better than VADER with misclassified sentiment of 11% compared to VADER with 21.5%. Both VADER and SentiNetWord did not exhibit a strong performance with regard to neutral and positive sentiments. Reason of the bad performance is caused by many code-mixed in Javanese and Indonesian language consist of words such as good, thanks or yes expressing positives values. Those positive values give result in false positive. Our results showed that the word good has been found 278 tokens from total of 665 identified tokens. On the overall performance, VADER performed better than SentiNetWord. This study suggested that VADER gave better results in classifying compared to SentiNetWord.

4. Conclusion
This experiment showed that the polarity in code-mixed sentiment analysis in Javanese and Indonesian polarity can be classified with additional pre-processing such as translating to English. VADER and SentiNetWord prevail to classify negative sentiments in a better way compared to positive and neutral sentiments. However, classification using simple arithmetic to compare negativity and positivity scores is suggested to be modified in the next research to prevent false positive. It would be interesting...
to conduct further research by using Javanese and Indonesian lexicon resources without any necessity to translate into English.

5. References

[1] Sadia A, Khan F and Bashir F 2018 An overview of lexicon-based approach for sentiment analysis Int. Electr. Eng. Conf. 1–6
[2] Srivastava V and Singh M 2020 IIT Gandhinagar at SemEval-2020 Task 9: Code-Mixed Sentiment Classification Using Candidate Sentence Generation and Selection
[3] Kiritchenko S, Zhu X and Mohammad S M 2014 Sentiment analysis of short informal texts J. Artif. Intell. Res. 50 723–62
[4] Mishra P, Danda P and Dhakas P 2018 Code-Mixed Sentiment Analysis Using Machine Learning and Neural Network Approaches
[5] Samih Y, Maharjan S, Attia M, Kallmeyer L and Solorio T 2016 Multilingual Code-switching Identification via LSTM Recurrent Neural Networks 50–9
[6] Prakash N 2019 A Comparative study of Lexicon based and Machine learning based classifications in Sentiment analysis 8–13
[7] Vu L and Le T 2017 A lexicon-based method for Sentiment Analysis using social network data Int’l Conf. Inf. Knowl. Eng. 10–6
[8] Jurek A, Mulvenna M D and Bi Y 2015 Improved lexicon-based sentiment analysis for social media analytics Secur. Inform. 4
[9] Musto C, Semeraro G and Polignano M 2014 A comparison of lexicon-based approaches for sentiment analysis of microblog CEUR Workshop Proc. 1314 59–68
[10] Khoo C S G and Johnkhan S B 2018 Lexicon-based sentiment analysis: Comparative evaluation of six sentiment lexicons J. Inf. Sci. 44 491–511
[11] Pamungkas E W and Putri D G P 2017 An experimental study of lexicon-based sentiment analysis on Bahasa Indonesia Proc. - 2016 6th Int. Annu. Eng. Semin. Ina. 2016 28–31
[12] Liapakis A, Tsiligiridis T and Yialouris C 2020 A Sentiment Lexicon-based Analysis for Food and Beverage Industry Reviews. The Greek Language Paradigm Int. J. Nat. Lang. Comput. 9 21–42
[13] Turner Z, Labille K and Gauch S 2020 Lexicon-Based Sentiment Analysis for Stock Movement Prediction 14 185–91
[14] Wunderlich F and Memmert D 2020 Innovative approaches in sports science-Lexicon-based sentiment analysis as a tool to analyze sports-related twitter communication Appl. Sci. 10
[15] Zabha N I, Ayop Z, Anawar S, Hamid E and Abidin Z Z 2019 Developing cross-lingual sentiment analysis of Malay Twitter data using lexicon-based approach Int. J. Adv. Comput. Sci. Appl. 10 346–51
[16] Hutto C J and Gilbert E E 2014 VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14).” Proc. 8th Int. Conf. Weblogs Soc. Media, ICWSM 2014