InSet Lexicon: Evaluation of a Word List for Indonesian Sentiment Analysis in Microblogs

Conference Paper · December 2017

DOI: 10.1109/IALP.2017.8300625

2 authors, including:

Fajri Koto
University of Melbourne
13 PUBLICATIONS **42** CITATIONS

SEE PROFILE

InSet Lexicon: Evaluation of a Word List for Indonesian Sentiment Analysis in Microblogs

Fajri Koto
Engineering Department
KMK Online
Jakarta, Indonesia
Email: fajri.phd@gmail.com

Gemala Y. Rahmaningtyas
Engineering Department
KMK Online
Jakarta, Indonesia
Email: yanuarita.gemala@gmail.com

***Abstract*—In this study, we propose InSet, an Indonesian sentiment lexicon built to identify written opinion and categorize it into positive or negative opinion, which could be utilized to analyze public sentiment towards particular topic, event, or product. Composed using collection of words from Indonesian tweet, InSet was constructed by manually weighting each words and enhanced by adding stemming and synonym set. As the result, we obtained 3,609 positive words and 6,609 negative words with score ranging between –5 and +5. Based on the experiment utilizing the InSet, our method outperforms other rarely found Indonesian lexicon that we used as baseline.**

***Keywords*- sentiment analysis; lexicon; indonesian; microblog; twitter**

I. Introduction

As Indonesia is an archipelago country with thousands of islands across the ocean and diverse culture as well as high population density, online social media has become a medium for most Indonesian citizen to communicate and send any thought or ideas across the country. Even these days, any microblogging tools such as *Facebook*, *Twitter* and *Instagram* have become an inseparable culture of modern Indonesian. And in fact, it has been very easy to create a viral news on the Internet and gather Indonesian public opinions towards certain topic.

With the massive number of social media users in Indonesia, it would be interesting, if not profitable, to analyze the netizens sentiment of any topics, or which is also known as Sentiment Analysis. Sentiment Analysis is a classification task to determine the text polarity and refers to broad area of natural language processing, computational linguistic and text mining [1]. According to [2], there are two kinds of sentiment classification task: 1) polarity classification with *classes* = {*positive, negative*}, and 2) subjectivity classification with *classes* = {*subjective, objective*}. It is clear that the positive or negative class indicates the positive or negative polarity of a sentence. On the other hand, the objective sentence means the utterance of information containing facts or news and less argumentation while the subjectivity reflects a private point of view, emotion or belief [3].

Though sentiment analysis has been one of the most studied field in computer science and specifically in natural language processing, there are not many researches conducted to improve Sentiment Analysis, or opinion mining, for Indonesian language. Therefore, in this work we tried to improve sentiment analysis technique specifically for Indonesian language by building a precise Indonesian lexicon which particularly aims microblogs.

Our sentiment lexicon was composed by words gathered from *Twitter*, as the representation of commonly used social media in Indonesia. We built the lexicon by classifying the polarity of each word and enhanced it with some previously proven methods. The result of tests and evaluations conducted in this study shows that InSet has a satisfactory performance as an Indonesian sentiment lexicon to predict the negative and positive polarity of shortly written opinions.

II. Related Works

In computer science, the study of sentiment analysis over microblogs has been one of the main focus in the field of natural language processing, information retrieval and data mining. These researches mainly focused on various construction for English Lexicon such as *SentiWordNet* [6], Liu Lexicon [11], AFINN Lexicon [12], *Opinion Finder* [9], Senti-Strength [13], HBE Lexicon [14], and also NRC Emotion Lexicon [15]. Research by Koto et al. has summarized the performance comparison among these features and reveal that AFINN and Senti-Strength are the current best features for English Twitter Sentiment Analysis [16].

Research on sentiment analysis over microblog Twitter has been done by Go et al. in which they utilized emoticons to annotate tweets with sentiment label [10]. The next study by Agarwal et al. used manually annotated tweets with sentiment and performed unigram model to do classification [17]. In other studies, Koto et al. analyzed sentence pattern of tweets with sentiment label, either in subjectivity and polarity domain [18].

Despite the large number of works related to English sentiment analysis, the research number that focuses on Indonesian language are limited. In [4], Wicaksono et al. proposed a methodology to automatically construct dataset for twitter sentiment analysis, while in [5] Lunando et al. tried to combine sarcasm and sentiment analysis task for Indonesian language. However, the features used for sentiment analysis were only produced from the translated *SentiWordNet* [6].

978-1-5386-1981-0/17/$31.00

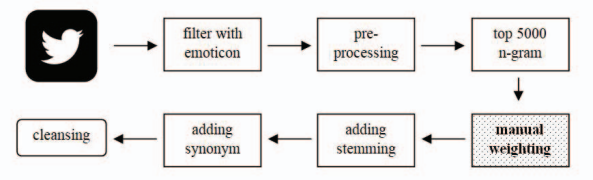

Figure 1. Stages of InSet Construction

In other work, Naradhipa et al. only used n-gram as the key feature to perform sentiment classification [7] which did not give a lot of improvement from the previous technique. The most satisfying research related to Indonesian sentiment analysis that we found was works by Vania et al. [8] in which they constructed Indonesian sentiment lexicon by translating it from *Opinion Finder* [9] and applying enhancement by seeding the words. Based on those previous works, here we focused on building lexicon for Indonesian language which is used to analyze public sentiment specifically in microblogging.

III. Indonesia Sentiment (InSet) Lexicon Construction

Our sentiment lexicon was constructed in 2017 and utilized the *Twitter* data stream around November 2016. The data was collected for three days and filtered with Bahasa Indonesia (Indonesian Language) and two kinds of emoticons that express positive ":)" and negative ":(" polarity. We grouped the tweet into positive and negative set by following works of Go et al. in which they utilized emoticons to annotate tweet with sentiment label [10].

In total we have around 10,000 tweets and then applied the preprocessing stages by 1) removing repetitive ads, 2) converting to the lowercase, 3) removing *url* and *Twitter* entities such as @*account*, and 3) removing special character and stopwords. To select the words candidate, we applied n-gram (where n = {1, 2, 3}) and removed words with frequency equals to 1. In this stage, we had 12,503 and 13,164 candidates for positive and negative words consecutively.

As AFINN [12] has shown a very good performance in sentiment analysis task for English twitter [16], we followed their work by scoring top-5000 words for each set, range from -5 (very negative) to +5 (very positive). To make the labelling process easier, we only scored for valence, leaving out subjectivity/objectivity. Manual weighting was done by two native Indonesian speakers which had been given the same instruction before conducting the scoring.

The result of scoring is shown in Table I and indicates that there are many words weighted as 0 by the annotators. The score column was calculated by averaging the scores given by two annotators and rounded into the greater nearest integer (*ceiling*). The agreement score between two annotators can be described as the average of sum of difference score. For positive set, the agreement score is 0.52 for whole words and 1.27 if we exclude the words with score 0, while for the negative set the agreement score is 0.45 and 0.97. Since the score ranges between 0 and +5 or –5, the agreement score can be classified as a good annotation result.

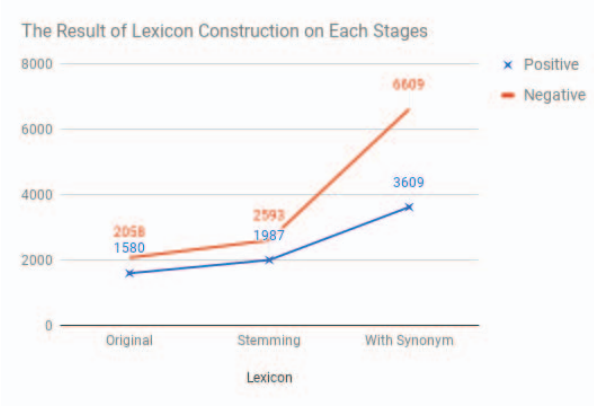

Figure 2. The Result of Lexicon Construction on Each Stages

Table I
The Manually Scoring Result of InSet Construction

Positive		Negative	
Score	Count	Score	Count
0	3421	0	2942
1	420	-1	197
2	529	-2	464
3	382	-3	719
4	205	-4	458
5	44	-5	520

As described in Fig 1, the word list construction was continued by conducting stemming for the resulting words in Table I. In this stage, we excluded words with score 0 since it can be classified as irrelevant for sentiment lexicon. Stemming was conducted by utilizing *Sastrawi*[1], library that implements stemming algorithm for Indonesian language, proposed by [19]. If the resulting words exist in the original set, then it will be excluded, otherwise the score of the stemmed words will follow the highest polarity of its original words. For instance "*memarahi*" (scold) and "*dimarahi*" (scolded) have polarity scores –4 and –2 consecutively. The stemming result of both words is "*marah*" (angry) will follow the score of "*memarahi*" word which contains higher polarity for negative set.

After enhancing through stemming stage, our lexicon grew into 1,987 positive and and 2,593 negative words as described in Fig 2. This word list was then used as the input of the next stage where we added the synonym to enhance the lexicon. Here, we used Indonesian synonym from *SinonimKata*[2] which consists of 35,711 unique words. Words which have similar synonym will be scored with the highest polarity score as well as the scoring of stemming stage.

[1]https://github.com/sastrawi/sastrawi
[2]http://www.sinonimkata.com/

We also conducted cleansing as the final stage, particularly for the result of synonym addition. Some words in Indonesian language may have different polarity with its synonym, so it may cause error if we add all of synonym sets. For instance, word "*hotel*" (hotel) is the synonym of "*gubuk*" (shack), but the polarity is totally different where "*gubuk*" is more suitable for negative word. Word "*bacot*" (words) and "*perkataan*" (words) are also the synonym to each other but "*bacot*" contains the strong negative polarity. Therefore, to exclude some irrelevant synonym addition, we conducted 2 steps: 1) we only selected words with high polarity and remove synonym words with score = {–2, –1, 1, 2}. 2) For each additional words, we excluded it from the list if it also exists in the additional list of its opposite polarity.

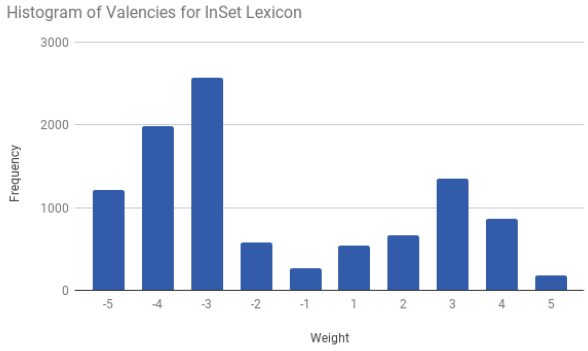

Figure 3. Histogram of Valencies for InSet Lexicon

Finally, we have the word list that comprises of 10,218 words and call them as InSet (Indonesian Sentiment) as described in Figure 3. The word list has a bias towards negative words (6,609, corresponding to 65%) compared to positive words (3,609). However, the bias apparently corresponds closely to the bias found in the *Opinion Finder* sentiment lexicon (4911 (64%) negative and 2718 positive words), and also AFINN Lexicon (1598 (65%) negative and 878 positive words).

IV. Experiment and Evaluation

A. Experimental Set-Up

To evaluate the InSet, we used a *Twitter* dataset from 2015 which has been manually annotated by a native Indonesian speaker. The data was crawled through twitter API and has 1,259 positive and 1,371 negative tweets. The InSet lexicon was used by constructing two values, called as *InSetPos* and *InSetNeg*, where *InSetPos* or *InSetNeg* is the sum of the scores for the positive or negative words that match the lexicon. After that, we performed binary classification by using several supervised algorithms such as: Logistic Regression (LR), Naive Bayes (NB), Support Vector Machine (SVM), and Neural Network (NN). The preprocessing was conducted before performing training and testing for the evaluation and it includes 1) tweet conversion to the lowercase, 3) removing *url* and twitter entities, and 3) removing special character.

Table II
List of Features used in the Experiment

Technique	#positive	#negative	#unique
TF	unigram + bigram of the texts		
TF-IDF	unigram + bigram of the texts		
Vania Lexicon [8]	414	581	994
Translated *SentiWordNet* [6]	17015	18028	29095
Translated Liu Lexicon [11]	1182	2402	3461
Translated AFINN [12]	878	1598	2476
InSet Lexicon	3609	6609	9075

As the baseline, we used some existing techniques such as n-gram, Vania Lexicon [8] and some translated well-known lexicons. Here we followed the works of Lunando et al. in which they use translated *SentiWordNet* to perform Sentiment Analysis task for Indonesian language [5]. By using Google translation, three translated lexicons were included, such as *SentiWordNet*, Liu Lexicon and AFINN Lexicon. The details of the word distribution in each polarity are described in Table 2. Vania and Liu Lexicon were used by calculating number of words that match with corresponding sentiment class, while features of *SentiWordNet* and AFINN were constructed by summing of the scores for the positive or negative words that match the lexicon.

B. Experiment Result

To evaluate our method, we performed cross validation (with k=10) and show the results in Table III. The value on each cell indicates the accuracy of each methodology with particular classifier. Our experiment results show that the traditional technique such as TF and TF-IDF do not work well in classifying the tweets for Indonesian language. The translated *SentiWordNet*, Liu and AFINN also show the similar result. It might be caused by 1) the error of the translation system, 2) the translation system which does not cover the OOV (Out of Vocabulary) or slang words of Indonesian language, and 3) the lexicon itself which contains too many uncommon words and rarely used for user-generated platform such as *Twitter*.

In Table III, the InSet has the highest accuracy for each classifier compared to the other baselines. It reaches 65.78% as the highest and better than the Vania Lexicon with 61.48% accuracy. Vania Lexicon is an Indonesian Lexicon that was built by translating *Opinion Finder* Lexicon and then seeding the words for the enhancement. Factors which may cause difference in performance are: 1) Vania Lexicon was produced by utilizing the translation system, while the word candidates of InSet were selected directly from Indonesian language source, 2) Vania Lexicon contains many formal words, while InSet contains formal and informal words as it was generated using Twitter data, 3) InSet was constructed by manually labeling of 2 native speakers where the bias and the precision of words polarity become the priority of InSet, while Vania Lexicon was constructed automatically. Even though our Lexicon is less reproducible, we argue that

Table III
THE ACCURACY (%) OF EACH FEATURE SET

Classifier	Baseline						InSet
	TF	TF-IDF	Vania Lex.	Translated English Lexicon			
				SentiWordNet	Liu Lex.	AFINN	
NB	59.52	58.71	61.48	52.66	61.48	57.18	**65.13**
LR	53.58	52.89	61.44	53.83	61.48	52.13	**65.78**
SVM	55.56	56.21	61.48	53.76	61.48	59.51	**65.51**
NN	57.64	54.68	61.41	52.51	59.84	58.97	**65.71**

the manual effort is still needed and required as it works very well like AFINN. 4) InSet has polarity score that indicates positive or negative sentiment of a word, while Vania Lexicon only a list of words in two different classes: positive and negative set.

V. CONCLUSION

In this study we have constructed InSet Lexicon that comprises of 3,609 positive and 6,609 negative words in Indonesian language. Each word was manually labeled based on its polarity and then enhanced by adding the stemming and synonym set. Our research contributions lay on two points. First, we constructed the sentiment lexicon for Indonesian language which is more suitable for microblogs. Second, our approach shows that the result outperforms all of the existing baseline methods where the highest accuracy is 65.78%. For the future works, the enhancement of lexicon can be conducted by incorporating the translated English lexicon with the InSet.

REFERENCES

[1] W. J. Trybula, "Data Mining and Knowledge Discovery". In *Annual review of information science and technology (ARIST)*, 1997, pp. 197–229.

[2] F. Bravo-Marquez, M. Mendoza, and B. Poblete, "Combining strengths, emotions and polarities for boosting Twitter sentiment analysis". In *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining*, 2013.

[3] S. Raaijmakers and W. Kraaij, "A Shallow Approach to Subjectivity Classification". In *ICWSM*, 2008.

[4] A. F. Wicaksono, C. Vania, B. Distiawan, and M. Adriani, "Automatically Building a Corpus for Sentiment Analysis on Indonesian Tweets". In *PACLIC*, 2014, pp. 185–194.

[5] E. Lunando, and A. Purwarianti, "Indonesian social media sentiment analysis with sarcasm detection". In *Advanced Computer Science and Information Systems (ICACSIS)*, 2013, pp. 195–198.

[6] S. Baccianella, A. Esuli, and F. Sebastiani, "*SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining*". In *LREC*, 2010, pp. 2200–2204.

[7] A. R. Naradhipa, and A. Purwarianti, "Sentiment classification for Indonesian message in social media". In *Cloud Computing and Social Networking (ICCCSN)*, 2012, pp. 1–5.

[8] C. Vania, M. Ibrahim, and M. Adriani, "Sentiment Lexicon Generation for an Under-Resourced Language". In *International Journal Comput. Linguistics Appl.*, 20154, pp. 59–72.

[9] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis". In *Proc. of the conference on human language technology and empirical methods in natural language processing*, 2005, pp. 347–354.

[10] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision." In *CS224N Project Report, Stanford*, 2009, pp. 1–12.

[11] B. Liu, M.and J. Cheng, "Opinion observer: analyzing and comparing opinions on the web". In *Proceedings of the 14th international conference on World Wide Web*, 2005, pp. 342–351.

[12] F. A. Nielsen, "A new ANEW: Evaluation of a word list for sentiment analysis in microblogs". 2011, Available at http://arxiv.org/abs/1103.2903

[13] M. Thelwall, K. Buckley, and G. Paltoglou, "Sentiment strength detection for the social web". In *Journal of the American Society for Information Science and Technology*, 2012, pp. 163-173.

[14] F. Koto, and M. Adriani, "HBE: Hashtag-Based Emotion Lexicons for Twitter Sentiment Analysis". In *Proceedings of the 7th Forum for Information Retrieval Evaluation*, 2015, pp. 31–34.

[15] S. M. Mohammad, P. D. Turney, "Crowdsourcing a wordemotion association lexicon". In *Computational Intelligence*, 2013, pp. 436–465.

[16] F. Koto, and M. Adriani, "A comparative study on twitter sentiment analysis: Which features are good?". In *International Conference on Applications of Natural Language to Information Systems*, 2015, pp. 453–457.

[17] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau, "Sentiment analysis of twitter data". In *Proc. of the Workshop on Languages in Social Media*, 2011, pp. 30–38.

[18] F. Koto, and M. Adriani, "The use of POS sequence for analyzing sentence pattern in Twitter sentiment analysis". In *Advanced Information Networking and Applications Workshops (WAINA)*, 2015, pp. 547–551.

[19] M. Adriani, J. Asian, B. Nazief, S. M. Tahaghoghi, and H. E. Williams "Stemming Indonesian: A confix-stripping approach". In *ACM Transactions on Asian Language Information Processing (TALIP)*, 2007, pp. 1–33.