IJCLA VOL. 5, NO. 1, JAN-JUN 2014, PP. 59–72
RECEIVED 07/01/14 ACCEPTED 06/02/14 FINAL 18/06/14

Sentiment Lexicon Generation for an Under-Resourced Language

CLARA VANIA, MOH. IBRAHIM, AND MIRNA ADRIANI

Universitas Indonesia, Indonesia

ABSTRACT

Sentiment analysis and opinion mining are actively explored nowadays. One of the most important resources for the sentiment analysis task is sentiment lexicon. This paper presents our study in building domain-specific sentiment lexicon for Indonesian language. Our main contributions are (1) methods to expand sentiment lexicon using sentiment patterns and (2) a technique to classify the polarity of a word using the sentiment score. Our method is able to generate sentiment lexicon automatically by using a small seed of sentiment words, user reviews, and part-of-speech (POS) tagger. We develop the lexicon for Indonesian language using a set of seed words translated from English sentiment lexicon and expand them using sentiment patterns found in the user reviews. Our results show that the proposed method can generate additional lexicon with sentiment accuracy of 77.7%.

KEYWORDS: Sentiment lexicon, natural language processing, under-resourced language, lexicon generation.

1 INTRODUCTION

Sentiment analysis or opinion mining is one of the most active research areas today. The rapid growth of social media such as Twitter, Facebook, forum discussions, etc., has made a huge amount of opinionated data available on the web. People share their opinion about things they like or dislike on the web. A person who wants to buy a particular product

searches for its review on the web. Organizations conduct survey or research to analyze public opinions. As a result, opinion mining has been used to track public opinions toward entities, i.e products, events, individuals, organizations, topics, etc.

One of the most important resources for sentiment analysis task is sentiment lexicon. Sentiment lexicon consists of words with its polarity, whether it is positive or negative. For example, "good" is considered as positive word and "bad" as negative word. While there are many English sentiment lexicons available on the web, sentiment lexicons in other languages can be considered very limited or even unavailable. This made research in sentiment analysis quite difficult for non-English documents. Therefore, developing sentiment lexicon in other languages is very important.

According to Liu [10], sentiment lexicon generation can be divided into three approaches, namely manual approach, dictionary-based approach, and corpus-based approach. The first approach is built manually by human and thus requires considerable resources. The second approach is dictionary-based approach, where a set of seed words is created manually and then expanded by using a dictionary (thesaurus, WordNet, etc). The corpus-based approach also uses manually labeled seed words and then expanded using available corpus data.

Many research works on sentiment lexicon generation have been done. Most of the research work is applied in English, while for other languages the research is still growing. Turney and Littman [18] use queries to find candidate English sentiment lexicons from Web search engine. Kanayama and Natsukawa [7] propose an unsupervised method to detect polar clause in domain-specific documents. Qiu et al. [16] use double propagation to expand the sentiment lexicon and extract opinion target in a document. Pérez-Rosas et al. [14] apply dictionary-based approach to build Spanish sentiment lexicon. Kaji and Kitsuregawa [5] uses massive HTML corpus to build Japanese sentiment lexicon. In their work, they use structural clues to find polar sentence from Japanese HTML documents. Banea et al. [1] propose a method for constructing sentiment lexicons for low-resourced language.

In this paper, we apply corpus-based approach to build Indonesian sentiment lexicon for a specific target domain. While most of sentiment lexicon generation techniques rely on the availability of WordNet, in our case it is not feasible because of the limitation of Indonesian language resources. Our proposed methods depend on the availability of English sentiment lexicon, machine translation, part-of-speech (POS) tagger and online user reviews. Our main contributions in this paper are:

1. Methods to expand the sentiment lexicon using automatic translation services and simple pattern-based approaches. We use available English sentiment lexicon and translate them into Indonesian language. To expand the lexicon, we use user reviews from user-generated content (UGC) and social media data, as they are available and can be collected easily.
2. Techniques to filter sentiment words and scoring function to determine the polarity of each word.

In this work, we show that although the language resources are limited, we can use other resources, which can be collected easily to build the lexicon. UGC and social media are quite popular nowadays and available in almost every language. Those data also contains many public opinions and very suitable for sentiment analysis research.

2 INDONESIAN SENTIMENT LEXICON GENERATION

2.1 *Seed Lexicon*

Many research about sentiment lexicon generation use seed words to build the lexicon. Some use manually built seed lexicon [9] and some others use seed words taken from dictionary (e.g., [2, 4, 6, 8]). In this study, we use an available English sentiment lexicon, which has been widely used in many sentiment analysis research works. The lexicon that we used in this experiment is OpinionFinder[1] [21] and SentiWordNet.[2] In the OpinionFinder, each word is assigned with its polarity; positive, negative, or objective. It also gives label strong or weak subjectivity to each word. SentiWordNet is another English sentiment lexicon developed by [4]. This lexicon is built in accordance with WordNet. Each synset is assigned with its subjectivity score. SentiWordNet defines three score for each synsets; positive, negative, and objective score.

In this study, we aim to build sentiment lexicon with positive and negative subjectivity. We begin by selecting initial seed words to building the lexicon. We select terms from OpinionFinder with *strong positive* / ***negative polarity***. For SentiWordNet, we select adjective synsets with *highest subjectivity score* (in this experiment we take terms with score

[1] http://mpqa.cs.pitt.edu/opinionfinder
[2] http://sentiwordnet.isti.cnr.it

above 0.7). These selection criteria help us to choose terms with strong polarity and use it as seed words.

As we want to build lexicon for Indonesian language, we translate those seed words into Indonesian. Several problems which occur are terms which do not have its corresponding terms in Indonesian and some English terms which have the same translation in Indonesian. The same problem also happened in the previous research by Wiebe et al. [21], Wan [19] in using translation to build sentiment lexicon. In this study, we simply eliminate terms that do not have its corresponding translation in Indonesian language.

In order to get expansion terms with high precision, we have to ensure that the selected seed words are opinion words. Therefore, we conduct two stages of manual evaluation which consist of translation and subjectivity evaluation. For translation evaluation, we eliminate words that have no translation in Indonesian. Duplicate translations and mistranslated word also removed from the seeds. To evaluate the subjectivity, we conduct manual evaluation for each word to check whether the translated word contains the same polarity with the English word.

Table 1 shows the statistics of our seed lexicon. After evaluation, there are 291 positive words and 517 negative words which we used as seed words.

Table 1. Statistics of Seed Words

Source Lexicon	#English words	#Translated Words	#Seed Words
Positive Words	2071	1161	**291**
Negative Words	4637	2392	**517**

2.2 *Sentiment Lexicon Expansion*

SENTI-PATTERN (SP). People tend to have similar patterns to express their opinion. For domain-specific sentiment analysis, these patterns are useful to analyze opinions about a particular entity. For example, in book reviews, we can find opinions such as "*This book is great*" or "*This book is awfull*". Although those opinions have opposite subjectivity, the sentences use the same pattern that clearly states opinions about the book.

In the first approach, we want to find sentiment patterns that are usually found in the user reviews. In the previous study, Pantel and Pennacchiotti [12] use generic patterns to extract semantic relations from raw text. In this study, our hypothesis is that sentiment patterns that are frequently used by many reviewers can be used to extract new sentiment

words. Fig. 1 shows the extraction process of SP. In the first step, documents (user reviews) will be divided into sentences. After that, we develop a list of n-grams (n = 3) along with its frequency that are found in the corpus. We filter the n-grams by only taking n-grams which contains seed words. Any seed word found in the n-grams is then replaced by the same tag, i.e. [SENT] to indicate sentiment word position in the n-gram. Top-N n-grams with highest frequency (we use N=50) are then used as senti-patterns. Fig. 2 shows example of sentiment patterns found in the corpus.

We expand the seed words by searching the senti-patterns in the corpus to find candidate sentiment words. At this step, we do not classify the word polarity as the patterns can contains opinion words with various polarities. The polarity classification will be done at the filtering step.

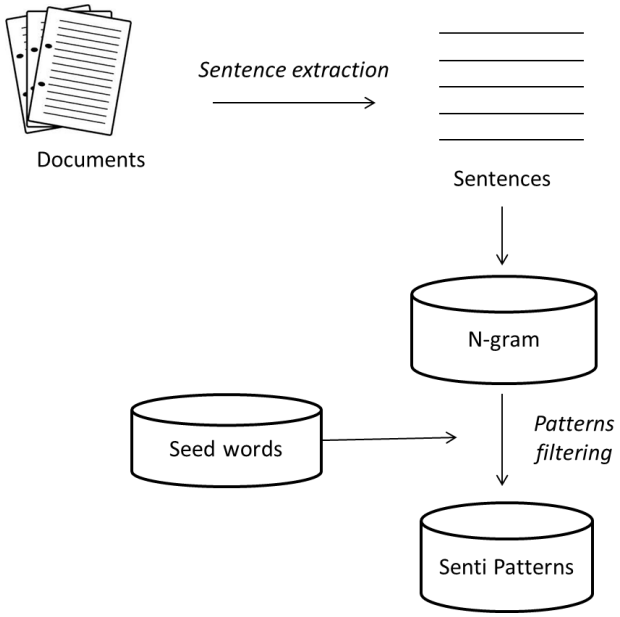

Fig. 1. Senti-Pattern (SP) extraction process

tempat yang [SENT]
('[SENT] place')

Fig. 2. Example of Senti-Pattern

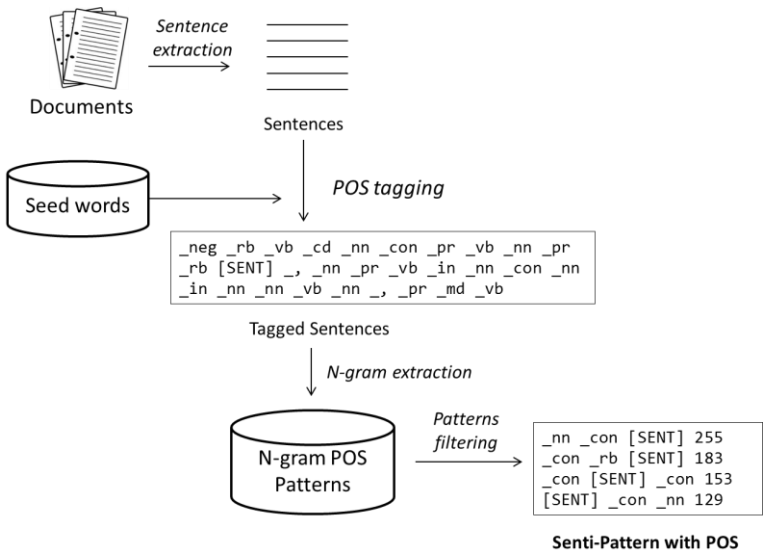

Fig. 3. Senti-Pattern with POS (SP-POS) extraction process

SENTI-PATTERN WITH PART-OF-SPEECH (SP-POS) The SP approach find new candidate words by using patterns that usually occurred in the documents. However, the approach depends on exact matching to find new words. In the second approach (SP-POS) we try to develop more general patterns by using the *Part-Of-Speech* (POS) information of the n-grams.

At the first step, we apply sentence extraction on the documents. Next, we conduct POS tagging in order to mark every word in the sentence with its part-of-speech, based on its definition and context. We use Indonesian POS Tagger developed by Pisceldo et al. [15], which uses probabilistic approach. In the tagging process, seed words found in the sentences are tagged with a special tag, '[SENT]'. After that, we extract n-grams (n = 3) from the tagged sentences. We ranks the n-grams based on their frequency and take top-N (N = 50) n-grams that contain [SENT] tag as SP-POS. Fig. 3 depicts the overall process to extract SP-POS.

After we develop the SP-POS patterns, we create parallel corpus, which consists of original sentences and its corresponding tagged sentences (without using seed words). We match the SP-POS patterns to the tagged sentences, and find n-gram that suitable with the SP-POS patterns, except the [SENT] tag, which can match any word. Finally, we look for the original words that fit the [SENT] tag in the parallel corpus and add them to the candidate lexicon.

Table 2. Negation and Transitional Words

Negation words	*Tidak* (not), *enggak* (not), *nggak* (not), *engga* (not), *ga* (no), *gak* (no), *gag* (no), *bukan* (not), *tiada* (no), *non* (not), *tak* (not), *kagak* (no), *kaga* (non)
Transitional words	*Tetapi* (but), *melainkan* (but), *padahal* (whereas), *sedangkan* (while), *tapi* (but), *namun* (however), *sebaliknya* (otherwise)

EXPANSION USING SENTENCE POLARITY (SPO) The next approach expands the seed lexicon using sentence polarity. (Terra and Clarke, 2003) propose technique to find words that have high similarity based on their co-occurrence in the corpus. Using the same idea, we try to find new sentiment words by its occurrences in polar sentences. A sentence is a polar sentence if it contains seed word(s). We assume that the occurrence information will implicitly define the relationship between seed words and candidate word.

EXTRACTING SENTENCE POLARITY To expand the seed lexicon, first we filter sentences that contain seed words. By default, the sentence polarity follows the seed word polarity. We also include some cases that may change the polarity of a sentence by searching transitional and negation words.

TRANSITIONAL WORDS We detect transitional words that appeared in the sentences. A subjective sentence may contain more than one polarity, as people can state what they like and dislike in one sentence. For example, "While the food is expensive, the taste is very delicious". In that sentence, we can find two kind of sentiment with different polarity. The reviewer likes the food but does not like the price of that food. Here, we list words that may change the polarity of a sentence. For this kind of sentence, we split the sentence into two sub-sentences with different polarity.

NEGATION WORDS We also detect negation words, such as "no" and "not" in the sentences. Negation words are used to detect polarity shifting.

SELECTING CANDIDATE WORDS After extract the polarity of sentence, we calculate polarity score of each word in the sentence. We adopt the

Pointwise Mutual Information (PMI) (Church and Hanks, 1989) to estimate polarity value. For each word *w* in the corpus, we calculate its two polarity score, positive polarity (*pos_polarity*) and negative polarity (*neg_polarity*). Sentiment polarity of a word *w* will have higher value when it frequently occurred in sentiment sentences. Sentiment polarity value is estimated as follows:

$$pos\_polarity(w) = \log_2 \frac{p(w,pos)}{f(w) \cdot \frac{f(pos)}{N}}, \quad (1)$$

$$neg\_polarity(w) = \log_2 \frac{p(w,neg)}{f(w) \cdot \frac{f(neg)}{N}}, \quad (2)$$

where *p*(*w, pos*) is the occurrence likelihood of word *w* in the positive sentences, *f*(*w*) is the frequency of sentences which contain word w, *f*(*pos*) is frequency of positive sentences, and N is total number of sentences. The same definition applied for negative polarity. We compute *p*(*w, pos*) *and p*(*w,neg*) as follows:

$$p(w,pos) = \frac{f_{(w,pos)}}{N}, \quad (3)$$

$$p(w,neg) = \frac{f_{(w,neg)}}{N}, \quad (4)$$

where $f_{(w,pos)}$ is the frequency of positive sentences that contain word *w*.

2.3 *Filtering Expansion Terms*

OPINION FILTERING We apply opinion filtering to remove non-sentiment words from the candidate sentiment words. Several study on sentiment analysis show that adjective words is effective to increase accuracy [3, 13]. In this phase, we simply remove non-adjective words based on our random observation that sentiment words are usually adjectives.

SENTIMENT DETECTION As lexicon expansion only collect candidate sentiment words without determining its polarity, in this step we detect sentiment of each candidate word. We detect the polarity of a word by calculate its two sentiment score, *sent_pos*(*w*) and *sent_ neg*(*w*):

$$sent_{pos(w)} = \frac{\frac{P(x|pos)}{P(pos)}}{\frac{P(x|pos)}{P(pos)} + \frac{P(x|neg)}{P(neg)}}, \quad (5)$$

$$sent_{neg(w)} = \frac{\frac{P(x|neg)}{P(neg)}}{\frac{P(x|pos)}{P(pos)} + \frac{P(x|neg)}{P(neg)}}, \quad (6)$$

where $P(x|pos)$ is the number of positive seed words in positive documents and $P(x|neg)$ is the number of positive seed words in negative documents. $P(pos)$ and $P(neg)$ is the number of positive and negative documents. A word is considered positive if its positive score is higher than negative score and vice versa.

3 Experimental Results

3.1 *Dataset*

In this study, we use three kind of dataset collected from social media data. We focus on domain specific sentiment lexicon, so we collect data from tourism domain. Dataset used in this experiment are collected from TripAdvisor[3], Twitter, and OpenRice[4]. TripAdvisor and OpenRice are user generated content (UGC) which contains reviews about Indonesian tourism and restaurants. We collect reviews from both sites and assign polarity value (negative or positive) based on the review ratings. As they use rating scale 1–5, we assign review with ratings (1–2) as negatives and (4–5) as positives. For Twitter data, we collect tweets using query about tourism sites in Indonesia. As building human annotated Twitter corpus requires considerable resources, we collect Twitter corpus using query that contains emoticons, i.e :-), :), :(, :-(, etc. We assume that a tweet is a subjective if it is contain emoticons and classify the tweets using positive and negative emoticons. Statistics of our dataset are shown in Table 3.

[3] http://www.tripadvisor.co.id
[4] http://id.openrice.com

Table 3. Dataset Statistics

Source	# positive reviews	# negative reviews
TripAdvisor	1139	229
OpenRice	3553	297
Twitter	8435	3381

3.2 *Lexicon Evaluation*

Lexicon evaluation was done manually by two annotators with Kappa value 0.729, which is considered substantial agreement. Both annotators judge the subjectivity and polarity for each candidate word. In subjective evaluation, annotators are asked to judge whether a candidate word is a sentiment word or not. Furthermore, for polarity evaluation, annotators are asked to judge whether a candidate word is positive or negative.

EXPANSION RESULTS From the lexicon expansion phase, all the three approaches can generate a number of candidate lexicons. SP and SP-POS generate a fair number of words as they use exact matching with patterns. SPo generates a large number of candidate words, because it uses word occurrences in sentences. The result of seed expansion process is shown in Table 4. This table shows the percentage (%inc) of lexicon increment relative to the initial lexicon (seed words).

Table 4. Seed Expansion Results

Dataset	%inc		
	SP	SP-POS	SPo
TripAdvisor	86%	132%	2624%
Twitter	203%	168%	5682%
Openrice	185%	172%	4740%

FILTERING RESULTS Tables 5 and 6 report the filtering result. We use two evaluation metrics; %inc to shows the number of new candidate words relative to the initial seed words and %acc to shows the accuracy of candidate words.

The opinion filtering results are shown in Table 5. As seen from the tables, after opinion filtering, SP generates candidate words with highest accuracy (89%) but with lowest expansion (23.98%). This is because SP generates specific sentiment patterns that not always occurred in the document (exact matching). On the other hand, SP-POS achieves 71.63%

accuracy but can expand the lexicon with the highest percentage at 105.33%. SP-POS can generate more candidate lexicon because it uses generalized patterns, which use part-of-speech information. SPo yields lowest accuracy (41.91%) with lexicon increment at 92.12%. SPo fails to generate candidate words with good accuracy because it rely on word occurrence in sentences, so that any words that frequently appear can become candidate sentiment words. The algorithm finds the correlation between words with assumption that a review sentence will contains more than one sentiment word. However, based on our observation, a review sentence does not always contain more than one sentiment words.

Table 5. Opinion Filtering Results

Dataset	SP		SP-POS		SPo	
	%inc	%acc	%inc	%acc	%inc	%acc
TripAdvisor	16.95%	97.5%	73.73%	75.60%	76.27%	41.51%
Twitter	16.75%	75.0%	101.83%	69.67%	86.13%	45.41%
Openrice	38.24%	94.5%	140.44%	69.63%	113.97%	38.80%
All	23.98%	**89.00%**	**105.33%**	71.63%	92.12%	41.91%

From the dataset perspective, SP and SP-POS generates best result with TripAdvisor dataset because it contains reviews with good sentence structure. SPo produces best result with Twitter because the algorithm does not count on the sentence structure and Twitter has the highest number of documents to construct correlation between seed words and candidate words.

The sentiment detection results are shown in Table 6. From the overall results, we can see that polarity detection accuracy for positive words is always better than negative words. This is because the dataset that we used in this study contains more positive documents then negative documents. Based on the results, we can see that our approach to detect polarity of a word produce consistent accuracy for all kind of dataset.

Table 6. Sentiment Detection Results

Dataset	TripAdvisor		Twitter		OpenRice	
	positive	negative	positive	negative	positive	negative
SP	91.20%	66.70%	90.50%	61.10%	91.60%	57.90%
SP-POS	77.70%	84.90%	76.20%	60.60%	84%	61%
SPo	56.50%	43.00%	52.90%	52.40%	50.30%	47.30%
All	**75.13%**	**64.87%**	**73.20%**	**58.03%**	**75.30%**	**55.40%**

4 CONCLUSION

In this paper, we propose our approaches in building domain specific sentiment lexicon for Indonesian language. Our main contributions are: (1) methods to expand sentiment lexicon using sentiment patterns; and (2) techniques to classify the polarity of a word using sentiment score.

The process start by translating English sentiment words to build seed lexicon. The seed lexicon is then expanded using senti-patterns (SP and SP-POS) and similarity with polar sentence (SPo) to produce candidate sentiment words. Finally, we apply two stages of filtering process, opinion filtering and sentiment detection to generate final list of expanded sentiment lexicon.

We test our proposed methods to build Indonesian sentiment lexicon for tourism domain with three kind of dataset which is different in the level of sentence structure. Yet, using the same techniques, it is also possible to implement this technique in other under-resourced languages, which can provide seed lexicon, POS tagger, and user reviews.

REFERENCES

1. Banea, C., Mihalcea, R., Wiebe, J.: A Bootstrapping Method for Building Subjectivity Lexicons for Languages with Scarce Resources. In: LREC (2008)
2. Brooke, J., Tofiloski, M., Taboada, M.: Cross-linguistic sentiment analysis: From English to Spanish. In: Proceedings of the 7th International Conference on Recent Advances in Natural Language Processing, Borovets, Bulgaria (2009) 50–54
3. Chesley, P., Vincent, B., Xu, L., Srihari, R.K.: Using verbs and adjectives to automatically classify blog sentiment. Training 580 (2006) 233
4. Esuli, A., Sebastiani, F.: SentiWordNet: A publicly available lexical resource for opinion mining. In: Proceedings of LREC (2006) 417–422
5. Kaji, N., Kitsuregawa, M.: Building Lexicon for Sentiment Analysis from Massive Collection of HTML Documents. In: EMNLP-CoNLL (2007) 1075–1083
6. Kamps, J., Marx, M.J., Mokken, R.J., De Rijke, M.: Using wordnet to measure semantic orientations of adjectives (2004)
7. Kanayama, H., Nasukawa, T.: Fully automatic lexicon expansion for domain-oriented sentiment analysis. In Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics (2006) 355–363

8. Kim, S.-M., Hovy, E.: Determining the sentiment of opinions. In: Proceedings of the 20th International Conference on Computational Linguistics (2004) 1367
9. Klebanov, B.B., Burstein, J., Madnani, N., Faulkner, A., Tetreault, J.: Building subjectivity lexicon(s) from scratch for essay data. In: Computational Linguistics and Intelligent Text Processing, CICLing 2012, LNCS, Springer (2012) 591–602
10. Liu, B.: Sentiment analysis and opinion mining. Synth. Lect. Hum. Lang. Technol **5** (2012) 1–167
11. Mihalcea, R., Banea, C., Wiebe, J.: Learning multilingual subjective language via cross-lingual projections. In: Annual Meeting of the Association for Computational Linguistics (2007) 976
12. Pantel, P., Pennacchiotti, M.: Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In: Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics (2006) 113–120.
13. Peng, W., Park, D.H.: Generate adjective sentiment dictionary for social media sentiment analysis using constrained nonnegative matrix factorization. Urbana **51** (2004) 61801.
14. Pérez-Rosas, V., Banea, C., Mihalcea, R.: Learning Sentiment Lexicons in Spanish. In: LREC (2012) 3077–3081
15. Pisceldo, F., Manurung, R., Adriani, M.: Probabilistic Part-of-Speech Tagging for Bahasa Indonesia. In: The Third International MALINDO Workshop, co-located Event ACL-IJCNLP (2009)
16. Qiu, G., Liu, B., Bu, J., Chen, C.: Opinion word expansion and target extraction through double propagation. Computational Linguistics **37** (2011) 9–27
17. Terra, E., Clarke, C.L.: Frequency estimates for statistical word similarity measures. In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, volume 1 (2003) 165–172
18. Turney, P., Littman, M.L.: Unsupervised learning of semantic orientation from a hundred-billion-word corpus (2002)
19. Wan, X.: Using bilingual knowledge and ensemble techniques for unsupervised Chinese sentiment analysis. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (2008) 553–561
20. Wan, X.: Co-training for cross-lingual sentiment classification. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, volume 1 (2009) 235–243
21. Wiebe, J., Mihalcea, R.: Word sense and subjectivity. In: Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics (2006) 1065–1072.

CLARA VANIA
FACULTY OF COMPUTER SCIENCE,
UNIVERSITAS INDONESIA,
INDONESIA
E-MAIL: <C.VANIA@CS.UI.AC.ID>

MOH. IBRAHIM
FACULTY OF COMPUTER SCIENCE,
UNIVERSITAS INDONESIA,
INDONESIA
E-MAIL: <MOCH.IBRAHIM@UI.AC.ID>

MIRNA ADRIANI
FACULTY OF COMPUTER SCIENCE,
UNIVERSITAS INDONESIA,
INDONESIA
E-MAIL: <MIRNA@CS.UI.AC.ID>