Advanced Science Letters
Vol. 4, 400–407, 2016

Non-English Sentiment Dictionary Construction

Khalifa Chekima[1] and Rayner Alfred[2]
Faculty of Computing and Informatics, Universiti Malaysia Sabah, Kota Kinabalu, Malaysia

Sentiment analysis (SA) in non-English world remains a challenging problem, as constructing language specific sentiment dictionary is an extremely resource intensive process. While different well-defined approaches have been defined for English SA, which resulted English language to be mature and resourceful when it comes to SA, the problem remains far from being solved for other languages such as Malay language, despite having more than 215 million Malay native speakers worldwide. To our knowledge, there is no publically available Malay sentiment dictionary. Most researchers dealing with lexicon based for Malay SA use a translated version of a well-known English lexicon called the SentiWordNet. In this paper, the lexicon gap is addressed by utilizing existing sentiment analysis resources and tools from English along with the automated machine translation capability to automatically build Malay sentiment dictionary which is of high quality and with large coverage of sentiment words. The architecture for constructing Malay sentiment dictionary is presented, comprises of 5 modules: (1) seed words extraction, (2) seed words propagation, (3) weak expression elimination, (4) expressions translation and (5) de-duplication. As a result, MySentiDic[1] was constructed comprises of 2,781 Malay sentiment words. To measure the accuracy of MySentiDic, lexicon data is compared with human annotator, the agreement score recorded high at 0.81. MySentiDic data is then compared with Malay translated version of SentiWordNet. The result was promising where MySentiDic recorded 0.22 more accuracy compared to translated version of SentiWordNet which recorded a lower accuracy of 0.58. The discussion and implication of these findings are further elaborated.

Keywords: Sentiment Analysis, Malay Sentiment Dictionary, Malay lexicon, Unsupervised Technique, Natural Language Processing, Data Analytics, SentiWordNet.

1. INTRODUCTION

Sentiment analysis (SA), also called opinion mining, is the field of study that analyses people’s sentiments, appraisals , evaluations, opinions, attitudes, and emotions towards entities such as services, products, individuals, organizations, , issues, events, topics, etc. [2]. There are two common approaches adopted by most researchers while dealing with SA, which may be classified into machine learning approach also known as supervised approach and lexicon based approach also known as unsupervised approach [1].

Researchers dealing with supervised approach mostly utilized one of machine learning techniques namely Support Vector Machine(SVM), Neural Network (NN), Naïve Bayes (NB), Maximum Entropy (ME) to build their classifiers [1]. According to [3], Classifiers built using supervised methods reach a high accuracy in detecting the polarity of a text as shown in [4][5][6][7]. However, even though such classifiers perform splendidly in the domain they are trained on, their performance drops significantly when the same classifier is used in different domains.

As for unsupervised approach, researchers used either Dictionary Based Approach or Corpus Based Approach. These techniques involve calculating orientation for a document from the semantic orientation of words or phrases in the document [8]. Dictionaries for lexicon-based approaches can be created manually, as described in [9], or automatically, using seed words to expand the list of words [10][8][11]. According to [2], even though lexical approach does not invariable outperform machine learning method, yet its overall track record is better.

1936-6612/2011/4/400/008 doi:10.1166/asl.2011.1261

Another researcher [12] claims that lexicon based method are robust, resulting in good cross-domain performance, and can be enhanced easily with multiple sources of knowledge.

The intention of this work is to create non-English sentiment dictionary of high quality, with large coverage of words that can contributes towards both supervised and unsupervised SA for Malay language. The rest of the paper is organized as follow. Section 2 discusses related work. Section 3 discusses the English lexicons/resource used in this research, including the preprocessing techniques involved. Section 4 discusses the process involved in constructing Malay sentiment dictionary (MySentiDic). Section 5 discusses the evaluation and result of MySentiDic. Finally, this paper is concluded in section 6.

2. RELATED WORK

Among the work conducted for non-English lexicon based sentiment analysis are Arabic [13] [20], Chinese [14] [22], French [15], German [16] [21], Japanese [17], Spanish [18], Romanian [23], Hindu and French [19] where some have manually constructed lexicons in their language, as for others, they adopted an automatic construction techniques, where they started with a list of words with known polarities like "good" and "bad", from there these words are automatically propagated to obtain new words that share the same polarities. Few techniques adopted in propagating seed words such as using thesaurus to obtain words synonyms and antonyms, others used Point-wise Mutual Information (PMI), etc.

As for Malay language, researchers such in [25] used the SentiWordNet to obtain word's polarity, by first translating Malay words into English, then utilized the SentiWordNet for classification purposes. One obvious drawback of using this technique is wrong polar classification, as the SentiWordNet accuracy is recorded to be low. For instance, the following words "kesalahan" which means "iniquity", "tidak bermoral" means "immorality", "jahat" means "evil" and "kejahatan" means "wickedness", are classified as positive expressions by the SentiWordNet, infect they are obviously negative. This will be further discussed in the coming sections.

3. LIST OF ENGLISH SUBJECTIVITY LEXICON

Five well known English lexicons were selected in this research to assist in constructing our golden seed words namely, GI (Harvard General Inquiries), Bing Liu's opinion lexicon, MPQA (Multi-perspective Question Answering), SentiWordNet and AFFIN-111.

These lexicons can be categorized into two categories based on the method they were built on, either automatic or manual construction. Table 1 lists down English lexicons, their corresponding number of terms/expressions and method of construction.

Table 1. English Subjectivity Lexicons and their Corresponding number of terms

Lexicon Name	Number of Expression	Construction Method
SentiWordNet	117,660	Automatic
GI	11,788	Manual
MPQA	8,219	Manual
Bing Liu's	6,788	Manual
AFINN-111	2,477	Manual

4. ENGLISH LEXICON PREPROCESSING

Before the English lexicons are used, an extra preprocessing (cleaning) step is needed to eliminate noise/unwanted data, as using these lexicons without further preprocessing leads to inaccurate translation results.

4.1. SYMBOL ELIMINATION

Symbols are characters that do not contribute towards an end results. Asides from being meaningless, these symbols tend to have high frequencies. From an observation made on the five lexicons, a list of common symbols/unwanted characters have been identified as follow "%#'|.:='".

4.2. UNWANTED COLUMN ELIMINATION AND DEDIPLICATION

Since only words and their corresponding values of being either positive or negative are crucial in constructing the Malay lexicons, the remaining columns are discarded from lexicons.

The final preprocessing step is getting rid of redundant words in lexicons. De-duplication process is crucial as it speeds up process by avoid reprocessing the same data more than once, as well as it saves space incase storage is an issue. The system keeps one copy of duplicated words and discards the rest.

Table 3. English Lexicons Number Positive and negative Expressions

Lexicon Name	Pos Expression	Neg Expression
SentiWordNet	5,691	3,586
G.I	1,635	2,015
MPQA	2,304	4,151
Bing Liu's	2,005	4,782
AFFIN-111	670	1,286

Table 3 lists down the clean/preprocessed version English lexicons and their corresponding final number of expressions. From now on, when the English lexicons are mentioned, we are actually refereeing to their clean version.

5. MALAY LEXICON CONSTRUCTION

This section discusses the techniques adopted in constructing Malay subjectivity lexicon as shown under Figure 1. First, weak expressions are removed from all of the English lexicons. Next, an intersection and union are performed on the five English lexicons. As a result, two

[*]Email Address: 1. k.chekima@gmail.com, 2. ralfred@ums.edu.my
1. Contact author at k.chekima@gmail.com for a copy of MySentiDic

new lexicons were produced, namely an intersection lexicon with high quality of seed words, and a union lexicon with high number of expressions (coverage). Next, a bootstrapping technique (based on word's synonym and antonym) were applied to all lexicons to propagate existing terms for expansion purposes. Once synonyms and antonyms terms obtained, next, all of the lexicons are translated from English to Malay.

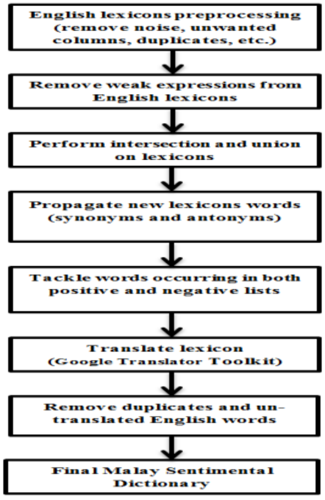

English lexicons preprocessing (remove noise, unwanted columns, duplicates, etc.)
Remove weak expressions from English lexicons
Perform intersection and union on lexicons
Propagate new lexicons words (synonyms and antonyms)
Tackle words occurring in both positive and negative lists
Translate lexicon (Google Translator Toolkit)
Remove duplicates and un-translated English words
Final Malay Sentimental Dictionary

Fig. 1. Steps Involved in Producing Potential Malay Lexicons

5.1. REMOVAL OF WEAK EXPRESSIONS

Since the aim is to construct a sentiment word dictionary, which is of high quality, removing weak expressions and keeping only those with strong values is essential. From SentiWordNet, expressions are stored in a form of synsets list, where each synset has negative and positive scores ranging from -1 to 1. Since SentiWordNet's expression score defines the level of an expression of being weak, average, fair or strongly positive/negative, a threshold was defined with value of 0.5 for positive expressions and -0.5 for negative expression. This threshold eliminates weak and average expressions and only keeps expressions which have fair or strong positive/negative expressions values.

Similar extraction technique applied to SentiWordNet was applied to AFFIN-111, where the threshold is set to 2 for positive, and -2 for negative scores.

For MPQA lexicon, expressions were extracted by looking at the 'priorpolarity' value of being either positive or negative. A similar technique were applied to extract expression from Harvard General Inquirer's lexicon, by looking at the 'Pos' and 'Neg' values. As for Bing Liu's lexicon, it was used as it is.

5.2. ENGLISH LEXICON INTERSECTION AND UNION

Two new lexicons produced by performing intersection and union procedure on existing English lexicons. The intersection was performed to extract agreed terms by different lexicons. Seed words present in this lexicon are of high quality, it is less likely to spot words with wrong classification. As for the union, it was performed to extract all possible terms from different lexicons. Since AFFIN-111 lexicon is relatively small compare to other lexicon, when intersection performed on the lexicons, a total number of 170 expressions retrieved, as for positive expression, a total of 103 positive expressions retrieved, for that, AFFIN-111's lexicon was ignored for intersection operation. Table 4 list down number of terms obtained as result of the intersection and union.

Table 4. Intersection and Union Lexicons and their Corresponding number of terms

Lexicon's Name	Pos Expression	Neg Expression
Intersection	297	456
Union	6,031	10,063

5.3. ENGLISH LEXICON PROPAGATION

To expand/propagate English lexicons, we hypothesize that synonyms of positive words are mostly positive and the antonyms are negative. A program was developed that scans through each of the lexicons' expressions, and automatically fetches synonym and antonym for each expression if available from the Thesaurus[1]. For instance, the word "happy" will have the following synonym words "overjoyed", "cheerful", "joyful", "elated", "contented" and "ecstatic". Figure 2 is a visualization of links between the word "happy" and its synonyms.

Synonyms of positive words are added to the positive list and the antonyms are added to the negative list and vice versa. Every time a new list of synonyms and antonyms retrieved, iteration through this list is carried out to extract the corresponding synonyms and antonyms. The iteration has been conducted for four times until an obvious overlap of expressions is noticed. One advantage of iterating through new list is to extract as much possible of new positive/negative expression to enrich existing lexicons, since one of the important characteristics of a lexicon is having a wider coverage.

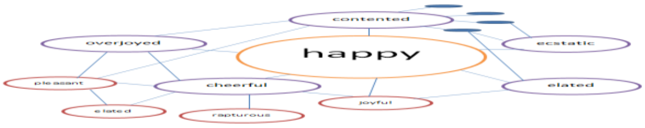


Fig. 2. Visualization of links between words and their corresponding synonyms

However, not all of the synonyms and the antonyms retrieved can be used, as some occurred in both positive and negative categories such as "strange", "overtake", "stout" which may lead to word ambiguity. To tackle this, two options have been identified; first to discard these words from both lists, the second option is to develop polarity strength measure by counting the number of occurrences for each of the word's synonyms in both positive and negative lists. The intuition of this technique is, the more a word's synonym occurring in one of the classes positive or negative the more likely it belongs. In this research, the second technique is deployed since the purpose is to maximize number of expression for each lexicon without lowering its quality.

First, the probability $P(c|w)$ of word w given class c

is measured, where class c can be either positive class or negative. The probability is measured by computing the occurrence of word w's synonyms in the list of class c divided by the total number of all synonyms $syns(w)$ for word w.

Formula to compute the probability:

$$P(c \mid w) = \frac{\sum_{i=1}^{n} \mathrm{count}(syn_i, c)}{\mathrm{count}\ (syns(w))} \qquad 1$$

Once the probability values retrieved for both negative and positive class, next a simple prediction method is used to decide to which category a word w should belong to. The polarity of a given word w that contains a list of positive synonyms P and a list of negative synonyms N is defined as follows:

$$polarity(w) = \begin{cases} positive & \frac{\sum_{i=1}^{n} count(syn_i,P)}{count(syns(w))} > \frac{\sum_{i=1}^{n} count(syn_i,N)}{count(syns(w))} \\ negative & \frac{\sum_{i=1}^{n} count(syn_i,P)}{count(syns(w))} < \frac{\sum_{i=1}^{n} count(syn_i,N)}{count(syns(w))} \\ discard & \frac{\sum_{i=1}^{n} count(syn_i,P)}{count(syns(w))} = \frac{\sum_{i=1}^{n} count(syn_i,N)}{count(syns(w))} \end{cases} \qquad 2$$

The probability values for each class is then compared, if positive probability value is greater than the negative probability value by exceeding certain threshold, word w belongs to positive list/class, if negative probability is higher, word w belongs to negative class, and in the case of draw, the word w will be discard from both lists. The default value for positive strength is 1 and -1 for negative strength. Table 5 shows several examples of words and their corresponding positive and negative strength obtain as result of using the above formula. The word "prodigy" for instance was classified as a strong positive expression with strength value of +0.937. The word "stout" on the other hand, was classified as fair positive expression as its strength value was only +0.66.

By using this technique, the chances of assigning wrong polarity to new words during lexicon propagation can be reduced. This technique can tackle the problem of SentiWordNet, where it assigns wrong polarity to words, which directly affected its accuracy. An example of this is when SentiWordNet assigns polarity of +0.75 for the synset {iniquity, immorality, evil, wickedness}, where this words are obviously negative. Using the technique above, "inequality" has negative score of -0.8, "immorality" = -1, "evil" = -1 and "wickedness" = -1, which changes its wrong polarity from positive +0.75 to the correct strong negative polarity with strength value of -0.95.

Table 5. Example of Lexicon's Expressions and their Corresponding Strengths

Word	No. of word's Synonyms	No. of synonym in each list		Word's Strength		Word's Class
		Positive list	Negative List	Positive	Negative	
prodigy	16	15	1	+0.937	-0.062	positive
stout	5	3	2	+0.66	-0.33	positive
kill	11	9	2	+0.18	-0.81	negative
shock	13	2	11	+0.15	-0.84	negative
abound	5	4	1	+0.8	-0.2	positive

5.4. ENGLISH LEXICON TRANSALTION

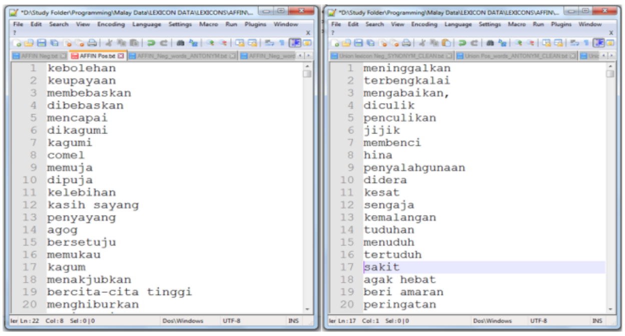
From a review conducted on several available machine translation, such as Moses, Google translation, Babylon, etc. Google machine translation has shown to perform quite well, due to that, Google translator toolkit was used as translation tool.

5.5. MALAY LEXICON PREPROCESSING

The final step in Malay lexicon construction is to remove duplicates and un-translated terms. Even though de-duplication has been performed earlier on English lexicons, yet it is essential to recheck for duplication after translation, as there is a chance for two or more English words to point to the same Malay word after translation, for instance, the words "happy", "overjoyed", "rapturous", "ecstatic" and "elated" when translated into Malay they refer to the same Malay word "gembira". Figure 3 shows data sample for one of the final Malay sentiment dictionary.

5.6. FINAL MALAY LEXICON

As result of expanding the intersection and union lexicons, a total of 2,028 new expressions added to intersection lexicon, which 917 of it were added to the positive list and 1,111 to the negative list, on the other hand, a total of 9,835 new expressions were added to union lexicon, which 3,895 of it were added to the positive list and 5,940 were added to the negative list as shown in Table 6. The following section discusses the evaluation of the newly produced lexicons


*D:\Study Folder\Programming\Malay Data\LEXICON DATA\LEXICONS\AFFIN\...
File Edit Search View Encoding Language Settings Macro Run Plugins Window ?
AFFIN Neg.txt
AFFIN Pos.txt
AFFIN_Neg_words_ANTONYM.txt
AFFIN_Neg_word
ler Ln : 22 Col : 8 Sel : 0 | 0 Dos\Windows UTF-8 INS
*D:\Study Folder\Programming\Malay Data\LEXICON DATA\LEXICONS\AFFIN\...
File Edit Search View Encoding Language Settings Macro Run Plugins Window ?
Union lexicon Neg_SYNONYM_CLEAN.txt
Union Pos_words_ANTONYM_CLEAN.txt
ler Ln : 17 Col : 1 Sel : 0 | 0 Dos\Windows UTF-8 INS

```
kebolehan
keupayaan
membebaskan
dibebaskan
mencapai
dikagumi
kagumi
comel
memuja
dipuja
kelebihan
kasih sayang
penyayang
agog
bersetuju
memukau
kagum
menakjubkan
bercita-cita tinggi
menghiburkan
```

```
meninggalkan
terbengkalai
mengabaikan,
diculik
penculikan
jijik
membenci
hina
penyalahgunaan
didera
kesat
sengaja
kemalangan
tuduhan
menuduh
tertuduh
sakit
agak hebat
beri amaran
peringatan
```

Positive List Negative List

Fig. 3. Sample of Final Malay Lexicon's lists

Table 6. Final Number of Positive Expressions in Malay Lexicons

Source Lexicon	Positive	Negative
Intersection	1,214	1,567
Union	9,890	16,003

6. MALAY LEXICON EVALUATION SETUP AND RESULTS

Since the intention of this evaluation is to measure the accuracy of sentiment dictionary's data for being positive or negative, a total of 500 expressions were randomly selected from each lexicon (250 Negative & 250 Positive) as an evaluation data. Next, two Malay native speakers were asked to classify expressions in each of the evaluation data as positive or negative. Their results are then compared to lexicons for evaluation purposes. To measure the agreements between the two human annotators H-H and the lexicons versus human L-H, Cohan’s Kappa[24] was used. Based on Cohan’s Kappa, if the raters are in complete agreement, then k value is 1. In case no agreement among the raters other than what would be expected by chance, k value will be $k \leq 0$.

Table 7. Human Annotators and Lexicon Versus Human Agreement Using Cohan’s Kappa

Lexicon	H-H	L-H
Intersection	**0.839**	**0.810**
Union	0.692	0.632
SentiWordNet	0.710	0.587

In this experiment, the Malay translated version of the SentiWordNet is included to evaluate its data polarity accuracy as it is one of the widely used lexicon in constructing non-English lexicons. As can be observed from Table 7, intersection lexicon shows a highest agreement from human annotators. Union lexicon on the other hand shows less agreement from human annotators, yet, its result is reasonably fair compared to human judgment, at the same time it outperform the SentiWordNet.

7. CONCLUSION

In this paper, a potential English resources that can contribute in building non-English sentiment dictionary has been identified. Asides from that, a technique that filters out non-relevant/weak polarity words during lexicons’ propagation process to keep its data at highest quality possible is introduced. As result of proposed techniques, more than 2,000 new words have been deduced to intersection lexicon without affecting its accuracy, and more than 9,000 new words were induced to union lexicon with reasonable accuracy drop, which at the end both of the union and intersection lexicons performed better than the translated Malay version of the SentiWordNet. The technique presented in this paper is believed to be applicable to other non-English lexicons construction.

REFERENCES

[1] Medhat, W., Hassan, A., Korashy, H.: Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal, Vol 5, Issue 4, pp. 1093--1113 (2014)

[2] Liu, B.: Sentiment Analysis and Opinion Mining. Claypool Publishers . (2012)

[3] Taboada, M., Brooke, B., Tofiloski, M., Stede., M.: Lexicon-Based Methods for Sentiment Analysis. Computational Linguistics archive. Vol. 37, Issue 2, pp. 267--307. MIT Press Cambridge (2011)

[4] Chaovalit, P., Zhou, L.: Movie review mining: A comparison between supervised and unsupervised classification approaches. In Proceedings of the 38th Hawaii International Conference on System Sciences, Hawaii (2005)

[5] Alistair. K., Inkpen, D.: Sentiment classification of movie and product reviews using contextual valence shifters. Computational Intelligence, 22(2):110--125. (2006).

[6] Boiy, E., Hens, H., Deschacht, K., Moens,. M.F.: Automatic sentiment analysis of on-line text. In Proceedings of the 11th International Conference on Electronic Publishing, pages 349–360, Vienna (2007)

[7] Bartlett, J., Albright, R.: Coming to a theater near you! Sentiment classification techniques using SAS Text Miner. In SAS Global Forum 2008, San Antonio, TX (2008)

[8] Turney, P.: Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In Proceedings of 40th Meeting of the Association for Computational Linguistics, pp. 41--424 (2002.)

[9] Stone, P.J., Dexter C. D, Marshall S. Smith, Daniel M.O.: The General Inquirer: A Computer Approach to Content Analysis. MIT Press (1966)

[10] Hatzivassiloglou, Vasileios, H., McKeown, K.: Predicting the semantic orientation of adjectives. In Proceedings of 35th Meeting of the Association for Computational Linguistics, pp. 174--181 (1997)

[11] Turney, P., Littman. M.: Measuring praise and criticism: Inference of semantic orientation from association. ACM Transactions on Information Systems, pp.315–346 (2003)

[12] Feldman, R.: Techniques and Applications for Sentiment Analysis. Communications of the ACM. 56(4): 82-89 (2013)

[13] Fawaz H.H. Mahyoub, Muazzam A. Siddiqui Mohamed Y. Dahab.: Building an Arabic Sentiment Lexicon Using Semi-supervised Learning. Journal of King Saud University Computer and Information Sciences. Journal of King Saud University – Computer and Information Sciences, pp. 417—424 (2014)

[14] He Y, Alani H, Zhou D: Exploring English lexicon knowledge for Chinese sentiment analysis. In: CIPS-SIGHAN Joint conference on Chinese language processing (2010)

[15] Ghorbel, H., Jacot, D.: Sentiment Analysis of French Movie Reviews. Advances in Distributed Agent-Based Retrieval Tools, vol 361 of the series Studies in Computational Intelligence, pp. 9--108 (2011)

[16] Remus, R., Quasthoff, U., Heyer, H.: Sentiws – a publicly available german-language resource for sentiment analysis, in Proceedings of the 7th International Conference on Language Resources and Evaluation, pp. 1168—1171 (2010)

[17] Kaji, N., Kitsuregawa, M.: Building lexicon for sentiment analysis from massive collection of HTML documents. In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 1075--1083 (2007)

[18] Brooke, J., Tofiloski, M., & Taboada, M.: Cross-linguistic sentiment analysis: 565 From english to spanish. In Proceedings of the RANLP Conference, pp. 50–54 (2009)

[19] Rao, D., & Ravichandran, D.: Semi-supervised polarity lexicon induction. In Proceedings of the 12th conference of the european chapter of the association for computational linguistics (pp. 675–682). Association for Computational Linguistics (2009)

[20] Abdul-Mageed, M., Diab, M. T., & Korayem, M.: Subjectivity and sentiment analysis of modern standard arabic. In ACL (Short Papers), pp. 587–591 (2011)

[21] Clematide, S., & Klenner, M.: Evaluation and extension of a polarity lexicon for german. In Proceedings of the first workshop on computational approaches to subjectivity and sentiment analysis (2010)

[22] Lu, B., Song, Y., Zhang, X., Tsou, B. K.: Learning chinese polarity lexicons by integration of graph models and morphological features. In Information retrieval technology (pp. 466–477). Springer (2010)

[23] Banea, C., Mihalcea, R., Wiebe, J.: A bootstrapping method for building subjectivity lexicons for languages with scarce resources. In LREC, pp. 2764-- 2767 (2008)

[24] Cohen, J. J.: Weighted Kappa; Nominal scale agreement with provision for scaled disagreement or partial credit. Psychological Bulletin, 70, 213--220 (1968)

[25] Shamsudin, N. F., Basiron, H., Saaya, Z., Rahman, A. F. N. A., Zakaria, M. H., & Hassim, N.: Sentiment Classification of Unstructured Data Using Lexical Based Techniques. Jurnal Teknologi, 77(18).(2015)