Sentiment Lexicon Construction Using SentiWordNet 3.0

CONFERENCE PAPER · AUGUST 2015

3 AUTHORS:

Nishantha Medagoda

Auckland University of Technology

2 PUBLICATIONS **0** CITATIONS

SEE PROFILE

Subana Shanmuganathan

Auckland University of Technology

41 PUBLICATIONS **66** CITATIONS

SEE PROFILE

Jacqueline L. Whalley

Auckland University of Technology

54 PUBLICATIONS **609** CITATIONS

SEE PROFILE

Sentiment Lexicon Construction Using SentiWordNet 3.0

Nishantha Medagoda, Subana Shanmuganathan, Jacqueline Whalley
School of Computer and Mathematical Sciences
Auckland University of Technology
Auckland, New Zealand

Abstract—**Opinion mining and sentiment analysis have become popular in linguistic resource rich languages. Opinions for such analysis are drawn from many forms of freely available online/ electronic sources, such as websites, blogs, news re-ports and product reviews. But attention received by less resourced languages is significantly less. This is because the success of any opinion mining algorithm depends on the availability of resources, such as special lexicon and WordNet type tools. In this research, we implemented a less complicated but an effective approach that could be used to classify comments in less resourced languages. We experimented the approach for use with Sinhala Language where no such opinion mining or sentiment analysis has been carried out until this day. Our algorithm gives significantly promising results for analyzing sentiments in Sinhala for the first time.**

Keywords- Opinion Mining; Sentiment Analysis; Sentiment Lexicons;

I. Introduction

Opinions are subjective expressions of human thoughts, emotions and feelings. The research area of analyzing opinions contained in texts is popularly known as opinion mining or sentiment analysis. For a customer interested in finding specific information on a certain product or service, an opinion mining system helps the individual enormously in the investigation. The relevant information can be gathered using opinion mining tools without depending upon the verbal comments of the clients who has already used the same product or service. Governments as well as political parties benefit immensely from the review analysis by predicting the election results based on the comments given by the public using the social networks as the media of raising their voice. The manufacturers or merchants will be benefited by using opinion mining systems. They might be interested in determining the success of a new version of a product or service based on its popularity or identifying the demographics that likes or dislikes the special features of the commodity before launching the new advertising campaign. Identifying such information systematically by opinion mining tools saves time and money significantly than conducting time consuming surveys or market research. In addition, the results are much accurate and reliable since the data has been created by real customers in ideal situations without forcing them.

A given opinion can be classified as either a positive or negative or objective one, depending upon the opinion's polarity towards or against the theme of the topic being talked about. Some occasions, the opinion does not say anything about the topic being talked about; such neutral opinions can be considered as objective opinions. This whole procedure is known and subjectivity classification [1].

The primary resource required for classifying an opinion based on the above de-scribed categories using a supervised method in a given language is, the lexicon type repository called Sentiment lexicon. Sentiment lexicon usually contains a special set of words for the language with polarity scores either positive or negative. The polarity score is a scale used to determine the sense of the word that is present in the opinion. Hence, an opinion can be categorized into positive, negative or objective by combining the polarity scores of the sentiment words in that opinion. The languages that do not have similar resources with other utilities, such as, text corpus, morphological analyzer, translation model, WordNet and other such word depositories can be considered as less resourced languages as far as language processing is concerned.

Based on Ethnologue: Of the 7105 living Languages in the World [2] 63% of them are spoken in Asia and Africa. The chances of having linguistic language processing resources for these languages are limited as most of the countries are being developed. Manual Compilation of such a subjective lexicon for a given language is a challenging task as it always consumes enormous time and manpower. The literature survey conducted for this research reveals that most of the languages use the Word-Net to construct a Subjective lexicon. This paper describes a fast and less complex method of building a subjective lexicon for less resourced languages. In this study, the approach is being tested using a case study of more than two thousand opinions from the Sinhala Language spoken by 70 % of the 22 million people in Sri Lanka.

The remainder of the paper is organized as follows. Previous related work on this theme and related topics by other researchers is explained in section 2. Section 3 proposes the methodology of building a sentiment lexicon and the classification methods being investigated for testing the applicability of the Sinhala lexicon in this study. The components of the experiment are given in section 4. The results of the case study in Sinhala with accuracy measurements are presented in Section 5. Section 6 discussed the paper with anticipations relating to future research directions. Finally, section 7 presented the conclusion.

II. Related Works

Many researchers have addressed the problem of constructing subjective lexicon for different languages in recent years. To compile a subjective lexicon Bing Liu [1], investigated three main approaches and they are outlined in this section.

A manual approach described to be the simplest form however, it is a very time consuming process. The accuracy of the words collected has been improved by combining an automated method to this manually generated lexicon. The automated approaches are two types; the main one is dictionary based and the other is corpus based. In the dictionary based method a compilation process is initiated using a small word list known as the seed list. Normally, this seed list is manually constructed using adjectives and adverbs with their orientation (polarity). The seed list is then propagated through an online dictionary, such as WordNet, to grow the list by adding new terms by searching for the synonyms and antonyms of the seed list words. The weakness of the subjective lexicon constructed in this nature is, it becomes domain specific in orientation.

There is an alternative approach to overcome the domain specificity is the corpus-based method. In the dictionary-based method, the seed list is searched through the corpus searching for any syntactic or co-occurrence patterns of the seed word. An additional adjective (adverb) of the seed word with its orientation is added to the list using a set of constrains or conventions on connectives. The most of the rules or constraints are designed using the connectives “and”, “or”, “but”, “either-or” and “neither-nor”. These linguistic rules are called sentiment consistency. One of the limitations of this method is, building a corpus that represents all the words in a language is impossible.

A Hindi subjective lexicon constructed and discussed in [3] consisted of a seed list of 45 adjectives and 75 adverbs. In the adjective list 15 of each positive, negative and objective adjective were considered. Similarly, the same polarities but 25 of each in adverb collection were included in the adverb seed list. The Breadth First search was performed to expand the seed list on a graph based WordNet where words were connected to each other to indicate their synonyms and antonym relations. A new word was appended to the list assigning the polarity of the word using an as-assumption that synonym carries the same polarity and antonym shows the opposite polarity of the root word (the seed list word). In the method, the authors had man-aged to build a Hindi subjective lexicon with 8,048 adjectives and 888 adverbs. The new subjective lexicon was then evaluated using two methods: by human judgment and by simple classification on pre-annotated product review data set. In the classification method, firstly, the authors identified the adjectives and adverbs using a shallow parser. The weighting of the review was calculated using the unigram: defined as a single adjective or adverb, with a positive, a negative and an objective polarity. The maximum count was used as the final score. The authors had commented the reason for the poor agreement as the ambiguity in Hindi words. However, later an approximately 80% accuracy rate had been achieved using the same proposed classification method but by stemming the words in the review that did not have a matching subjective lexicon. Negation also was handled in the classification method.

Huang [4] utilized the chunk dependency knowledge to extract the domain-specific sentiment lexicon based on constrained label propagation. They had divided the whole strategy into six steps. Firstly, detected and extracted domain-specific sentiment terms by combining the chunk dependency parsing knowledge and prior generic sentiment lexicon. To refine the sentiment terms some filtering and pruning operations were carried out by others. Then they selected domain-independent sentiment seeds from the semi-structured domain reviews which had been designated manually or directly borrowed from other domains. As the third step, calculated the semantic associations between sentiment terms based on their distribution contexts in the domain corpus. For this calculation, the point-wise mutual information (PMI) was utilized which is commonly used in semantic linkage in information theory. Then, they defined and extracted some pair wise contextual and morphological constraints between sentiment terms to enhance the associations. The conjunctions like “and” and “as well as” were considered as the direct contextual constraints whereas “but” was referred to as a reverse contextual constraint. The above constraints propagated though out the entire collection of candidate sentiment terms. Finally, the propagated constraints were incorporated into label propagation for the construction of domain-specific sentiment lexicon. The proposed approach showed an accuracy increment of approximately 3% over the baseline methods.

III. Methodology

In this research a sentiment lexicon for Sinhala Language has been developed with the aid of English sentiment lexicon (SentiWordNet 3.0) compiled by Esuli and Sebastiani [5]. When classifying lexicon opinions in Sinhala, they can be broken down into a set of lexicon words (adjectives and adverbs) along with a positive and a negative score for each word. A feature vector for the classification is constructed using the total positive and negative score extracted from the constructed lexicon along with other features.

The English SentiWordNet 3.0 used in this study, comprises more than 100,000 words that occur in different context along with their positive and negative scores. In addition, a part of speech (POS) tag for each word is also included in the SentiWordNet 3.0.

The English SentiWordNet 3.0 was mapped to an online Sinhala dictionary using the English word in the dictionary as the search key to build the Sinhala sentiment lexicon. The English/Sinhala dictionary contains synonyms for each Sinhala word and an English word as the direct translation for the original Sinhala word. Then the sentiment score for the English word in SentiWordNet is assigned to the Sinhala word and its synonyms. Through this approach, a Sinhala sentiment lexicon is generated by mapping the English words in the English/Sinhala dictionary to the sentiments in SentiWordNet 3.0. The advantage of this method is it does not require a translation tool or software unlike other multilingual sentiment approaches. The following assumptions were made in the construction of this Sinhala sentiment lexicon when combining the English/Sinhala dictionary with the English SentiWordNet 3.0 to simplify the process of combining two dictionaries from parallel languages. Firstly, the sense of the word in the two languages was assumed to be the same. Secondly, the sentiment score of an English word calculated for use in English opinions, was considered as same for the matched Sinhala word. Finally, POS in both languages were considered as equivalent.

In the initial mapping, each English/Sinhala dictionary word was searched for a matching English word in SentiWordNet 3.0; 72,049 matches were found for the 10,000 English words. These exhaustive searches consisted of several matching English words embedded in POS. But for this experiment only Adjectives and Adverbs were added to the list as the two are the most important language units (part of speech) when analyzing sentiments in any language [6]. Through this selection initially 10,778 Sinhala adjectives were obtained in different context (POS) where the corresponding English term occurred. Besides this count, for Sinhala adverb search, there were 1,364 matches in different POS in English. The experiment was further continued with another assumption that POS for Sinhala words were considered to be same as those of in English. With this assumption, complexities relating to POS within a Sinhala sentence have been avoided. Hence, the final Sinhala lexicon (adjective and adverb list) obtained without POS but with positive and negative sentiment scores same as the corresponding English word in SentiWordNet 3.0 consisted of unique 5,973 adjectives and 405 adverbs.

The constructed sentiment lexicon was then evaluated using 2,083 manually classified news article opinions collected from a leading Sinhala newspaper website. The opinions supportive of the article were classified as positive (P) whereas, criticizing the topic were marked as negative (N) and any unrelated to the topic (neutral) were classified as objective (O).

This lexicon matrix of adjective and adverb scores (a set of positive and negative for both) was used to calculate the scores for the 2,083 opinions already classified as positive, negative or neutral opinion. An parser was implemented to traverse through the opinions searching for the adjectives and adverbs in each opinion and then assigning the total positive and total negative scores for the lexicon words in that opinion. These total positive and negative scores calculated for all the adjectives and adverbs in an opinion were used as the input vector for that opinion in the classification analysis

A. *Classification Algorithms*

Classification is the procedure run to identify the properties that indicate the group to which each case belonged. The methods of determining the semantic orientation used for identifying the polarity of the sentence are categorized into two approaches: supervised and unsupervised classification techniques [1]. In this experiment, only supervised classification algorithms were tested as the category of the sample opinions were already in place with the comment. The most common supervised state-of-practice algorithms in sentiment analysis are Naïve Bayes and Support vector machine. Naïve Bayes algorithm is the most widely used one and it is a simple but effective supervised classification method [7]. On the other hand, Support Vector Machine (SVM) is also tested in this study as it is a more efficient algorithm in sentiment classification [8]. Along with this two algorithm, one decision tree method, namely, J48 (using WEKA data mining software tool) was also investigated. The purpose of testing the J48 algorithm is to find the rules in the classification of opinions using adjectives and adverbs.

Naïve Bayes Algorithm.

The Naive Bayes Classification technique is based on the so called Bayesian theorem and is particularly suited when the dimensionality of the inputs is high.

Let R = {R_1, R_2, R_3,….R_n} denote the set of training opinions, where each opinion is labeled with one of the cording in C = {P, N, O}. Given some new opinion, the aim is to estimate the probability of each code. Using Bayes rule, in general

$$p(c/r) = \frac{p(r/c)p(c)}{p(r)}$$

As we are interested in the relative order of codes for a given opinion r, p(r) is independent of codes, then we can consider

$$p(c/r) = p(r/c)p(c)$$

If F denotes the ordered sequence of the features that compose the opinion R then F= {w_1, w_2, w_3, w_4}
Where,

w_1=Adjective_Positive_Score, w_2=Adjective_Negative_Score, w_3=Adverb_Positive_Score, w_4=Adverb_Negative_Score,

$$p(c/r) = p(r/c)p(c) = p(c)\prod_{i=1}^{p} p(w_i/c)$$

And classify r into the most possible code c using

$$\underset{c}{\operatorname{argmax}}\, p(^{C}/_{r})$$

Support Vector Machine(SVM)

Support vector machine (SVM) is the best binary classification method [9] proposed by Vladimir Vapnik. SVM is a nonprobabilistic classification technique that looks for a hyperplane with the maximum margin between the positive and negative examples of the training opinions. Support Vector Machines are based on the concept of decision planes that define decision boundaries. A decision plane is one that forms a separation between a set of objects, which have different class memberships. Decision planes are the classifiers either a line or a curve. A simple classifier may use liner decision planes rather than more complex structures. Classification tasks based on drawing separating lines to distinguish between objects of different class memberships are known as hyperplane classifiers.

SVM is primarily a classification method that performs classification tasks by constructing hyperplanes in a multidimensional space that separates cases of different class labels [10]. SVM supports both regression and classification tasks and it can handle multiple continuous and categorical variables.

To construct an optimal hyperplane, SVM employs an iterative training algorithm; this is used to minimize an error function. According to the form of the error function, SVM models can be classified into distinct groups.

In the simplest SVM, training involves the minimization of the error function.

$$\frac{1}{2}w^T w + C\sum_{i=1}^{N}\xi_i$$

Subject to the constrains

$$y_i\left(w^T\phi(x_i)+b\right)\geq 1-\xi_i \text{ and } \xi_i \geq 0,\ i=1,\ldots,N$$

Where C is the capacity constant w is the vector of coefficients, b, a constant and ξ are parameters for handling non separable data (inputs). The index i labels the N training cases. Note that y (∈±1) is the class label and x_i is the independent variables. The kernel ϕ is used to transform data from the input (independent) to the feature space. It should be noted that the larger the C, the more the error is penalized. Thus, C should be chosen with care to avoid over fitting.

It is suggested in [11] that SVM does not depend on the dimensionality of the problem when compared with other machine learning methods. The success of SVM in text categorization lies in its automatic capacity tuning by minimizing ǁ , i.e. the extraction of a small number of support vectors from the training data that are relevant for the classification.

Decision Tree classification Algorithm.

Decision tree is a graph with branches that represent every possible outcome of a decision. The rules produced by a decision tree model are human readable and are easily interpretable. The classification task using decision tree technique can be performed without complicated computations and the technique can be used for both continuous and categorical variables [12]. In this work, a decision tree model was tested to classify comments broken down to Positive, Negative or Neutral and then the rules generated by the decision trees were investigated.

J48 –Decision Tree Algorithm.

J48 is a univariate decision tree classification method that creates trees based on the information gain [13]. Initially, it tests whether all cases belong to the same class; if true then the tree is a leaf and is labeled as a class. Next for each attribute calculate the information gain. The information gain can be calculated as

$$Gain(p,f) = Entropy(p - entropy(f\|p)$$

Where,

$$Entropy(f\|p) = \frac{pf}{p}\log\frac{pf}{p}$$

Finally, find the best fitting attribute based on the current selection criteria. Once the initial tree is constructed using the entropy then pruning is carried out in order to remove the outliers and address the over fitting.

IV. Experiment

A. *The Sample*

We performed an experiment for testing the usability of the constructed sentiment lexicon using a set of Sinhala opinions. 2,083 comments were extracted from a leading online newspaper called “lankadeepa” (http://lankadeepa.lk/). As this is a domain independent classification, a sample consisting of different news articles was chosen, the domains included in the sample were; Political, Criminal, Education, Religion, Medical and General. The sample consisted of general and political discussions rather than the Medical and Religion related news. The data set contains 44,426 words with average comment length of 21 words. The tested sample was comprised of 745 positive (P), 838 Negative (N) and 500 neutral (O) opinions.

B. *The constructed Lexicon*

The constructed sentiment lexicon contains 5,973 Adjectives and 405 Adverbs with their positive and negative scores. In the Sinhala sentiment lexicon, the majority of the adjectives have more than one synonym, the maximum being 12. Similarly, for adverbs, 4 synonyms were obtained as the maximum.

V. Results

The experiments were conducted using the resources and classification algorithms described in the methodology section and the accuracy was evaluated using precision and recall. These standard measures have significantly higher correlation with human judgments [14]. These are first defined for the simple case where a classification system returns the categories.

Precision (P) is the fraction of retrieved opinions that are relevant whereas Recall (R) is the fraction of relevant opinions that are retrieved.

In general, the measures of precision and recall are used to evaluate the accuracy of the methods used in the testing of an approach being investigated. In this case, true positives and negatives returned indicate the percentage of the relevant opinions being classified correctly. A single measure that indicates the tradeoff between precision versus recall is the F measure, which is the weighted harmonic mean of precision and recall. F score is a measure of a test's accuracy. There are different weights that can be used to calculate the F measure.

In the first attempt of classification, Naïve Bayes, SVM and J48 algorithms were tested for all the classes; i.e., Positive, Negative and Neutral. The accuracy and other evaluating indexes such as precision and recall are shown in the table 2.

TABLE I. Classification Accuracies approach 1

	Classification Method		
	Naïve Bayes	**J48**	**SVM**
Accuracy (%)	**44**	**44**	**43**
Precision	**0.357**	**0.352**	**0.34**
Recall	**0.441**	**0.438**	**0.432**
F Value	**0.392**	**0.39**	**0.29**

The accuracy was around 44%, less than the bench mark values so far observed for the English and other well-resourced languages including some of Asian languages. The confusion matrix reveals the reason for the poor classification results as, the inclusion of the Neutral (O) category along with the other two.

In view of the above fact, a second experiment of the same study was carried out using a binary classification approach. In this second trial, opinions classified as positive and negative were only trained and tested. Here again, Naïve Bayes, J48 methods, SVM were tested using the sample of 1,583 opinions, which categorized as Positive or Negative. The accuracy measurements are given in table 3.

TABLE II. Classification Accuracies approach 2

	Classification Method		
	Naïve Bayes	**J48**	**SVM**
Accuracy (%)	**60**	**58**	**56**
Precision	**0.593**	**0.581**	**0.541**
Recall	**0.598**	**0.577**	**0.55**
F Value	**0.538**	**0.578**	**0.412**

Based on the table 3, 12 to 16 % improvements in accuracy in all three algorithms were observed but F value was still remained lower than 50% in support vector machine algorithm.

A tree constructed using the J48 algorithm in order to investigate the rules generated by the algorithm is presented in figure 3.

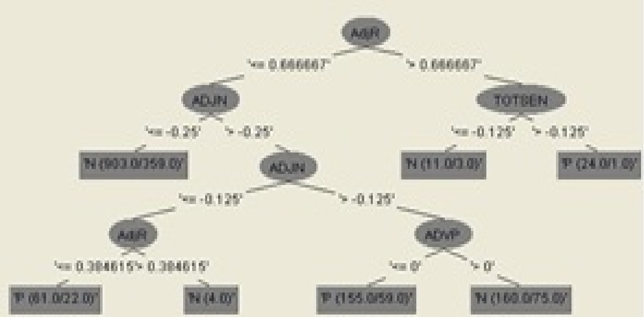

Figure 1. Rules generated by J48

In the J48 tree, the ratio of the adjectives to the total words in an opinion is used to generate the first classification rule (Figure 1). If the total adjective sentiment score is less than -0.25, the opinion is classified as negative (O). On the other hand, if the ratio of adjectives is greater than 0.666 and the total sentiment score of the opinion is greater than -0.125 then the opinion is positive (P). It is also observed that the adjectives play a key role in classification than the adverbs.

VI. Discussion

The paper investigated an approach to opining mining and sentiment analysis for less resourced languages using the resources in the English Language. The experiment was tested by constructing a Subjective lexicon for Sinhala language with the aid of English sentiment lexicon SentiWordNet 3.0. Then using this lexicon, a sample of Sinhala opinions were classified to see how the lexicon faired in classifying the opinions. Even though the accuracy of different classification methods were around 56-60%, the approach can be further optimized to improve the accuracy so far obtained in this initial investigation. In this work, negation of phrases that contain two or more words with negative meaning has not been considered. For example, the phrase like “වැරදි නෑ” meaning “not wrong”, gives total negative score if individual sentiment score of the two terms assigned to the weight vector. But, as multiword expression this is a positive expression. Handling such cases of negation to improve the methodology is being considered for future research. Some inaccuracies seen on the generated subjective lexicon scores may have been caused to classification accuracy. The word “ඉහළ” mapped to “above” with the negative score 0.125. It can be argued that this word can be of negative orientation in some context. But it would be a positively oriented word in most of the sentences.

VII. Conclusions

In this first ever attempt of a sentiment analysis in the Sinhala Language, we have achieved acceptable results maximum of 60% in Naïve base classification with a Sinhala sentiment lexicon developed using available resources such as the SentiWordNet (English). The bench mark accuracy level of

69% [15] achieved in similar work for the English Language can be achieved for Sinhala as well with the improvements being considered for the future as stated in the discussion section.

References

[1] B. Liu. (2010). *Hand book of Natural Language Processing*. CRC Press, Taylor and Francis Group.

[2] Ethnologue. (2014). *Statistical Summaries: Summary by world area*. Retrieved 04 05, 2014, from Ethnologue: Languages of the World: http://www.ethnologue.com/statistics

[3] A. Bakliwal, P. Arora and V. Vrma. (2012). Hindi Subjective Lexicon: A lexical Resource for Hindi Polarity Classification. *The eighth international conference on Language Resources and Evaluation (LREC). Hydeabad.*

[4] S. Huang, Z. Niu and C. shi. (2014). Automattic Construction of Domain Specific Sentiment Lexicon based on Constrained label propagation. *Knowledge-Based Systems, ELSEVIER, 56*, 191-200.

[5] A. Esuli and F. Sebastiani. (2006). SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining. *5th Conference on Language Resources and Evaluation (LREC'06)*, (pp. 417-422).

[6] F. Benamara, C. Cesarano, A. Picariello, D. Reforgiato and V. Subrahmanian. (2007). Sentiment Analysis: Adjectives and Adverbs are Better than Adjectives Alone. *International Conference on Weblogs and Social Media(ICWSM).* Boulder, Colorado. languages. *14th International Conference on Advances in ICT for Emerging Regions* (pp. 144-148). Colombo: IEEE.

[7] N. Medagoda, S. Shanmuganathan, J Whalley. (2013). A comparative analysis of opinion mining and sentiment classification in non-English M. Taboda, J. Brooke, M. Tofiloski, K. Voll, M. Stede. (2011). Lexicon-Based Methods for Sentiment Analysis. *Computational Linguistics, 37*, 267-307.

[8] R. Xia,C. Zong and S. Li. (2011). Ensemble of feature sets and classification algorithms for sentiment classification. *Information Sciences, 181*, 1138–1152.

[9] K. Vojislav. (2001). Learning and Soft Computing, Support Vector machines, Neural Networks and Fuzzy Logic Models. The MIT Press, Cambridge, MA.

[10] J. Kwok. (1998). Automated Text Categorization Using Support Vector Mechine. *International Conference on Neural Information Processing (ICONIP)*, (pp. 347-351).

[11] A. Rajput, R. P. (2011). J48 and JRIP Rules for E-Governance Data. *International Journal of Computer Science and Security (IJCSS), 5*, 201-207.

[12] N. Bhargava, G. Sharma, R. Bhargava. (2013). Decision Tree Analysis on J48 Algorithm for Data Mining. *International Journal of Advanced Research inComputer Science and Software Engineering, 3*, 1114-1119.

[13] Schütze, C. M. (1999). Text Categorization. In Foundations of Statistical Natural Language Processing. Cambridge. Cambridge: MIT Press.

[14] B. Ohana and B. Tierney. (1999). Sentiment classification of reviews using SentiwordNet. *9th IT & T Conference.* Dublin.