Building Machine Learning Based Senti-word Lexicon for Sentiment Analysis

Alaa Hamouda
Al_Azhar University / Department of Systems and Computers Engineering, Cairo, Egypt
Email:alaa_ham@gega.net

Mahmoud Marei and Mohamed Rohaim
Al_Azhar University / Department of Systems and Computers Engineering, Cairo, Egypt
Email :{ marie, m_rohaim}@azhar.edu.eg

***Abstract*— Sentiment analysis involves classifying opinions in text into categories like "positive" or "negative". One of approaches used to make sentiment classification is using sentiment lexicon. This paper aims to build a sentiment lexicon which is domain independent. We propose a Machine Learning Based Senti-word Lexicon (MLBSL) based on the Amazon data set which contains reviews from different domains. Our proposed MLBSL yields an improvement over previous published manual and automatic-built lexicons like SentiWordNet. We also provide an improvement in calculation method used in reviews sentiment analysis.**

***Index Terms*—Sentiment Analysis, Sentiment Lexicon, Machine Learning**

I. INTRODUCTION

Today, very large amounts of subjective text are available on the internet in the form of product reviews, blog posts and comments in discussion forums. Business analysts are turning their eyes on the internet in order to obtain factual as well as more subtle and subjective information (opinions) on companies and products. Opinion mining can assist in a number of potential applications in areas such as search engines, recommender systems and market research.

Classifying product reviews is a common problem in opinion mining and varieties of techniques have been used to address the problem. These techniques can be classified into two main approaches, approaches based on lexical resources and neutral language processing and approaches employing machine learning algorithms.

Machine learning either supervised or unsupervised methods using different aspects of text as sources of features have been proposed in the literature. Early work seen in [1] presents several supervised learning algorithms using bag-of-words features common in text mining research, with best performance obtained using support vector machines in combination with unigrams. Classifying terms from a review into its grammatical roles, or parts of speech has also been explored. In [2] part of speech information is used as part of a feature set for performing sentiment classification on a data set of newswire articles, with similar approaches attempted in [3], on different data sets. In [4] part of speech, words string and root information are used with various combinations for performing classification on various data sets of consumer reviews. Other studies used lexical resources like SentiWordNet to build a data set of features derived from its scores to be used as features for support vector machines classifier as done in [5] and [6].

Opinion lexicons are resources that associate sentiment polarity for words. Their use in opinion mining research stems from the hypothesis that individual words can be considered as a unit of opinion information, and therefore may provide clues to review sentiment and subjectivity. In [5] SentiWordNet lexicon was applied by counting positive and negative terms found in a review and the sentiment polarity was determined based on which class received the highest score. The below section describes the SentiWordNet lexicon and how it was used in the previous work.

In this paper we proposed a Machine Learning Based Senti-word Lexicon based on the Amazon data set which contains reviews from different domains. The remainder of the paper is organized as follows: Section 2 presents the previous work of classifying products reviews. Section 3 describes the required steps to build the proposed MLBSL, while section 4 shows the results of using MLBSL. Section 5 provides several improvements for the MLBSL lexicon. Finally, the conclusion is presented in Section 6.

II. RELATED WORK

There are several approaches for detecting sentiment in text present in literature. One of the most important is to use Opinion lexicons. Opinion lexicons are resources that associate sentiment orientation and words. Their use in opinion mining research stems from the hypothesis that

doi:10.4304/jait.2.4.199-203

individual words can be considered as a unit of opinion information, and therefore may provide clues to document sentiment and subjectivity [5]. Manually created opinion lexicons were applied to sentiment classification as seen in [1], where a prediction of document polarity is given by counting positive and negative terms. A similar approach is presented in [7], this uses an opinion lexicon based on the combination of other existing resources.

Building manual lists is a time consuming effort, and may be subject to annotator bias. To overcome these issues lexical induction approaches have been proposed in the literature with a view to extend the size of opinion lexicons from a core set of seed terms, either by exploring term relationships, or by evaluating similarities in document corpora [5]. In [8], it takes the construction of domain-oriented sentiment lexicon as clustering of sentiment words and extends the information-bottleneck clustering algorithm [9] by integrating more restriction for building an appropriate knowledge context of every sentiment word. While [10] proposed an expansion for domain sentiment lexicon using a novel propagation approach that exploits the relations between sentiment words and topics or product features that the sentiment words modify, and also sentiment words and product features themselves to extract new sentiment words.

One of popular sentiment lexicon is SentiWordNet[11] which contains opinion information on terms extracted from the WordNet database - by using a semi supervised learning method - and made publicly available for research purposes. SentiWordNet provides a readily available database of term sentiment information for the English language, and could be used as a replacement to the process of manually deriving opinion lexicons.

III. Porposed Sentiment Word Lexicon

We propose Machine Learning Based Senti-word Lexicon (MLBSL) which is based on the bag of words - generated from applying Support Vector Machine (SVM) to learn the significant senti-word- as a sentiment word lexicon. Our technique contains the following two phases to generate lexicon as shown in Fig. 1.

A. Data Preparation Phase

We used "Amazon Product Review Data Set" which is available at http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html. This data set has a round 5,000,000 reviews for various products (books, cameras, mp3s, etc.), which means that it has reviews from different domains. Each review consists of both a textual comment and a numeric rating score which takes a value between 1 and 5. The following data preparation steps are performed on this dataset before it is used.

1. Convert amazon data set to binary classes (positive and negative) by considering the reviews with rates (1, 2) as negative reviews and the reviews with rates (4, 5) as positive reviews and neglect the reviews with rate 3.
2. From the converted data set we choose 25,000 reviews randomly from 756,958 reviews, where numbers of positive and negative reviews are equal and review size is not more than 500 characters, Normal Reviews.
3. Apply the tokenization process on these reviews to split their text into very simple tokens such as numbers, punctuation and words of different types.
4. Apply the morphological analysis process to produce the roots of tokens. Root is the lemmatized, lower-case form of the token (for example, run is the root feature for run, runs, ran, and Running).

We use the General Architecture for Text Engineering (GATE) tool, a framework for the development and deployment of language processing technology in large scale [12], to implement steps 3 and 4.

B. Lexicon Development Phase

We use this prepared data set to train the support vector machine (SVM), based on n-grams of simple linguistic features of the text it contains. The machine-learning techniques used to infer the values of words implicitly from the training data. SVM was applied with the following parameters:

- Linguistic Feature: used is the 'string' of tokens. The 'string' is the original, unmodified text of the token
- N-gram: creates subsequences of tokens which will be considered in training. In our technique we use a combination between unigram (subsequence size is one token) and bigram (subsequence size is two tokens) with weights 1:5 respectively. The bigram with high weight is considered in training to include compound phrases like 'very good' in the output bag-of-words and gives high priority for them over single words 'very' and 'good'.
- Term weight calculation method: is term frequency – inverse document frequency (tf-idf). It is a multiplication of term frequency and inverse document frequency. Where term frequency refers to the number of occurrences of one term in a document. This count is usually normalized to prevent a bias towards longer documents to give a measure of the importance of the term t_i within the particular document d_j. Thus we have the term frequency, defined as follows.

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \qquad (1)$$

where $n_{i,j}$ is the number of occurrences of the considered term (t_i) in document d_j, and the denominator is the sum of number of occurrences of all terms in document d_j, that is, the size of the document $| d_j |$.

The inverse document frequency is a measure of the general importance of the term which is defined as follows:

$$idf_i = log \frac{|D|}{|\{d_j : t_i \in d_j\}|} \qquad (2)$$

Where |D| is the total number of documents in the corpus, and $| \{d_j: t_i \in d_j\}|$ is the number of documents in which the term ti appears.

The generated model from this training will contains a bag-of-words which has many tokens, with positive or negative weight values. We considered these weights as sentiment scores for these tokens and each token will have positive or negative meaning according to its polarity value.

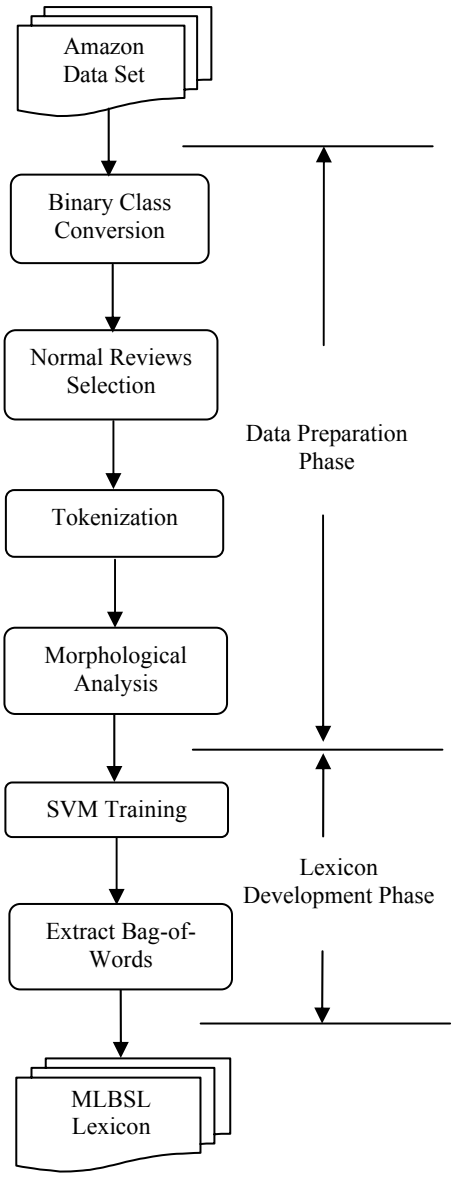

Figure 1. Creating Machine Learning Based Senti-word Lexicon Using Amazon Data set

IV. Using Machine Learning Based Senti-word Lexecion For Reviews Classification

From the bag-of-Words generated in the previous section we selected the highest tokens according to their absolute values and use them as a MLBSL. We applied this lexicon on two corpora. First corpus is the Amazon corpus, where we chose 4,000 reviews (2,000 positive and 2,000 negative). These reviews are selected to be different from training reviews to ensure realistic results. The second corpus is the movies corpus used in [1], [4], and [5]. This corpus contains 1,000 reviews positive and 1,000 reviews negative, and is available on-line at http://www.cs.cornell.edu/-people/pabo/movie-review-data.

We used 'Term Counting' calculation methods to apply MLBSL for reviews classification. In this method the MLBSL lexicon was applied by counting positive and negative words found in a review and determining sentiment polarity based on which class received the highest score. This method was applied with various sizes of MLBSL (lexicon size means number of tokens in it) and the results were as shown in table1. From these results we can see that MLBSL has a good performance (accuracy) when its size is between 15,000 and 20,000 tokens, while the highest performance is at 15,000 tokens.

TABLE 1: THE RESULTS OF AMAZON AND MOVIES REVIEWS CLASSIFICATION USING MLBSL BASED ON 'NORMAL REVIEWS' AND 'STRING' AS LINGUISTIC FEATURE

MLBSL Size (tokens)	Accuracy (%)		
	Amazon	Movies	Average
20,000	72.13	70.15	71.14
15,000	71.32	71.75	71.55
10,000	70.42	67.6	69.01
5,000	68.42	65	66.71

To compare our proposed MLBSL with the published lexicons, we used the same testing data set (Movies data set) used previously. The results showed that using MLBSL in reviews classification yielded a significant improvement over others as shown in table 2.

TABLE 2: COMPARING RESULTS

Method	Accuracy
MLBSL – Term Counting (this research)	71.75%
SentiWordNet – Term Counting [5].	65.85%
Term Counting - Manually built list of Positive/Negative words [1].	69.35%
Term counting from Combined Lexicon and valence shifters [4].	67.80%

V. Improving Machine Learning Based Senti-word Lexicon

In the previous section we use SVM to build MLBSL based on:

- Data Set : which is the 'Normal Reviews' extracted from Amazon data set (reviews with rates 1 and 2 are negative reviews and reviews with rates 4 and 5 are positive reviews)
- Feature: is the string of token, with n-gram is a combination between unigram and bigram.

We proposed several enhancements to improve the performance of using MLBSL lexicon. In each enhancement, the results are taken to investigate the performance improvement.

1. In the first enhancement, we used an alternative data set with strong meaning of positive and negative reviews. This data set is constructed by extracting reviews with rate 1 as negative reviews and reviews with rate 5 as positive reviews from Amazon data set, 'Strong Reviews'. Then the technique described in section 3 is applied to build improved MLBSL based on the following:

- Data Set: 25,000 'Strong Reviews' is extracted randomly from 1,260,499 reviews.
- Feature: the string of token, with n-gram as a combination between unigram and bigram with weight 1:5 respectively.

As shown in the previous section, the best performance for MLBSL was at a bag of words size of 15,000. So, we used a generated lexicon with 15,000 tokens to test the improvements of MLBSL. Applying this enhancement generates the results shown in table 3. It is evident that using strong data set in training produced small improvement with 0.33%.

TABLE 3: AMAZON AND MOVIES CLASSIFICATION USING MLBSL BASED ON STRONG REVIEWS

MLBSL Type / Calculation Method	Accuracy (%)		
	Amazon	Movies	Average
Based on Strong Reviews and String of Token / using Count Method	72.25	71.5	71.88

2. In the second enhancement, we used 'Term Score Summation' method, in which the summation of positive terms and the summation of negative terms in a review are calculated to get the positive and negative scores for all review terms. Then, the review sentiment is determined based on which score has the highest value. This method has an advantage over 'Term Counting' method that it takes into consideration the magnitude scores for tokens [13].

The results in table 4 show that using summation method in sentiment calculation produces significant improvement in the reviews classification performance with 5.87%. Actually, this improvement is not obtained from the MLBSL itself, but from how to use it in determining the sentiment polarity of the reviews.

TABLE 4: AMAZON AND MOVIES CLASSIFICATION USING MLBSL WITH SUMMATION METHOD

MLBSL Kind / Calculation Method	Accuracy (%)		
	Amazon	Movies	Average
Based on Strong Reviews and String of Token / using Summation Method	78.99	76.5	77.75

3. In the third enhancement, we used 'root' as a feature of tokens instead of 'string'. The 'root' is the lemmatized, lower-case form of the token. Then we applied the same technique described in section 3 to build improved MLBSL based on the following.

- Data Set: is 25,000 'Strong Reviews' selected randomly from 1,260,499 reviews with their sizes not more than 500 characters.
- Feature: is the root of token with n-gram, a combination between unigram and bigram with weight 1:5 respectively.

Table 5 shows the results of replacing the string feature with the root feature. 0.83% of increase in the accuracy is gained from this replacement.

TABLE 5: AMAZON AND MOVIES CLASSIFICATION USING MLBSL BASED ON ROOT FEATURE

MLBSL Kind / Calculation Method	Accuracy (%)		
	Amazon	Movies	Average
Based on Strong Reviews and Root of Token / using Summation Method	79.36	77.8	78.58

VI. Conclusion

Sentiment lexicons considered as one of valuable resources used for sentiment analysis. We propose a Machine Learning Senti-word Lexicon based on training a SVM using Amazon corpus which contains reviews from different domains. The training is based on the linguistic feature of text. Our proposed lexicon provides an improvement over previous published lexicon. We also provide an improvements for building this lexicon by using 'Strong Reviews' as the dataset and 'root' of tokens as the feature to be used by SVM. Another improvement

in reviews classification accuracy is provided by using 'Term Score Summation' as a calculation method. All these improvements yield an accuracy of 79.36 % and 77.8 % for Amazon and Movies corpora respectively.

REFERENCES

[1] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? Sentiment Classification using Machine Learning Techniques. Proceedings of EMNLP, 2002.
[2] T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. Proceedings of HLT/EMNLP, Vancouver, Canada, 2005.
[3] F. Salvetti, S. Lewis and C. Reichenbach. Automatic Opinion Polarity Classification of Movie Reviews. Colorado Research in Linguistics. Volume 17, Issue 1 (June 2004). Boulder: University of Colorado.
[4] A. Kennedy and D. Inkpen. Sentiment Classification of Movie Reviews Using Contextual Valence Shifters. Computational Intelligence, Vol. 22, 110–125, 2006.
[5] B. Ohana and B. Tierney. Sentiment Classification of Reviews Using SentiWordNet. 9th. IT&T Conference, Dublin Institute of Technology, Dublin, Ireland, 22nd.-23rd. October, 2009.
[6] H. Saggionα and A. Funk. Interpreting SentiWordNet for Opinion Classification. Proceedings of the Seventh conference on International Language Resources and Evaluation LREC10, 2010.
[7] A. Funk, Y. Li and H. Saggi, K. Bontchevaon and C. Leibold. Opinion Analysis for Business Intelligence Applications. Proceedings of the first international workshop on Ontology-supported business intelligence, 2008.
[8] W. Du and S. Tan. Building Domain-oriented Sentiment Lexicon by Improved, Information Bottleneck. ACM 978-1-60558-512-3/09/11, 2009
[9] N. Slonim and N. Tishby. Agglomerative information bottleneck. NIPS 1999.
[10] G. Qiu, B. Liu, J. Bu and C. Chen. Expanding Domain Sentiment Lexicon through Double Propagation. IJCAI'09 Proceedings of the 21st international jont conference on Artifical intelligence, 2009.
[11] A. Esuli and F. Sebastiani, SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining. Proceedings from International Conference on Language Resources and Evaluation (LREC), Genoa, 2006.
[12] Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02), 2002.
[13] A. Hamouda and M. Rohaim. Reviews Classification Using SentiWordNet Lexicon. World Congress on Computer Science and Information Technology, January, 2011.