International Journal of Computational Intelligence and Informatics, Vol. 5: No. 2, September 2015

Analysis of Classification Techniques for Mining Reviews Using Lexicon and WordNet Using R

Sharmista A
Research Scholar, Dept. of Computer Applications,
Madurai Kamaraj University
Tamil Nadu, India
ssharmistasaravananpkn@gmail.com

Ramaswami M
Associate Professor, Dept. of Computer Applications,
Madurai Kamaraj University
Tamil Nadu, India
mrswami123@gmail.com

***Abstract*- with the exponential growth of social media i.e. blogs and social networks, organizations and individual persons are increasingly using the number of reviews of these media for decision making about a product or service. Opinion mining detects whether the emotions of an opinion expressed by a user on Web platforms in natural language, is positive or negative. This paper presents extensive experiments to study the effectiveness of the classification of English type opinions in three categories: positive, negative and none. For this study, technological products corpora have been used. Furthermore, we have conducted a comparative assessment of the analysis of two classification techniques: J48 and C50 using the effect of both Opinion Lexicon and WordNet. Experimental results shows that the WordNet based sentiment classification perform well over Opinion Lexicon based classification. The proposed technique can also be used with any other language. The whole work is implemented using 'R' language.**

Keywords - Opinion Mining, Sentiment Analysis, Feature Extraction, Classification Algorithms, Opinion Lexicon and WordNet.

I. INTRODUCTION

The dramatic spread of the Internet in society has substantially changed the forms of communication, entertainment, knowledge acquisition and consumption. There is a constant increase in the number of people who consider the Internet as a medium for answering their queries [6], in addition to using it as a powerful means of communication. Indeed, on the one hand, the reviews expressed in forums, blogs and social networks are having greater importance to make a decision to buy a product, hire a service, and vote for a political party, among others. On the other hand, for providers, this information is also important to get some feedback about their clients' expectations and needs, clients' feelings about their products or services and then to improve them. However, the number of reviews has increased exponentially on the Web, therefore reading all the opinions are impossible for the users. On these grounds, different technologies to automatically process these reviews have lately arisen. These technologies are usually known as opinion mining.

With the advent of web 2.0, several types of social media sites such as blogs, discussion forums, review websites community websites and online shopping sites have emerged that have proved to be useful in determining the public; sentiment and opinion, towards the particular aspects of the products.

Thus with the growing use of Internet, use of social media sites have been increased to a much larger extent. Even these days' online shopping websites have a much larger gained speed as people mostly prefer these sites for shopping. There are various merchants offering millions of products online. For example, 5 millions of products have been indexed by Bing Shopping. Amazon.com archives a total of more than 36 million products. Shopper.com records more than five million products from over 3,000 merchants [8]. In this work we have collected the product reviews from Amazon.com.

Machine learning approaches are both supervised as well as unsupervised. But, opinion mining is basically a Supervised Approach where we need to train a classifier on the training set before it is to be applied on a test set. It combines the techniques of natural language processing, information retrieval, text analytics and computational linguistics.

Sentiment analysis or opinion mining is a type of subjectivity analysis, which aims at identifying opinions, emotions and evaluations expressed in natural language. The main goal is to predict the sentiment orientation (i.e. positive, negative or none) of an evaluation by analyzing sentiment or opinion words and expressions in sentences and documents. Three fundamental problems have to be solved which require at least linguistic

(lexical and syntactical) language analysis, or a richer and formal text characterization: aspect detection, opinion word detection and sentiment orientation identification [7]. The opinion mining task can be transformed into a classification task, so different supervised classification algorithms such as Decision Trees, Bayes Networks and Support Vector Machines (SVM), can be used to solve this task.

Opinion mining is carried out at document level, sentence level and aspect level mining based on the kind of information that is to be extracted from the opinionated text. In this work we have analyzed the sentence level reviews in order to predict the sentiment orientation.

II. LITERATURE REVIEW

Opinion Mining is the task of extracting the opinion expressed by the source on some target in a given set of sentences. In this paper opinion mining appears as a process of identifying and extracting a list of product features, and aggregating opinions about each of them from review sentences. Research on opinion mining started with identifying opinion bearing words, like great, amazing, wonderful, bad, poor etc. Many researchers have worked on mining such keywords and identifying their semantic orientations. In [8], a bootstrapping approach is proposed, which uses a small set of given seed opinion words to find their synonyms and antonyms in WordNet. A sizeable number of papers mentioning sentiment analysis focus on the specific application of classifying customer reviews as to their polarity – positive or negative [10, 7]. Researchers have also studied classification at the sentence-level, i.e., classifying each sentence as a subjective or objective sentence and/or as expressing a positive or negative opinion [7].

Many times the opinion holder writes the opinion without using proper English sentences but the inner meaning of the opinion is quite clear because users writes the adjective/adverb followed by the features which are usually nouns. To obtain detailed aspects, a feature based opinion mining is proposed in literature [8]. In [6] a supervised pattern mining method is proposed. The lexicon based approach [7, 8] uses opinion words and phrases in a sentence to determine the orientation of an opinion on a feature.

III. THE R ENVIRONMENT AND CLASSIFICATION TECHNIQUES J48 AND C50

R is a system for statistical computation and graphics. It provides, among other things, a programming language, high level graphics, interfaces to other languages and debugging facilities. R is an integrated suite of software facilities for data manipulation, calculation and graphical display. Among other things it has an effective data handling and storage facility, a suite of operators for calculations on arrays, in particular matrices, a large, coherent, integrated collection of intermediate tools for data analysis, graphical facilities for data analysis and display either directly at the computer or on hardcopy, and a well-developed, simple and effective programming language (called 'S') which includes conditionals, loops, user defined recursive functions and input and output facilities. (Indeed most of the system supplied functions are themselves written in the S language.). The term "environment" is intended to characterize it as a fully planned and coherent system, rather than an incremental accretion of very specific and inflexible tools, as is frequently the case with other data analysis software. With the advent of web 2.0, several types of social media sites such as blogs, discussion forums, review websites community websites and online shopping sites have emerged that have proved to be useful in determining the public; sentiment and opinion, towards the particular aspects of the products.

R is very much a vehicle for newly developing methods of interactive data analysis [1]. It has developed rapidly, and has been extended by a large collection of packages. However, most programs written in R are essentially ephemeral, written for a single piece of data analysis.

My proposed solution deals with the two main tasks of opinion mining: the pre-processing steps, applying decision tree classification techniques like – j48 and c50 is applied using R tool. Sentiment and the evaluation of the opinion are predicted and user can easily predict which product is best and buy according to their wish. The machine learning approach applicable to sentiment analysis mostly belongs to supervised classification in general and text classification techniques in particular. A number of machine learning techniques [2] have been adopted to classify the reviews. In this work, we compare decision tree based classifier models for the product review data sets obtained from Opinion Lexicon and WordNet.

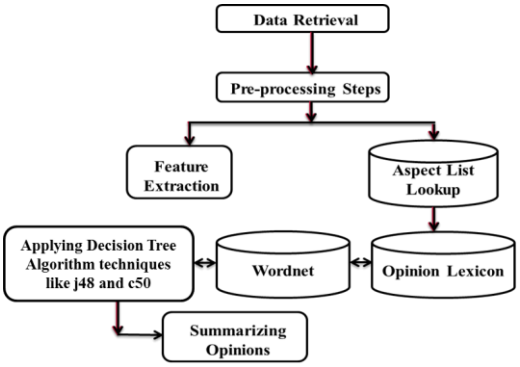

Figure 1. Framework for Opinion Mining

Opinion Lexicon is one of the most important factors determining the orientation of opinions is the opinion words that opinion holders use to express their opinions. Different entities may be modified by different opinion words. We can use their association information with entities (both objects and attributes) to identify their co-references. Most works use the prior polarity [4] of words and phrases for sentiment classification at sentence and document levels. Thus, the manual or semi-automatic construction of semantic orientation word lexicon is popular. It is used to match the positive and negative words in the review with the lexicon. It acts as a dictionary corpus.

WordNet is a lexical database which is available online, and provides a large repository of English lexical items. WordNet was designed to establish the connections between four types of Parts of Speech (POS) - noun, verb, adjective, and adverb. The smallest unit in a WordNet is synset, which represents a specific meaning of a word. It includes the word, its explanation, and its synonyms. The specific meaning of one word under one type of POS is called a sense. Each sense of a word is in a different synsets. Synsets are equivalent to senses = structures containing sets of terms with synonymous meanings. Each synsets has a gloss that defines the concept it represents. Synsets are connected to one another through explicit semantic relations. For one word and one type of POS, if there is more than one sense, WordNet organizes them in the order of the most frequently used to the least frequently used (Semcor). In our work we have computed both positive and negative sentiment strengths for each word and their relative opinion was compared for each product. Based on WordNet, we measured the semantic orientation [5] of words. We collected words and all their synonyms in WordNet, i.e. words of the same synset.

The WordNet package provides an R via Java interface to the WordNet lexical database of English which is commonly used in linguistics and text mining. Internally WordNet uses Jawbone2, a Java API to WordNet, to access the database. Thus, this package needs both a working Java installation, activated Java under R support, and a working WordNet installation. Thus by using WordNet matching with positive and negative words in the reviews gives more accuracy than lexicon because this dictionary corpus matches with all the synonyms for each word found in the customer reviews.

J48 or C4.5 is an algorithm used to generate a decision tree developed by Ross Quinlan. C4.5 is an extension of Quinlan's earlier ID3 algorithm. The decision trees generated by C4.5 can be used for classification, and for this reason, C4.5 is often referred to as a statistical classifier.

Algorithm: C4.5 builds decision trees from a set of training data in the same way as ID3, using the concept of information entropy. The training data is a set $S = s_1, s_2, \cdots$ of already classified samples. Each sample s_i consists of a p-dimensional vector $(x_{1,i}, x_{2,i}, ..., x_{p,i})$, where they x_j represent attributes or features of the sample, as well as the class in which s_i falls.

At each node of the tree, C4.5 chooses the attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. The splitting criterion is the normalized information gain (difference in entropy). The attribute with the highest normalized information gain is chosen to make the decision. The C4.5 algorithm then recurs on the smaller sub lists.

This algorithm has a few base cases.

- All the samples in the list belong to the same class. When this happens, it simply creates a leaf node for the decision tree saying to choose that class.
- None of the features provide any information gain. In this case, C4.5 creates a decision node higher up the tree using the expected value of the class.
- Instance of previously-unseen class encountered. Again, C4.5 creates a decision node higher up the tree using the expected value.

C5.0 algorithm is an extension of C4.5 algorithm which is also extension of ID3. It is the classification algorithm which applies in big data set. It is better than C4.5 on the speed, memory and the efficiency. C5.0 model works by splitting the sample based on the field that provides the maximum information gain. The C5.0 model can split samples on basis of the biggest information gain field. The sample subset that is get from the former split will be split afterward. The process will continue until the sample subset cannot be split and is usually according to another field. Finally, examine the lowest level split, those sample subsets that don't have remarkable contribution to the model will be rejected. C5.0 is easily handled the multi value attribute and missing attribute from data set [3].

Rule sets: An important feature of is to generate classifiers called rule sets that consist of unordered collections of (relatively) simple if-then rules. The Rule sets option causes classifiers to be expressed as rule sets rather than decision trees.

Each rule consists of:

- A rule number -- this is quite arbitrary and serves only to identify the rule.
- Statistics (*n*, lift *x*) or (*n*/*m*, lift *x*) that summarize the performance of the rule. Similarly to a leaf, *n* is the number of training cases covered by the rule and *m*, if it appears, shows how many of them do not belong to the class predicted by the rule. The rule's accuracy is estimated by the Laplace ratio *(n-m+1)/(n+2)*. The lift *x* is the result of dividing the rule's estimated accuracy by the relative frequency of the predicted class in the training set.
- One or more conditions that must all be satisfied if the rule are to be applicable.
- A class predicted by the rule.
- A value between 0 and 1 that indicates the confidence with which this prediction is made.

When a rule set like this is used to classify a case, it may happen that several of the rules are applicable (that is, all their conditions are satisfied). If the applicable rules predict different classes, there is an implicit conflict that could be resolved in several ways: for instance, we could believe the rule with the highest confidence, or we could attempt to aggregate the rules' predictions to reach a verdict. Rule sets are generally easier to understand than trees since each rule describes a specific context associated with a class. Furthermore, a rule set generated from a tree usually has fewer rules than the tree has leaves, another plus for comprehensibility. Another advantage of rule set classifiers is that they are often more accurate predictors than decision trees, since the rule set has an error rate of 0.5% on the test cases. For very large datasets, however, generating rules with the Rule set option can require considerably more computer time.

IV. DATA SOURCE

User's opinion is a major criterion for the improvement of the quality of services rendered and enhancement of the deliverables. Review sites, blogs and micro blogs provide a good understanding of the reception level of the products and services. The opinion of others played a fruitful role for the users in making a purchase decision. A large and growing body of user-generated reviews is available on the Internet. The reviews for products or services are usually based on opinions expressed in much unstructured format. The reviewer's data used in most of the sentiment classification studies are collected from the e-tailer websites like amazon (product reviews), CNET download (product reviews) and review Centers etc., which hosts millions of product reviews of consumers. In this work, the product reviews are collected from e-tailer amazon.com.

V. RESULT AND DISCUSSION

R is a sophisticated statistical software package, which provides new approaches to data mining. We have analyzed an effect of product review dataset obtained from both Opinion Lexicon and WordNet. The algorithm is executed to predict the best product by identifying the total number of positive opinions.

Table I. shows results of recursive decision tree with opinion lexicon for product reviews dataset run on R platform. It depicts the positive, negative and none reviews which are classified based on the number of instances 594. It is predicted that there are 97 negative reviews, 70 none classified reviews with 7 error rate and 410 positive reviews in case when matched with opinion lexicon.

TABLE I. CONFUSION MATRIX OF DECISION TREE J48 WITH OPINION LEXICON

		Predicted Class		
		Negative	None	Positive
Actual Class	Negative	**97**	3	1
	None	3	**70**	4
	Positive	3	3	**410**

Table II. Shows evaluation with 594 training cases (using Opinion Lexicon), it is predicted that there are 99 negative reviews, 72 none classified reviews with 5 error rate and 412 positive reviews in case when matched with opinion lexicon. It is predicted faster than J48.

TABLE II. CONFUSION MATRIX OF DECISION TREE C50 WITH OPINION LEXICON

		Predicted Class		
		Negative	None	Positive
Actual Class	Negative	**99**	1	1
	None	3	**72**	2
	Positive	3	1	**412**

Table III. shows evaluation with 594 training cases (using WordNet) , it is predicted that there are 99 negative reviews with 2 error rate , 75 none classified reviews with 2 error rate and 413 positive reviews in case when matched with WordNet.

TABLE III. CONFUSION MATRIX OF DECISION TREE J48 WITH WORDNET

		Predicted Class		
		Negative	None	Positive
Actual Class	Negative	**99**	1	1
	None	0	**75**	2
	Positive	2	1	**413**

Table IV. Shows evaluation with 594 training cases (using WordNet), it is predicted that there are 106 negative reviews, 70 none classified reviews and 415 positive reviews in case when matched with WordNet.

TABLE IV. CONFUSION MATRIX OF DECISION TREE C50 WITH WORDNET

		Predicted Class		
		Negative	None	Positive
Actual Class	Negative	**106**	0	1
	None	0	**70**	1
	Positive	0	1	**415**

One advantage of using decision tree is that it is easy to generate rules and easily understandable. On comparing predictive performance of decision tree techniques, decision tree j48 and c50 with WordNet gives

higher predictive accuracy than j48 and c50 with Opinion Lexicon review dataset. This is because WordNet matches with multiple synonyms for a single feature extracted from product reviews.

VI. CONCLUSION

This paper presents a method of sentiment analysis, on the review made by users to products. The purpose of a domain sentiment word extraction approach based on the propagation of both sentiment lexicon and WordNet, using decision tree classification algorithm like j48 and c50 exploits dependency relations to capture the association between features. When compared to product reviews with opinion lexicon, the error rate is reduced while using the product reviews with WordNet. Improving the efficient classification of text with other techniques is left as a future work.

REFERENCES

[1] Alexandra Trilla, Francesc Alias (2013). "Sentence-Based Sentiment Analysis for Expressive Text-to-Speech", IEEE Transactions on Audio, Speech, and Language Processing, Vol. 21, No. 2, pp.223-233.

[2] Angulakshmi G, Dr.R.ManickaChezian, (2014). International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 7, July 2014 - An Analysis on Opinion Mining: Techniques and Tools.

[3] Balamurali A.R., Aditya Joshi, Pushpak Bhattacharyya, (2011) Robust Sense Based Sentiment Classification, ACL WASSA, Portland, USA.

[4] Dave K, Lawrence S, Pennock D. (2003). "Mining the peanut gallery: opinion extraction and semantic classification of product reviews". Proceedings of the 12th international conference on World Wide Web, ACM, New York, NY, USA.

[5] Esuli, A. & Sibastiani, F. (2006). SentiWordNet: A publicly available lexical resource for opinion mining. In Proceedings of the 5th International Conference on Language Resources and Evaluation, Itely, 417-422.

[6] García-Crespo A, Colomo-Palacios R, Gómez-Berbís JM, and Ruiz-Mezcua B. SEMO: a framework for customer social networks analysis based on semantics. Journal of Information Technology, 2010; 25(2): 178-188.

[7] S.Kim, and E.Hovy, "Determining the Sentiment of Opinions", in Proceedings of the 20th International Conference on Computational Linguistics (COLING'04), Switzerland, 2004, pp. 1367-1373.

[8] B.Pang, L.Lee, and S.Vaithyanathan, "Thumbs up? Sentiment Classification Using Machine Learning Techniques", in Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing(EMNLP'02), USA, 2002, pp. 79 –86.

[9] Thet TT, Cheon J, and Khoo C. Aspect-based sentiment analysis of movie reviews on discussion boards. Journal of Information Science, 2010; 36: 823-848.

[10] Zheng-Jun Zha, *Member, IEEE*, Jianxing Yu, Jinhui Tang, *Member, IEEE*, Meng Wang, *Member, IEEE*, and Tat-Seng Chua "Product Aspect Ranking and Its Applications" Ieee Transactions On Knowledge And Data Engineering, Vol. 26, No. 5, May 2014.