Aspect Based Sentiment Analysis using Support Vector Machine Classifier

Raisa Varghese
Dept. of Computer Science
Govt. Engineering College
Trichur, India
Email: raisa1905@gmail.com

Jayasree M
Dept. of Computer Science
Govt. Engineering College
Trichur, India
Email:mjayasree.arun@gmail.com

***Abstract*—Sentiment Analysis involves the process of identifying the polarity of opinionated texts. Lots of social networking sites are being used for expressing thoughts and opinions by users to rate products. These user opinionated text is highly unstructured in nature and thus involves the application of various natural language processing techniques. In aspect based sentiment analysis, the various features of a product is identified through the training process. For e.g. the aspects of a camera are picture quality, size, resolution etc. The quantitative analysis of each aspect is done using support vector machine classifier. In most of the previous works, a product review is analysed as a whole rather than considering each aspect of it. Aspect based opinion mining is tedious since the identification of individual features is in itself a challenging task.**

***Keywords*-Sentiment Analysis, Aspect Selection, Machine Learning, Natural Language Processing**

I. Introduction

A huge amount of user opinionated data are flooding in the internet now a days. The customers can review their products through social networking sites, forums, blogs etc. These opinionated data are of large amount and hence reading of each opinion and finding the essence of it is a tedious job for a reader. Therefore an automated approach is essential. Machine can be trained on analysing text pieces and to classify the text to have positive or negative polarity. This is very useful because a lot of user opinions can be considered within a very short amount of time which is impossible for a human to do[1].

E-shops play a very important role in the current product marketing and many customers started buying products online. They try to make decisions to buy a particular product by reviewing the previous consumer reviews. This decision-making process can be done automatically by a machine which has knowledge processing capabilities. This knowledge processing ability can be incorporated in a machine by using Natural Language Processing Techniques and Quantitative analysis methods. Sentiment Analysis is helpful for a customer who intend to buy a particular product while aspect based (or feature based) sentiment analysis in particular can be used both by manufacturers and customers of the products. In the proposed work, a sentiment analysis model which does aspect based sentiment analysis is accomplished.

In our work, sentiment analysis on different aspects of the product is done. We propose a different approach which combines the use of dependeny parsing, coreference resolution and Sentiwordnet together for the sentiment analysis. The training of the system is done using the support vector machine. The paper also presents an architecture of the analysis model that includes the description about the actual sequence of steps by which the work is carried out.

II. Related works

The research work is inspired from the recent innovations in the field of machine learning approaches. The sentiment analysis process can be accomplished in various levels as aspect based, word level, sentence level and document level. In most of the product reviews, customers express mixed opinions on the different aspects of the product eventhough they rate whole product as good or bad. Sentiment analysis done in the aspect level will be able to mine such mixed opinions and to obtain more helpful results. Hence aspect based sentiment analysis is the one preferred in this work.

In [2] the word level feature extraction is done using Naive Bayesian Classifier. The semantic orientation of the individual sentences is retrieved from the contextual information. This machine learning approach on average claims an accuracy rate of 83%. Another significant work is the implementation of both Natural Language understanding and Generation in Sentiment analysis [3]. A couple of algorithms to search and predict the orientation of opinions are specified in this research work. Summarization of text is also done as a subsystem. But this summarization work is truly dependent on the features and hence is far from the automatic summarization work in the field of NLP. The paper also describes the need of pronoun resolution in opinion mining even though it is not addressed. In our work, we address the problem of coreference resolution and it is done in the processing module of our model.

In [4], a method of sentiment analysis which does not use conventional natural language rules is specified. The work uses a machine learning approach (Naive Bayesian)for classification. The f-measure is used as metric for evaluation, and claims efficiency upto 70%. In their paper, the review sentences are divided into various classes according to the association rules. The classification of the opinionated text is done using both class association rules and naive Bayesian classifier.

In [5], the authors present an approach for opinion mining which relies on natural language processing techniques. The work is accomplished by the sentiment lexicon and a pattern database. The two feature selection algorithms discussed in this work are based on mixture model and the likelihood ratio. They propose a sentiment pattern based analysis for the sentiment classification work. In [6], an in-depth study of dependency relations among the words of a sentence is discussed. In their work, the dependencies are classified as short range and long range dependencies. They use a clustering approach after the parsing is done. In our work, the aspect-opinion identification subprocess in the processing module uses both the dependency parsing and natural language processing techniques. But we do not use an in-depth analysis of dependencies in our work as the former does.

III. The Proposed Method

Figure 1 gives an overview of the overall sentiment analysis system.

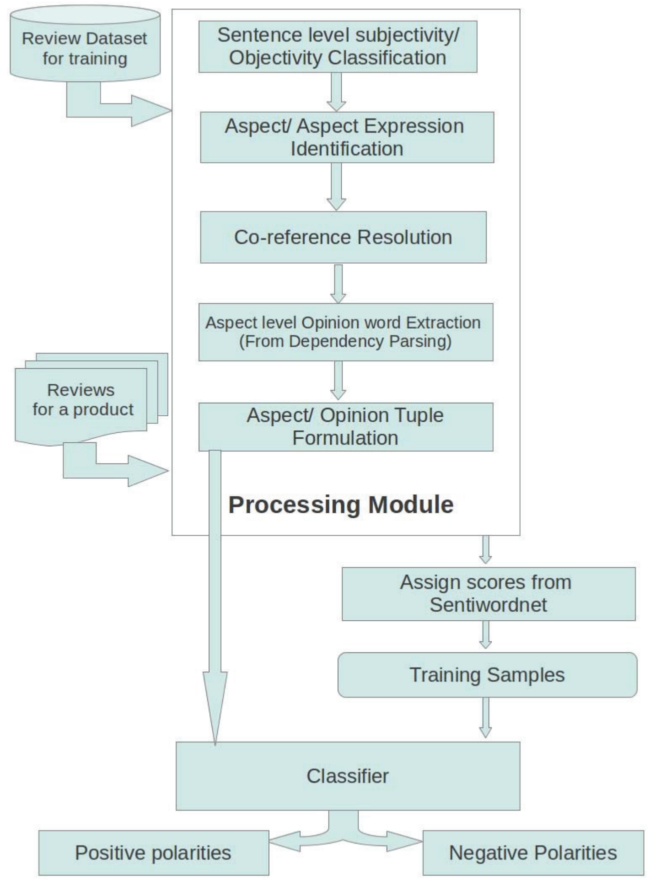

Figure 1. The Sentiment Analysis model

A. Dataset

The data set chosen for the work is a collection of reviews of digital cameras. A review set consisting of around 2500 user reviews from Amazon.in, ebay.in and other popular sites is used for training and feature identification purposes. The reviews of Canon A2300, Canon Powershot SX260, Sony DSC-RX100, Nikon D5100 and Canon A70 were used for training purpose. For testing, camera reviews from the customer review dataset[1] which is annotated with suitable polarity values are used. For mining the opinions about a particular product, the analysis system is designed to collect data from the web instantly. On an average, 200 reviews are collected to perform aspect based sentiment analysis on the reviews of one particular product.

B. Sentence Level Subjectivity/ Objectivity Classification

The proposed technique involves the subjectivity/ objectivity classification where the sentences in the review are categorized as useful or not. A review text is useful if it contains an opinion about the product being reviewed, else it is just a vague sentence that occurs during the normal human conversations. A question asked by a customer during the reviewing discussions may not contain any opinion about the product. These sentences (objective) must be found out and eliminated from the further steps to avoid unneccessary processing overhead. This classification can be done by using a Sentiwordnet which contains a collection of opinion words. Absence of any opinion words or any of the extracted features of the product can be used to help this categorization. Sometimes a customer simply asks a question in between a review sentence. The wh-determiner tag[2] in the Part-Of-Speech tagging can be used to find the wh-words like what, whom, when, where,

[1]http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#datasets

[2]http://nltk.org/api/nltk.tag.html

which etc in a sentence. This increases the probability that the sentence is a question. In our work, subjectivity/ objectivity classification is a part of pre-processing steps and this classification problem is addressed nominally.

C. Aspect/ Aspect Expression Identification

The aspects or features of a product can appear as a single word or a phrase. For example, picture quality of a camera is one among its aspects while size is another aspect. Therefore, simultaneous searching for proper nouns and such phrases is needed. This can be done by Part-Of-Speech Tagging and thus identifying the nouns and noun phrases , for example, words with tags NN (noun), NNS (noun plural), NNP (proper noun, singular)[2] etc. In most of the cases, the noun phrases in the sentences emerge to be the features of the particular product. Such phrases are identified during the training process and infrequent ones are discarded.

D. Coreference Resolution

In this work, we have incorporated coreference resolution to resolve all the anaphora that appears in the opinion sentences. For this purpose the Stanford Deterministic Coreference Resolution System [7], [8], [9] is used. In our system, the opinions about aspects are collected through a sentence level search. For instance, consider two short sentences that appeared in a review,

"Optical zoom of this camera is 5x. It is wonderful."

When the analysis of this review is done without coreference resolution, the opinion word in the second sentence could not be related to the aspect in first sentence. On the other hand, if coreference resolution is done this difficulty can be alleviated. The coreference resolver produces an output that "Optical zoom " in first sentence and "It" in second sentence corefer. So in the processing module, we replace the pronouns that got resolved, with the corresponding nouns. This replacement is limited to the pronouns that got resolved to aspect names of the product. Other coreference resolutions are avoided since they help little in the analysis work.

E. Opinion word Extraction for Aspects

The opinion word extraction for aspects is aided by the Stanford Dependencies Parser[3] [10], [11]. It helps to extract the grammatical relationships between words in a sentence. In our work, we use this dependency parser to find out the words that are related to the aspects of the product. In the Aspect Expression identification step, some important aspects of the product are found. Now the opinion sentences are fed to dependency parser and word dependencies are found out. In the dependency parsing certain relations are found to be more useful than the others. They were nsubj(nominal subject), amod(adjectival modifier), advmod(adverbial modifier), xcomp(clausal component with external subject) and negation modifer. But all direct dependencies and transitive ones (within a distance of one dependency relation) are used here.
For example, consider the sentence below,

"I am satisfied with this camera since it takes great pictures and is easy to use. "

The dependency parsing of the above sentence yields the folowing,

nsubjpass(satisfied-3, I-1), auxpass(satisfied-3, am-2), root(ROOT-0, satisfied-3), prep(satisfied-3, with-4), det(camera-6, this-5), pobj(with-4, camera-6), mark(takes-9, since-7), nsubj(takes-9, it-8), advcl(satisfied-3, takes-9), amod(pictures-11, great-10), dobj(takes-9, pictures-11), cc(takes-9, and-12), cop(easy-14, is-13), conj(takes-9, easy-14), aux(use-16, to-15), xcomp(easy-14, use-16)

Here in this sentence, two aspects and the camera as a whole are reviewed. From the dependency parsing, all the direct dependencies and transitive dependencies are considered. From the direct dependency, we get opinion words for the aspect 'picture' and 'use'. But the opinion about camera is extracted using transitive dependency from the dependencies, pobj(object of a preposition) and prep(prepositional modifier). In the similar manner, all the sentences are parsed and a collection of opinion words for each aspect in various sentences are collected. When negation is found as either direct relation or transitive relation, the opinion word will be appended with -1. This will be used for negation in the further steps.

F. Usage of Sentiwordnet

In most of the previous works, classification using only Sentiwordnet[12] has not been considered as a useful method to follow. The initial works done by us also substantiate this finding. Eventhough the whole task cannot be accomplished using Sentiwordnet alone, it is not necessary that we discard it completely. Therefore, throughout the sentiment analysis work, the classifier is also given access to the Sentiwordnet. The Sentiwordnet is particularly developed for opinion mining applications. For each synset in Wordnet[13], we could find in Sentiwordnet,

[3]Stanford Parser v1.6.4

three different polarties associated with them and they are positive, negative and objectivity. The score of the word 'small' in Sentiwordnet is -0.375. If in an opinion it appears that the battery life of a camera is small, it is definitely a negative opinion. But when the usage of the word 'small' is about size of the camera, it is normally positive one. Since our work is focussed on aspect based sentiment analysis, such a difference matters. So we assign different polarity scores to such words depending on the context in which they appear. This is done manually while training the classifier.

G. Support Vector machine classifier

Support vector machines (SVM) are primarily classifiers, that can classify by constructing hyperplanes that separate cases that belong to different categories. In our application, the SVM classifier should first be trained on a set of user reviews so that the machine gains knowledge for effective categorization. For the application, we have an iterative training algorithm to define a hyperplane that effectively separates two classes.

We adapt the SVM classifier to our terminologies as below:

Let X $\rightarrow$ Y be a mapping. Here X corresponds to the extracted opinion of the aspect of a product; Y can take real values and they are positive or negative scores. These scores are either from Sentiwordnet[12] or manually assigned as described in the previous step. The simplest case of classification is the binary classification. In addition to this, we can also include the third class which includes reviews with neutral polarity.

In the mapping X $\rightarrow$ Y is a single feature value pair where Y $\in$ **R**. Y can take negative and positive polarities. The training of the system can be done by using the feature set $x_1 : y_1, x_2 : y_2, ..., x_m : y_m$. After the training process, machine learns a classifier. The separation of the positive and negative opinions done by SVM classifier can be graphically represented as in figure 2.

In figure 2, various hyperplanes are used to separate the review texts into two categories. The SVM classifier must be trained to find out that hyperplane which considerably classifies samples. In our work, we used SVM^{light}[14] and the type of kernel used is linear. All the parameters of the classifier are set to their default values. The result of the classification is the polarity assignment to each feature of the product.

H. Aspect-Opinion tuple identifications

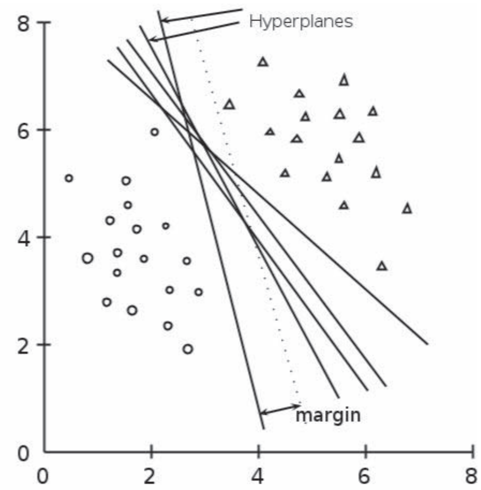

Figure 2. Hyperplanes and margin

The intermediate result of the work is a set of tuples that contains the aspect and the opinion itself. The classifier output is stored in the form of following tuple,

$$P\_orient = (r_i, a_{ij}, pol_{ij})$$

where,

r_i is the ith review, a_{ij} is the feature j in review i, pol_{ij} is the polarity(positive or negative) of jth feature in ith review.

I. Aggregation of Opinion Orientation

After the identification of polarities of aspect-opinion tuples, the polarities of the opinions in all the reviews can be aggregated together using the following formula,

For each aspect j of the product,

Orient_aggregate[j] = $\sum_i pol_{ij}$

The total polarity value for each feature is normalised using the total number of reviews considered.

Orient_norm[j] = $\frac{Orient\_aggregate[j]}{\sum_i}$

Orient_norm values are used to plot the resultant graph in Figure 3.

IV. Results

The result of the work is a quantitative analysis conveniently depicted using graphical plotting utilities. The result of the sentiment analysis involves the polarity orientations of each aspect of the product. In the figure 3, typical graphical plot for a product is shown. The Y-axis of the graph plots the polarity scores obtained by aggregating and normalising the polarity score output of SVM classifier. The aspect based sentiment analysis is particularly useful for the manufacturers of products to analyse the opinions about the individual aspects of a product.

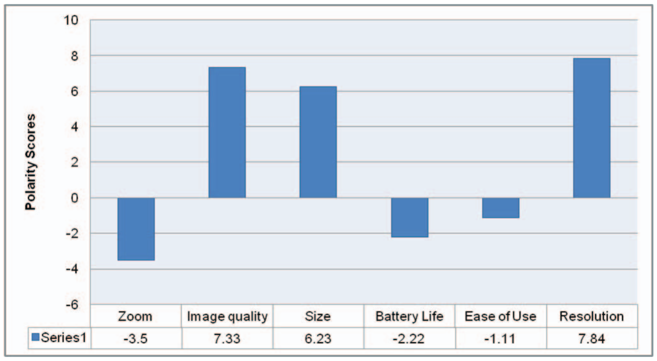

Figure 3. Graphical representation

V. Evaluation

For the evaluation of the system, we used the precision, recall and accuracy rates on a sentence basis. The test set included reviews of digital camera[1] which are already annotated using positive and negative polarities. The system which also uses coreference resolution produces results with an average accuracy rate of 77.98%. Table 1 gives the details of the evaluation results on the precision and recall metrics. The accuracy values are also depicted. Since we did this work in a single domain, tests are done only for reviews about digital cameras.

Table I
Evaluation Results

Product	Precision	Recall	Accuracy
Camera 1	85.94%	87.30%	78.48%
Camera 2	93.30%	77.38%	74.82%
Camera 3	91.30%	84.00%	80.65%
Average	90.18	82.89	77.98

As shown in the table, the method is promising. The novelty in the work lies in the use of combination of coreference resolver for pronoun resolutions and dependency parsing for relation extractions. We feed only 2-3 consequtive sentences to the resolver for accurate resolution. It is observed that coreferences which are far apart could not be resolved using the tool[7], [8], [9] we used. The system gets more data to be considered when the coreference resolution is done. The coreferences beyond 3 sentences are avoided here.

[1]http://www.cs.uic.edu/˜liub/FBS/sentiment-analysis.html#datasets

VI. Conclusion and Future Work

The sentiment analysis problem is a research area in which many hard problems are to be addressed. In the work done here, the task of finding the aspects or features of a particular product is done automatically. With sufficient supervised training, reliable results with an average accuracy rate of 77.98% could be achieved. The pre-processing steps are relativey not easy, because of the unstructured nature of natural language text. This is a serious issue to be considered and almost effective solutions can be provided to solve it. In our work, we used the coreference resolver, dependency parsing and Sentiwordnet together to produce the intended results.

In general, the customer reviews may contain statements that have opinions about aspects eventhough the aspect name may not be present in the sentence. Sentences with sarcasms may also appear. Such sentences need a second order interpretation by the reader apart from the mere meaning conveyed by the words in the sentence. Such cases occur in review texts, and it is extremely difficult for a machine to extract information from them. So in the current work done, such cases are not considered.

The next aim is to consider the comparitive sentences also in sentiment analysis work. In many reviews, the writers compare the product being reviewed with some other products. Opinions from such reviews can be mined with more knowledge and classification capabilities. Challenge like Named Entity Recognition should be considered for further improvement of the work.

References

[1] Bing Liu, *Sentiment Analysis and Subjectivity*, Handbook of Natural Language Processing, 2010.

[2] Khairullah Khan, Aurangzeb Khan, Bharum B,*Sentence Based Sentiment Classification from Online Customer Reviews*, ACM, 2010.

[3] Minqing Hu and Bing Liu, *Mining and Summarizing Customer Reviews*, Washington: ACM, 2004.

[4] C.C.Yang, Y.C. Wong, Chi-ping Wei, *Classifying Web Review Opinions for Consumer Product Analysis*, ICEC'09, Taiwan, ACM, 2009.

[5] Jeonghee Yi, T Nasukawa, Razvan B, Wayne Niblack, *Sentiment Analyzer: Extracting Sentiments about a given topic using Natural Language Processing Techniques*, ICDM'03, IEEE 2003.

[6] Subhabrata Mukherjee, Pushpak Bhattacharyya, *Feature Specific Sentiment Analysis for Product Reviews.*

[7] Heeyoung Lee, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu and Dan Jurafsky. *Deterministic coreference resolution based on entity-centric, precision-ranked rules*. Computational Linguistics 39(4), 2013.

[8] Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky. *Stanford's Multi-Pass Sieve Coreference Resolution System at the CoNLL-2011 Shared Task*. In Proceedings of the CoNLL-2011 Shared Task, 2011.

[9] Karthik Raghunathan, Heeyoung Lee, Sudarshan Rangarajan, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky, Christopher Manning *A Multi-Pass Sieve for Coreference Resolution* EMNLP-2010, Boston, USA. 2010.

[10] Marie-Catherine de Marneffe, Bill MacCartney and Christopher D. Manning. 2006. *Generating Typed Dependency Parses from Phrase Structure Parses*. In 5th International Conference on Language Resources and Evaluation (LREC 2006).

[11] Marie-Catherine de Marneffe and Christopher D. Manning. 2008. *The Stanford typed dependencies representation*. In COLING 2008 Workshop on Cross-framework and Cross-domain Parser Evaluation.

[12] Andrea Esuli and Fabrizio Sebastiani, *Sentiwordnet: A Publicly Available Lexical Resource for Opinion Mining*, in LREC 2006.

[13] George A. Miller, *WordNet: A Lexical Database for English*, Communications of ACM, November 1995/Vol. 38, No. 11.

[14] T. Joachims, *Making large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning, B. Schlkopf and C. Burges and A. Smola (ed.)*, MIT-Press, 1999.