International Journal of Engineering Research and Development
e-ISSN: 2278-067X, **p**-ISSN: 2278-800X, www.ijerd.com
Volume 7, Issue 6 (June 2013), PP. 16-19

A Corpus Approach for Opinion Mining to Improve the Performance Using Averaging

T.Janani[1], B.Subramani[2]

1 M.Phil Research Scholar, Dr. N.G.P Arts and Science College, Coimbatore,
2 Head of the Department (IT), Dr. N.G.P Arts and Science College, Coimbatore,

Abstract:- Opinion mining is one of the Natural Language Processing (NLP) which helps user to interact with the computer in user (i.e. natural) languages. The customer's reviews and opinion mining has become one of the wealthily areas in data mining. Nowadays as the enormous development of using web applications and sites provides a good platform for the customers to express their opinions directly on online shopping and company web sites like Cnet.com, Amazon.com etc., customers opinion becomes the helpful tools to manufacturers for assessment, finding satisfying proportion and limitation of the products. Recently many works are processed on this area of opinion mining, using different techniques. The urbanized techniques are good but still there are many challenges and obstacles found. In this paper, we collected opinions of various users from various review sites and constructed a corpus to perform classification and the challenges that face the opinion mining. This approach is tested on social networking reviews such as product reviews, movie reviews and MySpace comments. The classification approach can improve the effectiveness in terms of micro averaging and macro averaging.

Keywords:- Opinion mining, NLP, corpus, customer review, classification, data mining, social networks.

I. INTRODUCTION

The Internet contains important information on its user's opinions and the extraction of such unstructured web data is known as opinion mining and also sentiment analysis, a recent and volatile emerging research field widely employed by the industry for purposes such as marketing, customer service, and financial prediction. Mining opinions from natural language is an extremely difficult task which involves a deep understanding of most of the explicit and implicit information expressed by language structures [1], from single words to the entire document. The growth of the Social Web and the availability of a dynamic corpus of user-generated contents such as product review data makes essential to deal with the cognitive and affective information conveyed by expressive texts which reflects user responses.

The opinions found within comments, feedback and critiques provide useful indicators for many different purposes. These opinions can be categorized into three categories: positive, negative and neutral. For instance good, awesome, bad, disgusting, and satisfactory [2]. An opinion analysis task can be interpreted as a classification task where each category represents an opinion. Opinion analysis provides the level of product acceptance and to determine the strategies to improve product quality [3]. It also assists marketers or politicians to analyze public opinions with respect to public services or political issues. One important information need to be shared by many people, to find out opinions and perspectives on a particular topic.

II. USERS OPINION

Many works has recently focused on opinion mining of reviewers on social networks in order to get lunge on what people think about products and what are the features that they prefer with by using NLP [2]. Still opinion mining are opinionated and written as text and the available text mining systems are originally designed for regular kinds of texts of opinion.

A novel method may need to be adapted to deal with this type of text. The Natural Language Processing and its relevance's represent some useful tools for opinion mining and it also faces some difficulties in some aspects of documents, because each user takes up different style of opinion, thinking and way of writing. This paper will try to identify some of these aspects.

2.1 Customer Opinions

Each customer expresses their opinion on their own perspective, skill of writing, and thinking [5]. Some objective entities can be divided into the following categories.

2.1.1 **Direct opinion:** This type of opining is explicit if a feature or any of its synonyms appears in a sentence. This feature could be identified as explicit or direct opinion and they appear directly in a review. **E.g.:** "The accuracy of the iPod is slow".

2.1.2 Indirect opinion: This type of opining is implicit if a feature or any of its synonyms does not appear in a sentence. This feature could be identified as explicit or indirect opinion and they do not appear directly in review. **E.g.:** "My companion said that you lost your money by purchasing this iPod".

2.1.3 Comparative opinion: This type of opinion is done by comparing more than one entity. This kind of opinion is useful for the customers or reviewers to make a comparison of similar products.
E.g.: "Apple iPod is better than Samsung."

2.2 Opinion polarity and classification

All the customer comments and reviews about some products will be classified into polarity such as positive, negative or neutral. This is termed as opinion polarity [6]. Opinion can be classified into the following categories.

2.2.1 Document level: This level classifies a whole opinion document (a review) based on the overall sentiment of the opinion holder to check the polarity of the opinion.

2.2.2 Sentence level: This level classifies the whole document into sentence and determines the polarity of each sentence to detect the overall opinion polarity.

2.2.3 Dictionary based approach: This approach is based on the use of synonyms and antonyms in WordNet to determine opinions based on a set of propagating opinion [4]. The co-occurrence of vocabulary or phrases is developed in a corpus based approach.

III. CORPUS FOR OPINION MINING

A corpus is developed by three main steps: collection, annotation and analysis. In *Fig: 1.* the phases of a corpus are revealed. Each of them is strongly inclined by the others. The analysis and exploitation of a corpus can reveal limits of the annotation.

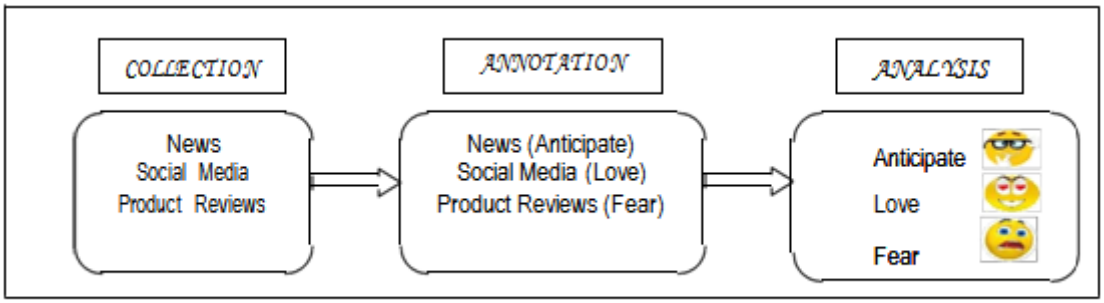

Figure 1: Phases of a corpus

3.1 Collection

The collection phase mainly refers to the selection of data and composition of the corpus (what), the choice of the data source (from where) and also to the collection methodologies applied (how). It is the task for which the resource is developed that usually drives the decisions about *what* data to collect and *from where* it should be collected. Most of the corpora designed are collected from web services [8]. Others are extracted from blogs and micro-blogs in order to provide insights about people's opinions and also about celebrities or politics.

3.2 Annotation

This annotation phase includes the explanation of a system and its application to the collected data but also the assessment effort of the material by the evaluation of inter-annotator agreement [8]. The design of the system is an to the perspective of data classification which makes theoretical assumptions to be annotated. It defines what kind of information to be annotated.
This is especially challenging because an agreed representation about these massively complex phenomena is missing. Modeling emotions and opinions can be done with three approaches the categorical, the dimensional, and the appraisal-based approach.

3.3 Analysis

The analysis phase is useful in training and testing for the classification of emotions and opinions. The results are strongly influenced by both the quantity and quality of data. Error detection and quality control techniques have been developed.

A strategy that can give very useful hints about the reliability of the annotated data is the comparison between the results of classification and human annotation. Labeling schemes are constructed by different uses of the annotated material. This motivates the efforts loyal to the definition and propagation of standards for the annotation of data for several NLP tasks [8]. By using these three phases the data collected is been developed into a corpus.

IV. CLASSIFICATION OF OPINION MINING

The goal of classification is to accurately predict the objective for each case in the data. In this approach a classification model is used to identify opinions as positive, negative or neutral [7]. Here the supervised learning mechanism is adopted and SVM (Support Vector Machines) classifier is used for classification and its one of the useful technique for data classification. The following procedure is used for classifying the polarity.

- Transform data into suitable format of an SVM package
- Conduct simple scaling on the data
- Extract the opinions expressed
- Extract the product features
- Extract relations between opinion expressed and product features with SVM.
- Train a SVM on data annotated with products features, opinion expressed and relations.

As a result a classified polarity of opinions is extracted. As a target lexicon and source of polarity information for our polarity-based concept similarity measure, WordNet is used. WordNet is a widely affective common sense resource for computing semantic web, and affective computing techniques to better identify, interpret, and process natural language opinions over the web [4]. It's a dictionary that assigns polarity values. The dataset consists of some wordlists of basic opinions for an overview of different sets of emotions proposed in the literature.

Review Sites	**Sample(S)**
Product Reviews	S_1: (1000+)
Movie Reviews	S_2: (1000+)
MySpace Comments	S_3: (1000+)

Table 1: The Data Sets

The data sets are collected from three different review sites as shown in *Table: 1* and sample S is chosen to develop a graph which is used for measuring the performance of the classifier. Based on the data set the following graph (*Fig: 2*) is constructed.

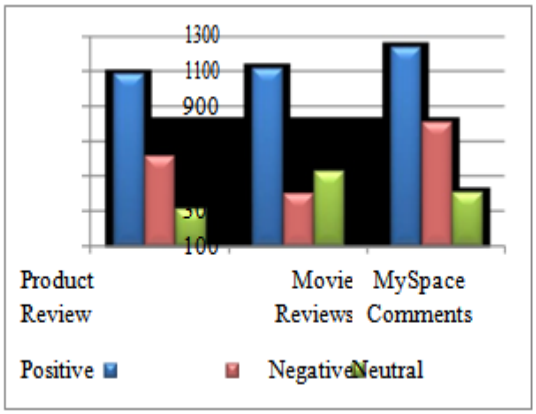

Figure 2: Polarity of review sites

V. MEASURING THE PERFORMANCE

It's the process of collecting and analyzing information regarding the performance of an individual or

group [9]. Measuring is one of the significant tasks to calculate the average performance of classifiers and it can be done by two different ways Micro averaging and Macro-averaging.

Micro averaging: A set of graph with polarity is given. Each part in the graph represents the sum of the number of documents extracted from review sites. In the graph, the average performance of a classifier in terms of its precision and recall is measured. Micro averaging treats each document equally. That is it results in averaging over a set of documents. The performance of a classifier is inclined to be dominated by common classes.

Macro averaging: Given a polarity based graph from which values are generated. Each value represents the precision or recall of an automatic classifier for each category. With these values, the average performance of a classifier in terms of its precision and recall is measured. In contrast macro averaging treats each class equally. The macro averaging results in averaging over a set of classes as a result.

VI. CONCLUSION

In this paper we examined the polarity classification showing that the subjectivity detection can compress reviews into much shorter extracts that still preserve polarity information at a level comparable to that of the full review. The opinion of people is gathered and a corpus is built for opinion mining and SVM classifier is used for classifying the data. The averaging methods are used to measure the performance of polarity. The macro averaged performance is lower than micro averaged performance. The use of classifiers can result in a better effectiveness in terms of micro averaged analysis than any individual classifier.

REFERENCES

[1]. Boyan Bonev, Gema Ramrez Sanchez, Sergio Ortiz Rojas, “Statistical sentiment analysis performance in Opinum”, in arXiv, 2013

[2]. Saifee Vohra, Jay Teraiya, “Applications and Challenges for Sentiment Analysis : A survey”, in IJERT, 2013

[3]. Erik Cambria, Yangqiu Song, Haixun Wang, Newton Howard, “Semantic Multi-Dimensional Scaling for Open-Domain Sentiment Analysis” in IEEE Intelligent Systems, 2012

[4]. J. Kamps, M. Marx, R. Mokken, and M.de Rijke, “Using WordNet to measure semantic orientation of adjectives,” in *LREC*, 2012

[5]. Xiaohui Yu, Yang Liu, Jimmy Xiangji Huang, Aijun An, “Mining Online Reviews for Predicting Sales Performance: A Case Study in the Movie Domain”, in IEEE transactions on knowledge and data engineering, 2012

[6]. E. Cambria, Y. Song, H. Wang, and Hussain, “Isanette: A common and common sense knowledge base for opinion mining,” in *ICDM*, 2011

[7]. Chenhao Tan, Lillian Lee, Jie Tang, “User-Level Sentiment Analysis Incorporating Social Networks”, in KDD, 2011

[8]. Janyce Wiebe, Ellen Riloff, “Finding Mutual Benefit between Subjectivity Analysis and Information Extraction”, in IEEE transactions on affective computing, 2011

[9]. Jie Tang, Jimeng Sun, Chi Wang and Zi Yang, “Social Influence Analysis in Large-scale Networks”, in KDD, 2010

[10]. L. Barbosa, J. Feng, “Robust sentiment detection on twitter from biased and noisy data”, in *COLING*, 2010