International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 03 Issue: 07 | July-2016 www.irjet.net p-ISSN: 2395-0072

Sentiment Analysis and Sentiment Classification using NLP

G.Divya, R.Suresh, Professor

G.Divya, PG Scholar, Dept. of CSE, CREC, Tirupati, AP, India
R.Suresh, Professor., Dept. of CSE, CREC, Tirupati, AP, India

ABSTRACT: The latest advancement in technology and growing data there is a need to extract information in a more efficient and quicker manner using queries. This gives rise to the need for a more easy-to-use query interface. So far, the typical query interfaces are GUI based visual query interfaces. Visual query interfaces however, have limitations especially when they are used for accessing large and complex datasets. The ease of expressing ones queries is limited due to language barrier and the knowledge of precise key words. Therefore, we are developing a novel query interface where users can use natural language expressions to help author visual queries and address the knowledge gap. The opinions and reviews posted in the websites are analyzed mostly based on static data. Till date, the methods applied to determine whether the opinions are positive or negative. But, now a days, as products are increasing day by day and a huge amount of data reviews are available online. So, this mechanism not considers the opinions at comprehensive level.

Keywords: Cross-Domains, Sentiment classification, Feature Expansion

I. INTRODUCTION

Sentiment analysis is a type of natural language processing for tracking the mood of the public about a particular product or topic. Sentiment analysis, which is also called opinion mining, involves in building a system to collect and examine opinions about the product made in blog posts, comments, reviews or tweets. Sentiment analysis can be useful in several ways. For example, in marketing it helps in judging the success of an ad campaign or new product launch, determine which versions of a product or service are popular and even identify which demographics like or dislike particular features.

There are several challenges in Sentiment analysis. The first is a opinion word that is considered to be positive in one situation may be considered negative in another situation. A second challenge is that people don't always express opinions in a same way. Most traditional text processing relies on the fact that small differences between two pieces of text don't change the meaning very much. In Sentiment analysis, however, "the picture was great" is very different from "the picture was not great". People can be contradictory in their statements. Most reviews will have both positive and negative comments, which is somewhat manageable by analyzing sentences one at a time. However, in the more informal medium like twitter or blogs, the more likely people are to combine different opinions in the same sentence which is easy for a human to understand, but more difficult for a computer to parse. Sometimes even other people have difficulty understanding what someone thought based on a short piece of text because it lacks context. For example, "That movie was as good as its last movie" is entirely dependent on what the person expressing the opinion thought of the previous model.

The user's hunger is on for and dependence upon online advice and recommendations the data reveals is merely one reason behind the emerge of interest in new systems that deal directly with opinions as a first-class object. Sentiment analysis concentrates on attitudes, whereas traditional text mining focuses on the analysis of facts. There are few main fields of research predominate in Sentiment analysis: sentiment classification, feature based Sentiment classification and opinion summarization. Sentiment classification deals with classifying entire documents according to the opinions towards certain objects. Feature-based Sentiment classification on the other hand considers the opinions on features of certain objects. Opinion summarization task is different from traditional text summarization because only the features of the product are mined on which the customers have expressed their opinions. Opinion summarization does not summarize the reviews by selecting a subset or rewrite some of the original sentences from the reviews to capture the main points as in the classic text summarization.

Languages that have been studied mostly are English and in Chinese .Presently, there are very few researches conducted on sentiment classification for other languages like Arabic, Italian and Thai. This survey aims at focusing much of the work in English and a few from Chinese. The emergence of sentiment analysis dates back to late 1990's, but becomes a major emerging sub field of information management discipline only from 2000, especially from 2004 onwards, which this survey focuses.

For the sake of convenience the remainder of this paper is organized as follows: Section 2 presents the data sources used for opinion mining. Section 3

introduces machine learning and semantic orientation approaches for sentiment classification. Section 4 presents some applications of sentiment classification. Then we present some tools available for sentiment classification in section 4. The fifth section is about the performance evaluation done. Last section concludes our study and discusses some future directions for research.

II. DATA SOURCE

User's opinion is a major criterion for the improvement of the quality of services rendered and enhancement of the deliverables. Blogs, review sites, data and micro blogs provide a good understanding of the reception level of the products and services.

2.1. Blogs

With an increasing usage of the internet, blogging and blog pages are growing rapidly. Blog pages have become the most popular means to express one's personal opinions. Bloggers record the daily events in their lives and express their opinions, feelings, and emotions in a blog (Chau & Xu, 2007). Many of these blogs contain reviews on many products, issues, etc. Blogs are used as a source of opinion in many of the studies related to sentiment analysis (Martin, 2005; Murphy, 2006; Tang et al., 2009).

2.2. Review sites

For any user in making a purchasing decision, the opinions of others can be an important factor. 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-commerce websites like www.amazon.com (product reviews), www.yelp.com (restaurant reviews), www.CNET download.com (product reviews) and www.reviewcentre.com, which hosts millions of product reviews by consumers. Other than these the available are professional review sites such as www.dpreview.com , www.zdnet.com and consumer opinion sites on broad topics and products such as www .consumerreview.com, www.epinions.com,www.bizrate.com (Popescu& Etzioni ,2005 ; Hu,B.Liu ,2006 ; Qinliang Mia, 2009; Gamgaran Somprasertsi ,2010).

2.3. Data Set

Most of the work in the field uses movie reviews data for classification. Movie review data are available as dataset (http:// www.cs.cornell.edu/People/pabo/movie-review-data). Other dataset which is available online is multi-domain sentiment (MDS) dataset .(http://www.cs.jhu.edu/mdredze/datasets/sentiment). The MDS dataset contains four different types of product reviews extracted from Amazon.com including Books, DVDs, Electronics and Kitchen appliances, with 1000 positive and 1000 negative reviews for each domain. Another review data available is http://www.cs.uic.edu/liub/FBS/CustomerReviewData.zi p. This dataset consists of reviews of five electronics products downloaded from Amazon and Cnet (Hu and Liu ,2006; Konig & Brill ,2006 ; Long Sheng ,2011; Zhu Jian ,2010 ; Pang and Lee ,2004; Bai et al. ,2005; Kennedy and Inkpen ,2006; Zhou and Chaovalit ,2008; Yulan He 2010; Rudy Prabowo ,2009; Rui Xia ,2011).

2.4. Micro-blogging

Twitter is a popular micro blogging service where users create status messages called "tweets". These tweets sometimes express opinions about different topics. Twitter messages are also used as data source for classifying sentiment.

III. S E N T I M E N T CLASSIFICATION

Much research exists on sentiment analysis of user opinion data, which mainly judges the polarities of user reviews. In these studies, sentiment analysis is often conducted at one of the three levels: the document level, sentence level, or attribute level. In relation to sentiment analysis, the literature survey done indicates two types of techniques including machine learning and semantic orientation. In addition to that, the nature language processing techniques (NLP) is used in this area, especially in the document sentiment detection. Current- day sentiment detection is thus a discipline at the crossroads of NLP and Information retrieval, and as such it shares a number of characteristics with other tasks such as information extraction and text-mining, computational linguistics, psychology and predicative analysis.

3.1. Machine Learning

The machine learning approach applicable to sentiment analysis mostly belongs to supervised classification in general and text classification techniques in particular. Thus, it is called „„supervised learning". In a machine learning based classification, two sets of documents are required: training and a test set. A training set is used by an automatic classifier to learn the differentiating characteristics of documents, and a test set is used to validate the performance of the automatic classifier. A number of machine learning techniques have been adopted to classify the reviews. Machine learning techniques like Naive Bayes (NB), maximum entropy (ME), and support vector machines (SVM) have achieved great success in text

categorization. The other most well- known machine learning methods in the natural language processing area are K-Nearest neighborhood, ID3, C5, centroid classifier, winnow classifier, and the N-gram model. Naive Bayes is a simple but effective classification algorithm. The Naive Bayes algorithm is widely used algorithm for document classification (Melville et al., 2009; Rui Xia, 2011; Ziqiong, 2011; Songho tan, 2008 and Qiang Ye, 2009). The basic idea is to estimate the probabilities of categories given a test document by using the joint probabilities of words and categories. The naive part of such a model is the assumption of word independence. The simplicity of this assumption makes the computation of Naive Bayes classifier far more efficient.

Support vector machines (SVM), a discriminative classifier is considered the best text classification method (Rui Xia, 2011; Ziqiong, 2011; Songho tan, 2008 and Rudy Prabowo, 2009). The support vector machine is a statistical classification method proposed by Vapnik . Based on the structural risk minimization principle from the computational learning theory, SVM seeks a decision surface to separate the training data points into two classes and makes decisions based on the support vectors that are selected as the only effective elements in the training set. Multiple variants of SVM have been developed in which Multi class SVM is used for Sentiment classification (Kaiquan Xu, 2011).

The idea behind the centroid classification algorithm is extremely simple and straightforward (Songho tan, 2008). Initially the prototype vector or centroid vector for each training class is calculated, then the similarity between a testing document to all centroid is computed, finally based on these similarities, document is assigned to the class corresponding to the most similar centroid.

The K-nearest neighbor (KNN) is a typical example based classifier that does not build an explicit, declarative representation of the category, but relies on the category labels attached to the training documents similar to the test document. Given a test document d, the system finds the k nearest neighbors among training documents. The similarity score of each nearest neighbor document to the test document is used as the weight of the classes of the neighbor document (Songho tan, 2008). Winnow is a well-known online mistaken-driven method. It works by updating its weights in a sequence of trials. On each trial, it first makes a prediction for one document and then receives feedback; if a mistake is made, it updates its weight vector using the document.

During the training phase, with a collection of training data, this process is repeated several times by iterating on
the data (Songho tan, 2008). Besides these classifiers other classifiers like ID3 and C5 are also investigated (Rudy Prabowo, 2009).

Besides using these above said machine learning methods individually for sentiment classification, various
comparative studies have been done to find the best choice of machine learning method for sentiment classification. Songbo Tan (2008) presents an empirical study of sentiment categorization on Chinese documents. He investigated four feature selection methods (MI,IG, CHI and DF) and five learning methods (centroid classifier, K-nearest neighbor, winnow classifier, Naive Bayes and SVM) on a Chinese sentiment corpus. From the results he concludes that, IG performs the best for sentimental terms selection and SVM exhibits the best performance for sentiment classification. When applying SVM, naive Bayes and n-gram model to the destination reviews, Ye et al. (2009) found that SVM outperforms the other two classifiers.

Rudy Parabola (2009) described an extension by combining rule-based classification, supervised learning and machine learning into a new combined method. For each sample set, they carried out 10-fold cross validation. For each fold, the associated samples were divided into training and a test set. For each test sample, a hybrid classification is carried out, i.e., if one classifier fails to classify a document, the classifier passes the document onto the next classifier, until the document is classified or no other classifier exists. Given a training set, the Rule Based Classifier (RBC) used a Rule Generator to generate a set of rules and a set of antecedents to represent the test sample and used the rule set derived from the training set to classify the test sample. If the test sample was unclassified, the RBC passed the associated antecedents onto the Statistic Based Classifier (SBC), if the SBC could not classify the test sample; the SBC passed the associated antecedents onto the General Inquirer Based Classifier (GIBC), which used the 3672 simple rules to determine the consequents of the antecedents. The Support vector machine (SVM) was given a training set to classify the test sample if the three classifiers failed to classify the same.

An ensemble technique is one which combines the outputs of several base classification models to form an integrated output. Rui Xia (2011) used this approach and made a comparative study of the effectiveness of ensemble technique for sentiment classification by efficiently integrating different feature sets and classification algorithms to synthesize a more accurate classification procedure. In his work, two types of feature sets are designed for sentiment classification, namely the part-of-speech based feature sets and the word-relation based feature sets. Then, three text classification algorithms, namely naive Bayes, maximum entropy and support vector machines, are employed as base-classifiers for each of the feature sets to predict classification scores. Three types of ensemble methods, namely the fixed combination, weighted combination

and meta-classifier combination, are evaluated for three ensemble strategies namely ensemble of feature sets, ensemble of classification algorithms, and ensemble of both feature sets and classification algorithms.

In most of the comparative studies it is found that SVM outperforms other machine learning methods in sentiment classification. Ziqiong Zhang (2011) showed a contradiction in the performance of SVM. They focused their interest on written Cantonese, a written variety of Chinese. there are many other words that invert the polarity of an opinion expressed, such as valence shifters, connectives or modals. "I find the functionality of the new mobile less practical", is an example for valence shifter, "Perhaps it is a great phone, but I fail to see why", shows the effect of connectives. An example sentence using modal is, "In theory, the phone should have worked even under water". As can be seen from these examples, negation is a difficult yet important aspect of sentiment analysis.

Kennedy and Inkpen (2005) evaluate a negation model which is fairly identical to the one proposed by Polanyi and Zaenen (2004) in document-level polarity classification. A simple scope for negation is chosen. A polar expression is thought to be negated if the negation word immediately precedes it. Wilson et al. (2005) carry out more advanced negation modeling on expression- level polarity classification. The work uses supervised machine learning where negation modeling is mostly encoded as features using polar expressions. Jin-Cheon Na (2005), reported a study in automatically classifying documents as expressing positive or negative.He investigated the use of simple linguistic processing to address the problems of negation phrase.

In sentiment analysis, the most prominent work examining the impact of different scope models for negation is Jia et al. (2009). They proposed a scope detection method to handle negation using static delimiters, dynamic delimiters, and heuristic rules focused on polar expressions Static delimiters are unambiguous words, such as because or unless marking the beginning of another clause. Dynamic delimiters are, however, rules, using contextual information such as their pertaining part-of-speech tag. These delimiters suitably account for various complex sentence types so that only the clause containing the negation is considered. The heuristic rules focus on cases in which polar expressions in specific syntactic configurations are directly preceded by negation words which results in the polar expression becoming a delimiter itself.

3.4.Feature based sentiment classification

Due to the increasing amount of opinions and reviews on the internet, Sentiment analysis has become a hot topic in data mining, in which extracting opinion features is a key step. Sentiment analysis at both the document level and sentence level has been too coarse to determine precisely what users like or dislike. In order to address this problem, sentiment analysis at the attribute level is aimed at extracting opinions on products' specific attributes from reviews.

Hu's work in (Hu, 2005) can be considered as the pioneer work on feature-based opinion summarization. Their feature extraction algorithm is based on heuristics that depend on feature terms‟ respective occurrence counts. They use association rule mining based on the Apriori algorithm to extract frequent item sets as explicit product features. Popescu et al (2005) developed an unsupervised information extraction system called OPINE, which extracted product features and opinions from reviews. OPINE first extracts noun phrases from reviews and retains those with frequency greater than an experimentally set threshold and then assesses those by OPINE‟s feature assessor for extracting explicit features. The assessor evaluates a noun phrase by computing a Point-wise Mutual Information score between the phrase and meronymy discriminators associated with the product class. Popescu et al apply manual extraction rules in order to find the opinion words.

Kunpeng Zhang (2009), proposed a work which used a keyword matching strategy to identify and tag product features in sentences. Bing xu (2010) , presented a Conditional Random Fields model based Chinese product features identification approach, integrating the chunk features and heuristic position information in addition to the word features, part-of-speech features and context features. Khairullah Khan et al (2010) developed a method to find features of product from user review in an efficient way from text through auxiliary verbs (AV) {is, was, are, were, has, have, had}. From the results of the experiments, they found that 82% of features and 85% of opinion-oriented sentences include AVs. Most of existing methods utilize a rule-based mechanism or statistics to extract opinion features, but they ignore the structure characteristics of reviews. The performance has hence not been promising.

Yongyong Zhail (2010) proposed a approach of Opinion Feature Extraction based on Sentiment Patterns, which takes into account the structure characteristics of reviews for higher values of precision and recall. With a self constructed database of sentiment patterns, sentiment pattern matches each review sentence to obtain its features, and then filters redundant features regarding relevance of the domain, statistics and semantic similarity.Gamgarn Somprasertsri (2010) dedicated their work to properly identify the semantic relationships between product features and opinions. His approach is to mine product feature and opinion based on the consideration of syntactic information and semantic information by applying dependency relations and ontological knowledge with probabilistic based model.

IV CONCLUSION

Visual query interfaces however, have limitations especially when they are used for accessing large and complex datasets. The ease of expressing ones queries is limited due to language barrier and the knowledge of precise key words. Therefore, we are developing a novel query interface where users can use natural language expressions to help author visual queries and address the knowledge gap. The opinions and reviews posted in the websites are analyzed mostly based on static data.

References

[1] Andrea Esuli and Fabrizio Sebastiani, "Determining the semantic orientation of terms through gloss classification", Proceedings of 14th ACM International Conference on Information and Knowledge Management, pp. 617-624, Bremen, Germany, 2005.

[2] Bai, and R. Padman, "Markov blankets and meta-heuristic search: Sentiment extraction from unstructured text," Lecture Notes in Computer Science, vol. 3932, pp. 167–187, 2006.

[3] Bing xu, tie-jun zhao, de-quan zheng, shan-yu wang, "Product features mining based on conditional random fields model " , Proceedings of the Ninth International Conference on Machine Learning and Cybernetics, Qingdao, 11-14 July 2010

[4] Chaovalit, Lina Zhou, Movie Review Mining: a Comparison between Supervised and Unsupervised Classification
Approaches, Proceedings of the 38th Hawaii International Conference on System Sciences – 2005.

[5] Chau, M., & Xu, J. (2007). Mining communities and their relationships in blogs: A study of online hate groups. International Journal of Human – Computer Studies, 65(1), 57–70.

[6] Christopher Scaffidi, Kevin Bierhoff, Eric Chang, Mikhael Felker, Herman Ng and Chun Jin, "Red Opal: product- feature scoring from reviews", Proceedings of 8th ACM Conference on Electronic Commerce, pp. 182-191, New York, 2007.

[7] Chunxu Wu, Lingfeng Shen , "A New Method of Using Contextual Information to Infer the Semantic Orientations of Context Dependent Opinions" , 2009 International Conference on Artificial Intelligence and Computational Intelligence.

[8] Gamgarn Somprasertsri, Pattarachai Lalitrojwong , Mining Feature-Opinion in Online Customer Reviews for Opinion Summarization, Journal of Universal Computer Science, vol.16, no. 6 (2010), 938-955.

[9] Gang Li , Fei Liu , "A Clustering-based Approach on Sentiment Analysis" ,2010, 978-1-4244-6793-8/10 ©2010 IEEE.

[10] Go,Lei Huang and Richa Bhayani , "Twitter Sentiment Analysis", Project Report, standford,2009.

[11] Go,Lei Huang and Richa Bhayani , "Twitter Sentiment Classification using Distant Supervision", Project Report, Standford,2009.