ISSN:0975-9646

Information Retrieval Using Subjective Analysis on Social Blogs

Haritha. A [1], Ch. Manohar Venkat[2], J.S.N.S. Sneha Priya[3] P.V.S.Lakshmi[4], G.Lakshmi[5]

[1, 5] Asst. Professor, [2, 3] Student, [4] Professor
Dept of IT, PVP Siddhartha Institute of Technology

***Abstract*—Number of blogs is increasing at a rapid pace and many potential applications for opinion detection and monitoring are arising as a result. Perspectives vary from one individual to another. In subjective analysis of various domain related reviews, one important problem is to produce a summary of opinions based on attributes. People leave on the web their opinions on domains and services they have known, it has become important to develop methods of (semi-)automatically classifying and gauging them. Subjective analysis helps in classifying adjectives which are commented in blogs into good, bad and neutral sets. Basing on these words one can decide whether the specific domain which is being searched has a positive or a negative review. We represent the blog posts in the form of bags of sentiments and use SVM(Support Vector Machine).When we use SVM, the objects are represented as vectors . A hyper plane will be chosen so that the margin or separation between two classes is distinct and clear. This separates objects of one class to the other .In this way ,categorizing the opinions turns to be easy than the rest of approaches. Moving one step ahead of this procedure, we intend to develop an application which elucidates the customer to know whether the domain they are tending to know about is better than the rest of domains. We retrieve the information about the domains which are given as comments in social blogs, based on these comments we give the results in a comparative form with rest of the domains which makes it much easier for the user to choose the best. Not only this, but it helps analytics to identify the domains which are unsuccessful and try in improvising their attributes so that there would be exciting improvement and achievements.**

I. Introduction

Information which is available either gives factual statements or opinions which are based on perceptions. Every person has his or her own opinion to judge about things like products or domains such as political parties or movies or educational institutions or anything under the sun, these opinions vary from one individual to another. At every point of time a person has to take a decision regarding each and every aspect in their lives. Earlier people gathered information about what they have to choose and decided what and which to choose. With the increasing pace of evolution of internet, users started to know about the domain by browsing about them in the internet. This process got advanced by collecting opinions in social blogs. A social blog is considered as a web site which would have collection of opinions which are either posted by experts or by people who is familiar with the domain. In the present day we can find several social blogs which are having lot of information posted for certain domains such as film reviews, public opinion about political parties, reviews about advanced technologies, electronic devices, automobiles. Though there are opinions posted we find several blogs posting opinions about a single issue which leads to collecting positive and negative opinions , so people are intended to develop a system that can identify and classify opinion or sentiment as represented in an electronic text which is in general tough for the user to find which is relevant data and which is not. For making this easy we have a major area of expertise that is Data Mining. As discussed earlier, several issues arise when it comes in the aspect of opinions. These are totally based on personal sentiments and perceptions of the individual. It turns to an easy task if we apply the concept of subjective analysis.

Keywords: Data mining, opinion mining, subjective analysis.

Data Mining: This entirely aims at extracting knowledge from large amount of data in an understandable structure that is useful for companies and individuals. This uses intelligence techniques , neural networks, and advance statistical tools to reveal trends, patterns and relationships. This can also be referred as data surfing.

Opinion Mining: It is an automated extraction of subjective content fromk text and identifying the orientation such as positive, negative or neutral in that text. It aims aims to explore feelings of a person who write the text. These opinions can be evaluated in two ways:

1. *Direct opinion*: It gives positive or negative opinion about the object directly.
2. *Comparison*: To compare the object with some other similar objects.

Basic components of an opinion are :

1. *Holder*: The person that gives a specific opinion on an object.
2. *Object*: Entity on which an opinion is expressed by user.
3. *Opinion*: A view, sentiment, or perception of an object done by a user.

Subjective analysis: It is the technology to evaluate a text and predicate the text's subjectivity or sentiment .It first finds out the keywords from the text that are of evaluating a feature or sentiment orientation to represent the text.

This process follows few sequential steps. First we have to collect the opinions which are posted in social blogs in the form of comments. This electronic data will be extracted in data sets formats. We give these comments as input, for example when we visit a social blog which is about film reviews; we get to know the opinions of people who have already watched or by critics who analyze films.

These opinions are refined into categories which are positive, negative and neutral. The collected data will be subjected into categories based on the key words by sending them through the adjective list. According the adjectives such as "good", "wonderful", "awful" and many others which are related to the subject, the product or the item which is acting as object will be decided whether it is good enough or bad.

After the classification is done, the predicted the output can be visualized. The object is declared to be good or bad by first identifying the number of bad and good objectives. If the number of comments which say that the object is good are more than it would represent the output that it is positive and the same is the negative output is predicted by observing the negative comments. We all know that a picture can express an opinion thousand times better than text. So, the obtained output would be represented in the form of graph which will show the features of the object along with the prediction whether it is positive or negative or neutral.

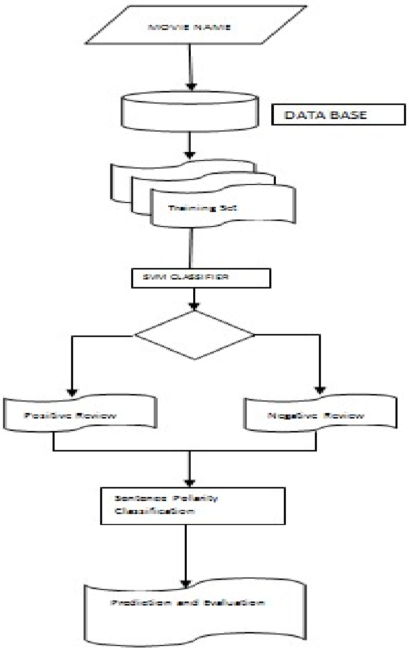
This would help users to understand the object in a better way, keeping several perspectives in mind. It would also help business analytics who analyze the features of products, and will help in rectifying the flaws by estimating the predictions so that there would be higher success rate.

II. RELATED WORK

At the initial stage, we used natural language processing in order to classify opinions. In this process, we collect the opinions and create the adjective list based on the words present in the opinions. If we consider a product, we have particular adjectives which decide about the features of the products, certain adjectives can mean entirely different depending on the context or the feature being considered. Let us consider the adjective "high", when we are talking about features like " high pixel value", high can be taken as a positive word whereas when we are talking about cost, "high cost" would tend to be a negative feature. So, it is difficult to judge certain words whether to classify them into positive or negative or neutral, a lot of ambiguity would come into existence if we consider the adjective lists for certain contexts or domains. The corpus which we consider for a product review and a movie review might consist of same set of adjectives but would lead to lot of differences. Eventually we are developing further systems which would resolve these problems.

III. PROPOSED SYSTEM

To overcome the existing problems, we are using Support Vector Machine (SVM). In SVM we treat each object or opinion as a vector, a hyper plane will be found out which deals with the classification of the objects into classes. The vector value of the hyperplane is compared with the vector value of the object, now it is a simple task of comparing the values and decides to categorize the objects into classes. We are now considering the domain of movie reviews. This is the key step for the entire process. This classification will have the following sequence.



A. *Consider the training set*: First we collect the raw data, we collect entire data sets that is, we take all the possible opinions which are existing. These consider entire positive, negative and neutral words all together, without any segregation. For the considered domain, we collect all comments about the existing movies in the blogs and maintain them as a set. Though we have tons of opinions in the data set, it is highly impossible to judge which is positive and negative, so we need to refine the opinions using SVM.

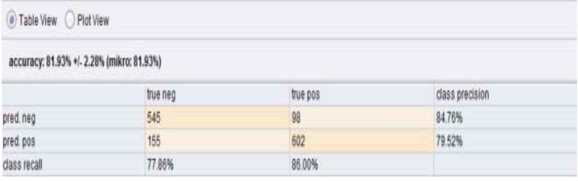
B. *SVM Classifier*: This plays a significant role in refining the collected opinions. First each opinion is considered as a vector with a value, now fix the hyper plane which leads to the classification of objects into respective classes. There are several types of classifiers, out of which we are two:

- *Linear SVM (my SVM*): This is a simple way to classify, we will have only one hyper plane which would classify the opinions into two classes only. The vector value of the object and hyper plane will be compared and then the objects will be placed into their classes.
- *SVMlib*: This consists of multiple hyper planes, which leads to classification of the corpus into several classes. This can be used to the maximum extent in issues such as feature based classifications.

In the present context, the classes would be positive and negative. This segregation helps us to identify whether the comments in multiple blogs pred

ict whether the given film is good or bad.

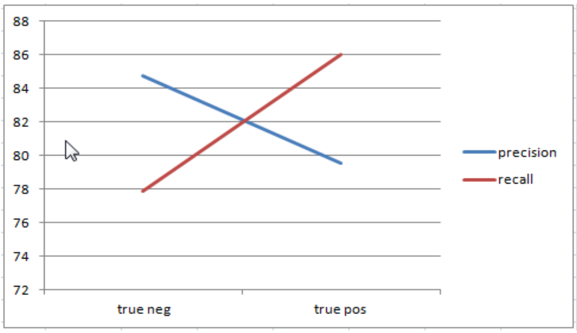
C. *Prediction and Evaluation*: After the refinement of raw data sets into relevant classes, we start to know the final prediction about the movie. As we consider all the opinions in several numbers of blogs, it would contain possible sentiments which exist about a particular movie. We obtain whether the movie has a positive or a negative response; this is based on the frequency count and document count of the vectors or words considered. Frequency of the word is not the entire criteria on which we decide the verdict, but the importance of the word. After evaluation, the result will be in both tabular and graphical forms along with the accuracies which makes user to feel comfortable.

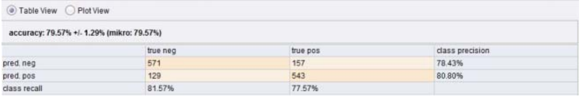


accuracy: 81.93% +/- 2.28% (mikro: 81.93%)

	true neg	true pos	class precision
pred. neg	545	98	84.76%
pred. pos	155	602	79.52%
class recall	77.86%	86.00%	

Fig 1. SVM linear vector list





accuracy: 79.57% +/- 1.29% (mikro: 79.57%)

	true neg	true pos	class precision
pred. neg	571	157	78.43%
pred. pos	129	543	80.80%
class recall	81.57%	77.57%	

Fig 2. - SVMlib vector list

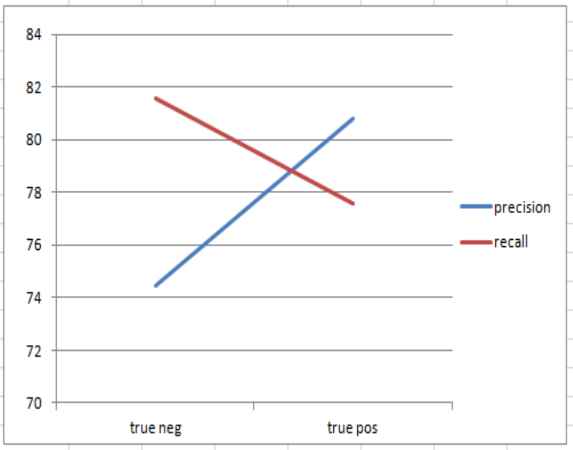


Fig 3. SVM lib

IV. Conclusion

With the revolutionary advancements in every aspect in our daily routines, there is a drastic change in the way decisions are being made by an individual. Irrespective of the importance of the issue, people started gathering paramount opinions from several blogs regarding the issue. The proposed system retrieves all the comments posted in social blogs about the domain and segregates them into different classes, that is either they are positive or negative. This system is not only applicable for just analyzing a movie, but can be applied for any domain of a person's interest. This system is driven with a motivation that, it has to be easily understood and can be analyzed by any novice user to an analytical expert. We have compared two systems, SVM linear and SVM lib, which helped us in proving that SVM linear is more accurate than SVM lib, the graphical representation gives crystal clear prediction.

REFERENCES

[1] G. Conrad and Frank Schilder Opinion Mining in Legal Blogs

[2] Zhongwu Zhai, Bing Liu, Peifa Jia and Hua Xu Clustering Product Features for Opinion Mining.

[3] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In VLDB'94, 1994.

[4] A. Andreevskaia and S. Bergler. Semantic tag extraction from wordnet glosses. 2007.

[5] K. Church and P. Hanks. Word association norms, mutual information, and lexicography. In Computational Linguistics, volume 16, pages 22–29, 1990.

[6] D. Downey, M. Broadhead, and O. Etzioni. Locating complex named entities in web text. In Proceedings of IJCAI'07, pages 2733–2739, 2007.

[7] V. Hatzivassiloglou and K. McKeown. Predicting the semantic orientation of adjectives. In Proceedings of 35th Meeting of the Association for Computational Linguistics, Madrid, Spain, 1997.

[8] M. Hu and B. Liu. Mining and summarizing customer reviews. In Proceedings of KDD'04, ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, WA, 2004.

[9] http://www.corequant.com/?p=1

[10] [10] M. Roche and V. Prince. AcroDef: A Quality Measure for Discriminating Expansions of Ambiguous Acronyms. In Proceedings of CONTEXT, Springer-Verlag, LNCS, pages 411–424, 2007.

[11] [11] http://fimi.cs.helsinki.fi/fimi03/