International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 04, April-2017

A Survey on Various Approaches for Sentiment Analysis and Performance Optimization

Rohini Jadhav
Assisstant Professor
Department of Computer Engineering
Bharati Vidyapeeth Deemed University
College of Engineering, Pune, India

Ambesh Bajpai
B Tech. Student
Department of Computer Engineering
Bharati Vidyapeeth Deemed
University College of Engineering,
Pune, India

Animesh Sinha
B Tech. Student
Department of Computer Engineering
Bharati Vidyapeeth Deemed University
College of Engineering, Pune, India

Abhishek Kumar Singh
B Tech. Student
Department of Computer Engineering
Bharati Vidyapeeth Deemed University
College of Engineering, Pune, India

Abstract:- **Sentiment analysis, which is also known as opinion mining and is a part of text mining were we mine data to extract their opinion. People provide their opinion, comments, feedbacks and these are very important indicators. From opinion mining a company may get a feedback for their products from there users, so that they can improve it further on the basis of feedbacks. Politicians can also use it for analyze sentiment about their policies or some other political issues.**
This paper presents various technique used in the whole process of opinion mining and how different choices will affect our result. Here we will also see the problems that are still there in extracting the sentiment about a product.

Keywords: Sentiment Analysis, Support Vector Machine, Online Commerce, Opinion Mining

I. INTRODUCTION

Sentiment analysis studies the sentiment of people towards certain entities. There are millions of people on social networking sites expressing their opinions about various products and their features. On internet we find a large amount of data which can be very useful if we can simply extract opinions or emotions from it. It can act as active feedback for the companies developing the products.
Sentiment analysis task is to retrieve opinion about certain product and features and to classify them as positive or negative. Generally review is not explicitly negative or positive. It consists of mixed emotion like, "Samsung S6 camera is good but it has battery issues". It's a mixed review where the customer is appreciating the camera and criticizing the battery life.

These sentiments can be categorized as positive, negative or neutral; or into an n point scale and it can be like:-very good, good, satisfactory, bad, very bad.

On internet several flaws hinder the process of sentiment analysis as people are free to post their opinion in the forum. Spammers post spam which are irrelevant to the topic and are known as fake comments.

II. LITERATURE SURVEY

A. Selecting a Template

Lina Zhou et al., [1] investigated movie review mining using machine learning and semantic orientation. Supervised classification and text classification techniques are used in the proposed machine learning approach to classify the movie review. A corpus formed that represents the data in the documents and using this corpus training of all the classifiers is done. Thus, the proposed technique is more efficient. Though, the machine learning approach uses supervised learning, the proposed semantic orientation approach uses "unsupervised learning" because it does not require prior training in order to mine the data. Experimental results showed that the supervised approach achieved 84.49% accuracy in three-fold cross validation and 66.27% accuracy on hold-out samples. The proposed semantic orientation approach achieved 77% accuracy of movie reviews. Thus, the study concludes that the supervised machine learning is more efficient but requires a considerable amount of time to train the model. On the other hand, the semantic orientation approach is slightly less accurate but is more efficient to use in real time applications. The results confirm that it is practicable to automatically mine opinions from unstructured data.

Bo Pang et al., [2] used machine learning techniques to investigate the effectiveness of classification of documents by overall sentiment. Experiments were done that demonstrated that the human produced baseline is less better than machine learning techniques for sentiment analysis on movie review data. The experimental setup consists of movie-review corpus with randomly selected 700 positive sentiment and 700 negative sentiment reviews. Features based on unigrams and bigrams are used for classification. Learning methods Naïve Bayes, maximum entropy classification and support vector machines were employed. Inferences made by Pang et al., is that human produced baseline is less better than machine learning techniques for sentiment classification. The accuracy achieved in to topic based categorization is much more than sentiment classification when compared.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 04, April-2017

Zhu et al., [3] proposed aspect based opinion polling from free form textual customers reviews. The aspect related terms were learnt using a multi-aspect bootstrapping method and were used for aspect identification. A proposed aspect-based segmentation model, segments the multi aspect sentence into single aspect units which was used for opinion polling. They tested on a Chinese restaurant reviews and achieved 75.5 percent accuracy in aspect-based opinion polling tasks. They used a opinion polling algorithm for this. This method is easy to implement and are applicable to other domains like product or movie reviews.

Jeonghee Yi et al., [4] proposed a Sentiment Analyzer to extract opinions about a subject from online data documents. Sentiment analyzer uses natural language processing techniques. The Sentiment analyzer calculates on the subject all the references and determines the sentiment polarity of each reference. The sentiment analysis conducted by the researchers utilized the sentiment lexicon and sentiment pattern database for extraction and association purposes. Online product review articles for digital camera and music were analyzed using the system with good results.

Alekh Agarwal et al., [5] proposed a machine learning method which contained knowledge gathering through synonymy graphs, which helped in effective opinion classification. This approach helps in knowing the degree of influence on their sentiment analysis among relationships of documents. This is brought about by the use of graph-cut technique and opinion words got through synonymy graphs of Wordnet. The proposed approach also improves the accuracy of predictions in classification task. Experiments using the system have given results with an accuracy of over 90%, with an added advantage of reduction in processing time, with minimal difference in final accuracies. The proposed methodology from the authors resulted in the following conclusions:

1. Automated mining of linguistic information is possible, so demonstrated with the structure of links in Wordnet.
2. Generic method of using graph-cut technique for efficient opinion classification.

Ahmed Abbasi et al., [6] proposed novel sentiment analysis methods to classify web forum opinions in multiple languages. The sentiment analysis method that is proposed here evaluated the sentiment in English and Arabic content, for which, it utilized the function of stylistic and syntactic features. The Entropy weighted Genetic Algorithm is included for enhancing the performance of the classifier and also to find out the key features. Experiments were conducted using movie review data set and the results demonstrated that the proposed techniques are efficient.

Anidya et al., [7] ranked the product reviews based on customer-oriented and manufacturer ranking mechanism. The expected helpfulness of the review is used for the ranking and also ranking is based on the expected effect on sale. The proposed methods identify the reviews which have the most impact. For feature based products, he reviews that confirm the information contained in the product description are used, and reviews with subjective point of view are useful for experience goods. Econometric analysis with text mining techniques and with subjectivity analysis is used in the proposed method. Product prices and sales ranking publicly available on amazon.com were used to compile the data set. The product and sales data are the two sets of information collected for each product. Products such as audio and video players, digital cameras were used to form the data set. The empirical analysis is performed using the compiled data set.

Michael et al., [8] presented „Pulse" a prototype system for mining topics and sentiment orientation from free text customer feedback. Blogs, newsgroups, feedback email from customers, and web sites that collect product reviews are all source of free text customer feedback. The proposed system is designed to handle the free form information of the customer feedbacks as the sources of information are less structured than traditional surveys. A clustering technique and machine learned sentiment classifiers were used in the proposed method. Sentiment and topic detections are performed at the sentence level not at the document level. The Pulse was evaluated using car reviews database, and the sample data contains 4, 06,818 customer car reviews written over a four year period. The data set contained almost 900,000 sentences in total. Sentiment analysis was performed using 3000 randomly selected sentences. Each sentence is classified as positive, negative and others. The other category contained both positive and negative sentiment and sentences with no complex sentiments. Training of the sentiment classifier was done using 2500 sentences and the remaining 500 sentences are reserved for test set. Results reflect the efficiency of the proposed system.

Miniqing Hu et al., [9] performed mining and summarization process to all the customer reviews of a product. The proposed process is carried out in three steps:

1. The product features commented by the customer in the review are mined. Natural language processing and Data mining techniques are used for mining.
2. The opinions in the review are identified and the opinions are classified as positive or negative. Set of adjectives words called opinion words are identified and semantic orientation of the opinion words is determined. WordNet can be used to identify the semantic orientation and the opinion orientation of each sentence is decided.
3. Summarize the results. The objective of the study is to perform feature based summary of a large number of customer reviews of a product sold online.

Qui et al., [10] analyzed the problems related to opinion mining such as opinion lexicon expansion and opinion target extraction. Opinion targets are entities and their attributes on which opinions have been expressed. The list of opinion words such as good, bad, excellent, poor used to indicate positive and negative sentiments is Opinion lexicon. The links between the opinion words and targets Syntactic relations are identified using dependency parser based on bootstrapping. The process uses semi-supervised methods, opinion word seeds are used in the initial opinion lexicon.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 04, April-2017

Bootstrapping process is started using the initial opinion lexicon. Double propagation method is used as information is propagated back and forth between opinion words and targets.

Lei Zhang et al., [11] identified domain dependent opinion words. Noun and noun phrases that indicate the product feature which implies opinions are found using a feature based opinion mining model. Two steps are used to identify the noun product feature which means the positive or negative opinion. Sentiment context of each noun feature is determined in the Candidate identification step. And also a list of candidate features with positive opinions and list of candidate features negative opinions is produced. Noun product feature is directly modified into positive and negative opinion words in pruning step. Opinion lexicon complied by Ding et al. (2008) was used to identify the opinion polarity on each product feature in a sentence. For a sentence s which contains a product feature f, opinion words in the sentence are first identified by matching with the words in the opinion lexicon. An orientation score for f is computed and the semantic orientation of the positive word is assigned the score of +1, and a negative word is assigned the score of -1. On summing up of all the scores, if the final score is positive, then the opinion in s on the feature is positive. If the score obtained is not positive, then the opinion in s on the feature is negative.

Xiaowen Ding et al., [12] proposed a holistic lexicon-based approach which uses external indications and linguistic conventions of natural language expressions to determine the semantic orientations of opinions. Advantage of this approach is that opinion words which are context dependent are easily handled. The algorithm used uses linguistic patterns to deal with special words, phrases. Researchers built a system called Opinion Observer based on this technique. Experiments using product review dataset was highly effective. It was shown that multiple conflicting opinion words in sentence are also dealt with efficiently. This system shows better performance when compared to existing methods.

III. PROBLEMS IN SENTIMENT ANALYSIS

Sentiment analysis continues to spread across politics, industries etc. Clients are looking for more substance, and benefitting as multifaceted topic, it's complicated.

A. Degree of accuracy

The degree of accuracy is a issue and is hard to answer. It depends on number of data sets involved, level of text we are analyzing, the voice sound, quality of videos and many other variables.

e.g.:- An article from The Atlantic about Anne Hathway was drives Berkshire Hathway's stock prices. This was due to some hedge funds using primitive data sets where Hathway's mentions aren't properly differentiated.

B. Machine do analytics and humans do analysis

Machine learning is isolated and humans don't learn in isolation.

Human have some prior knowledge which they have gained from their experience.

It's important to have humans in the loop for continuous training (and to avoid disasters like we have seen in previous example).

C. Keep an open mind about findings

People generally have a specific hypothesis in mind, and they selectively extract data that proves their theories.
They don't consider the wealth of information they can get from the internet.

D. Ambiguous data

In our data sets there is data which is not clear, it has ambiguity.It can be a sarcasm which could be taken as a positive sentence.

e.g.:- This movie was as good as Cat Women.
There can also be syntactic ambiguity in a sentence like:-

A man saw a boy with a telescope. Here it is not clear whether the telescope was with the boy or man.

IV. METHODS

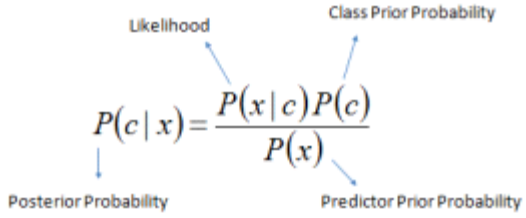
To overcome the above problems involved with sentiment analysis, we can use the following methods.

Naive Bayesian classifier

Naïve Bayes is a concept of classification techniques which is based on Bayes' Theorem including an assumption of independence among various predictors. Basically a Naive Bayes classifier assumes presence of a specific feature in a text which is unrelated to presence of other features present in it. For example, a fruit may be considered to be an Guava if it is green in color, round, and about 3 inches in diameter. These features basically depends upon each other or upon the existence of the other features, all of these properties independently contribute to the probability that this fruit is an Orange and that is why it is known as 'Naive'.

Naïve Bayes model is easy to build and particularly useful for very large data sets. It's a very simple algorithm and is known to outperform even highly sophisticated classification methods.

Bayes theorem provides a path of calculating posterior probability P(c|x) from P(c), P(x) and P(x|c). Look at the equation below:



$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Above

P(C|X) = posterior probability (C, target), given predictor (X, attributes).

P(C) = prior probability.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 04, April-2017

P(X|C) = likelihood which is the probability of predictor.

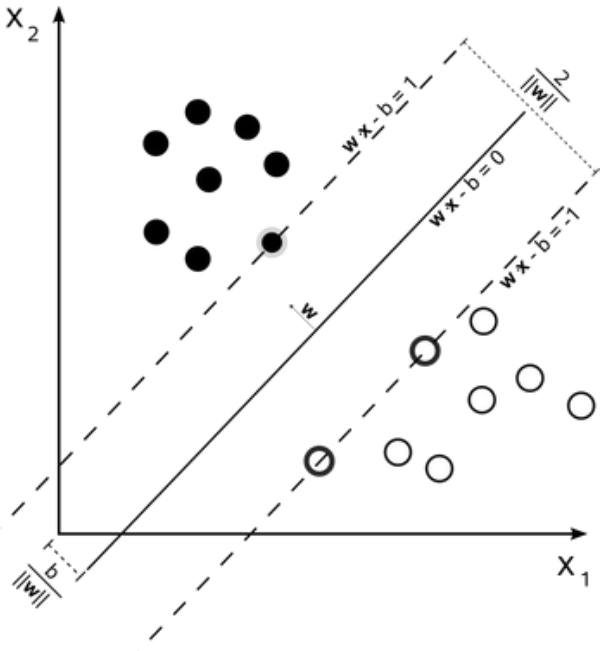
P(X) = prior probability of predictor.

Support vector machine

Support vector machine is a method for the classification of both linear and nonlinear data. The SVM searches for the linear optimal separating hyperplane (the linear kernel), if the data is linearly separable, which is a decision boundary that separates data of one class from another.

If the data is linearly inseparable, the SVM uses nonlinear mapping technique to transform the data into a higher dimension, then it solves the problem by finding a linear hyperplane. Functions to perform such transformations are called kernel functions.

Although it is not immediatly obvious from the name, the SVM algorithm is a 'simple' linear classification/regression algorithm[6]. It tries to find a hyperplane which seperates the data in two classes as optimally as possible.
Here is (as optimally as possible) means that as much points as possible of label A should be seperates to one side of the hyperplane and as points of label B to the other side, while maximizing the distance of each point to this hyperplane.

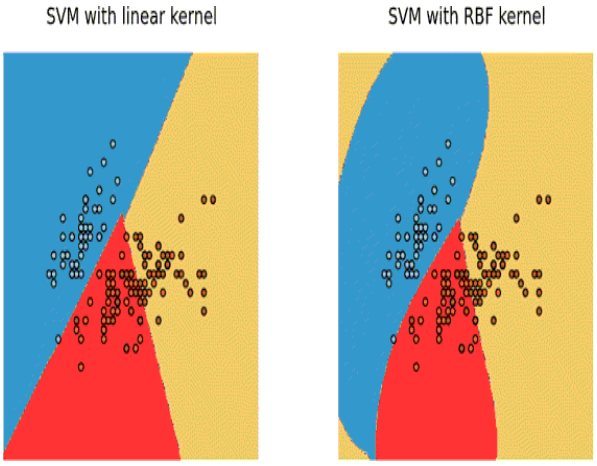


In the image above we can see this illustrated for the example of points plotted in 2D-space. The set of points from the graph are labeled with two categories (illustrated with black and white points) and Support Vector Machine basically chooses that hyperplane which maximizes the margin between the two given classes. This hyperplane is given by

$$(\overrightarrow{w}.\overrightarrow{x}) + b = \sum_i y_i \propto_i (\overrightarrow{x_i}.\overrightarrow{x}) + b = 0$$

$$\overrightarrow{x_i} = (x_{i1}, x_{i2}, \ldots . x_{in})$$

where y_i is a n-dimensional input vector, is its output value, $\overrightarrow{w} = (w_1, w_2, \ldots . w_n)$ is the weight vector (the normal vector) defining the hyperplane $\propto_i$ and the terms are the Lagrangian multipliers.



A classification example of SVM

V. APPLICATIONS

Online Commerce

The most basic general use of sentiment analysis is in e-commerce. Companies allows their users to submit their experience about their products. Users provides rating to the products and provide summary. Customers can view opinions and recommendation about the product.

Brand Reputation Management

Brand Reputation Management is basically involves managing reputation of the brand of the company. Rating and opinions about the product from customers can enhance or degrade the reputation. So basically Sentiment Analysis helps in determining products an company's brands, which affects the reputation of the brand.

Government Sentiment

Government uses technique sentiment analysis to assess their strength and problems by analyzing from public opinions.. It tracks citizen's opinions, identifying candidate's in a recruitment in a government job, assessing success of tax returns, or many other areas, we can see the potential for sentiment analysis.

VI. CONCLUSION

Sentiment Analysis is basically a study for collecting and analyzing users emotions, sentiments and attitudes for various applications.

The paper starting from the analysis of different studies provided in the literature survey, with discussion of various methods for sentiment analysis following by its applications and future scope.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 04, April-2017

This paper tackles with various basic problems of sentiment analysis. Different methods of sentiment analysis for various applications and how they affect the results. Two basic methods are discussed above Naive Bayesian classifier and Support Vector Machine both of which are machine learning approaches.

Any future scope of improvement :

- Product review based on opinions which are in multiple languages.
- Deal with problems of analyzing slangs.
- Deal with problems of analyzing sarcastic opinions.
- Deal with comparative opinions and find which one of the product is best.

VII. REFERENCES

[1] Lina Zhou, Pimwadee Chaovalit, "Movie Review Mining: Comparison between Supervised and Unsupervised Classification Approaches", Proceedings of the 38th Hawaii International Conference on system sciences, 2005.

[2] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan, "Thumbs up? Sentiment classifications using machine learning approaches techniques", In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2002.

[3] Zhu, Jingbo Wang, Huizhen Zhu, Muhua Tsou, Benjamin K. Ma, Matthew, "Aspect-Based Opinion Polling from Customer Reviews", IEEE Transactions on Affective Computing, Volume: 2,Issue:1 On page(s): 37. Jan-June 2011.

[4] Yi, J., T. Nasukawa, R. Bunescu, and W. Niblack: 2003, "Sentiment Analyzer: Extracting Sentiments about a Given Topic using Natural Language Processing Techniques", In: Proceedings of the 3rd IEEE International Conference on Data Mining (ICDM-2003). Melbourne, Florida.

[5] Alekh Agarwal and Pushpak Bhattacharyya, "Sentiment analysis: A new approach for effective use of linguistic knowledge and exploiting similarities in a set of documents to be classified", In Proceedings of the International Conference on Natural Language Processing (ICON), 2005.

[6] Ahmed Abbasi, Hsinchun Chen, And Arab Salem, "Sentiment Analysis in Multiple Languages: Feature Selection for Opinion Classification in Web Forums", ACM Trans. Inf. Syst., Vol. 26, No. 3. (June 2008), pp. 1-34.

[7] Anindya Ghose, Panagiotis G. Ipeirotis, "Designing Novel Review Ranking Systems: Predicting Usefulness and Impact of Reviews", Proceedings of the Ninth International conference on Electronic commerce ICEC07 (2007), pp: 303-310.

[8] Michael Gamon, Anthony Aue, Simon Corston-Oliver, and Eric Ringger, "Pulse: Mining Customer Opinions from Free Text Natural Language Processing", Microsoft Research, Redmond, WA 98052, USA.

[9] Minqing Hu and Bing Liu, "Mining and Summarizing Customer Reviews", Proceedings of the tenth ACM SIGKDD International conference on knowledge discovery in data mining (KDD-2004), August 22-25.

[10] Guang Qiu, Bing Liu, Jiajun Bu and Chun Chen. "Opinion Word Expansion and Target Extraction through Double Propagation." Computational Linguistics, March 2011, Vol. 37, No. 1: 9.27.

[11] Lei Zhang and Bing Liu. "Identifying Noun Product Features that Imply Opinions." ACL-2011 (short paper), Portland, Oregon, USA, June 19-24, 2011.

[12] Xiaowen Ding. A Holistic Lexicon-Based Approach to Opinion Mining- WSDM'08, February 11-12, 2008, Palo Alto, California, USA.2008 ACM 978-1-59593-9279/08/0002.