ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 4, Issue 2, February 2017

Sentiment Analysis and its Challenges

[1] Ravneet Kaur [2] Gaurav Gupta [3] Gurjit Singh
[1]M.tech Student, Department of Computer Engineering, Punjabi University, Patiala, Punjab
[2][3]Department of Computer Engineering, Punjabi University, Patiala, Punjab

***Abstract*— Sentiment analysis is the wide study area in educational and commercial fields. The word sentiment speaks of about the emotion or views of the public towards the particular area or field. So it is also called as opinion mining i.e. to mine the opinion of public to collect the required knowledge. Sentiment analysis can be used in any field. Earlier research is carried out to find the public sentiments. With enhancement sentiment analysis can become a useful area for research. Here we analyze previous research and interpret their results and also explain the various classification algorithms that are widely used for sentiment classification like Naive Bayes, Support Vector Machines (SVM) and K-NN classifier. Also the methodology for the sentiment analysis is explained using proper steps and examples. This paper also discusses the application areas and challenges in the field of sentiment analysis.**

***Index Terms:* - Data mining, Opinion mining, Sentiment Analysis, Text mining.**

I. INTRODUCTION

The use of internet is increasing day by day. People are more interested to use social media these days because of the services provided by the social media. Each person is willing to give the review on the social media about any particular product or services, which generates a huge amount of data on the internet. The feedback of the people helps other persons to develop their view point about that particular product or services. An enormous amount of data is available from different fields like medical, science, engineering, marketing and many more.

A huge amount of data is available now days. So a number of different techniques or methods are available to store that data & extract the meaningful information and hidden pattern from that database. Knowledge discovery in database is the complete procedure to convert data into useful information according to the user requirement. Data mining is the part of knowledge discovery in databases and is used to find useful and valid patterns. Text mining is the method to find the information from text which is in the form of natural language. Web mining is the sub part of text mining which considers web pages to collect and extract useful information. Web mining can be divided into three parts. Web usage mining is used to find the number of times a user use a particular web page. Web structure mining is used to identify the basic structure of the web pages. Web content mining is procedure to get the useful content from the various web pages or web sites. Web content mining is further classified into opinion mining.

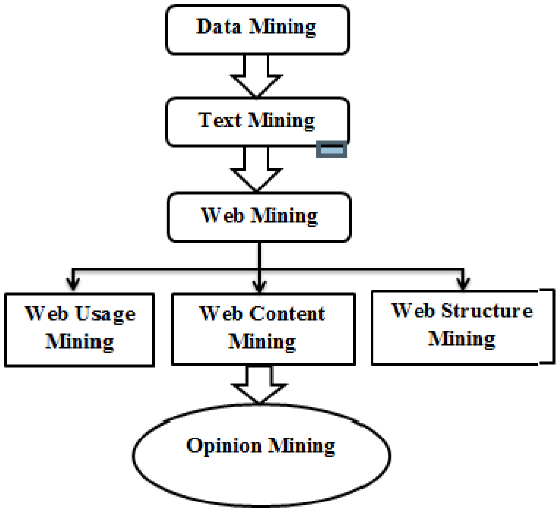

Figure 1: Data Mining Hierarchy [10]

II. SENTIMENT ANALYSIS

Sentiment analysis or Opinion Mining is the way of discovering the feelings and empathy of individuals to exact areas of attention. It may be an item or a movie, reviews of persons really matters. These reviews also affect any other person's policy making process. If a user wants to buy a new product, firstly he would see the reviews or comments of other individuals. Depending upon the polarity of reviews he decides whether to buy the product or not. Social networking websites such as Facebook, twitter are the platform where individuals update their status. Individuals tweet on their twitter account regarding any specific topic of their interest. Sentiment analysis is used to forecast the stock market, to predict the result of certain polls, to identify the effectiveness of any product or in many more.

Sentiment analysis is a process to classify the opinion of the person mainly expressed in the form of tweets. Tweets can be classified as positive, negative or neutral. For example, the tweet "I am very happy today because i topped in my class" is a positive text and the text "I hate this" is a negative text. Consider another example "Rustom is very good movie i suggest everybody to definitely watch this movie", it is clear that user review is totally positive towards the movie Rustom. Occasionally it is not easy to interpret whether the tweet is positive or negative, then we call the tweet as neutral. "Rustom is not bad but i don't understand why people put it as number one movie" these types of tweets considered as neutral. The tweets given above are all about the particular topic which is a movie named Rustom.

Twitter is a most commonly used social networking website that provides its users to update a 140 characters status. It stores a huge amount of data set about the specific topic. World Wide Web made it easier for people to share their ideas over the internet. Sentiment analysis makes use of natural language processing and text processing to complete the whole task i.e. to identify the opinion of the public. For example, if one wants to know - if the Chief Minister of Punjab is doing his job properly or not? The best way to answer this is view any social networking site. It is easy to find out about the work done by Punjab CM by viewing the tweets of user. But the problem is that there a huge number of tweets how we recognize that how many people are positive or negative towards the CM work. The best possible solution is to use the sentiment analysis on the tweets and find out what people say about Punjab CM.

Classification of Sentiment Analysis

Sentiment analysis can be classified into three forms as follows [1].

1) Positive Sentiment: It is the collection of good words in the sentiment. If the number of good words increased it is referenced as a Positive sentiment. For example, if reviews of a product have more positive comments then it is sure to be bought by many customers.

2) Negative Sentiment: It is the collection of bad words in the reviews. If the number of bad words increases then it becomes a negative sentiment. For example, if the total reviews or tweets about any product have more negative reviews then the product is not so useful then it is bought by very less number of people.

3) Neutral Sentiment: If the tweet is neither classified as negative nor positive then it is treated as neutral sentiment.

III. LITERATURE SURVEY

A huge amount of work has been carried out in Sentiment Analysis which is done in order to get the opinion of the public towards any specific topic or trends. Opinion got from the public is fu

1) Opinion Mining and Sentiment Analysis

Rushlene and Ravneet [2] have worked on opinion mining and sentiment analysis, they developed a new algorithm that has a good accuracy in predicting the sentiment of any data. The paper intends to predict the rise and fall of the stock prices. Effects of the public tweets on variation of stock prices are determined. It includes data extraction, data cleansing and use of suitable algorithm in the sentiment analysis.

Dataset: Tweets from multinational firm Samsung Electronics Ltd.

Features:

1. A new algorithm is developed that can be easily implemented in the java programming.
2. MS Access database is used to create the dictionary for positive, negative and neutral words.
3. The algorithm takes the tweet to analyze as inputand output given in the form of Sentiment score for the given tweet.

Results:

Tweets using the algorithm are analyzed with the correctness of 80.6%.

2) Sentiment Analysis of English Tweets Using Rapid Miner

Pragya and Santosh [3] used the data mining techniques for classification of the tweets. Text mining techniques like tokenization, stemming to change them into useful form to predict the sentiments of the tweets.

Dataset: Tweets collected from twitter.com website

Features:

1. Different datasets are taken in the form of text files.
2. Text mining operations are applied.
3. Naive Bayes and K-NN classifier are used to predict the sentiment of given tweets.
4. Classifiers are trained with training data and results are given on test data set.

Results:

Results of both the classifiers are compared and the results of K-NN classifier are more correct than the Naive Bayes classifier. Accuracy of Naive Bayes is 63.33% and K-NN is 70.00%.

3) Sentiment Analysis of E-Commerce and Social Networking Sites

Shubhiand Ashnaanalyse [4] the effect of user views over the social networking website. They lead a survey on 100s of replies obtained by the customers and determine if the whole reply is positive, negative or neutral. List of sentiments of emoticons, interjections and social acronyms are also given.

Dataset: Sentiments extracted from Facebook Post "One of the hardest lessonsin life is letting go whether its guilt, anger, love, loss or betrayal. Change is never easy. We fight to hold and fight to let go :D"

Features:

1. On behalf of Facebook Lexicons are categorized into different types: Lexicon of emoticons, Lexicon of interjections and social acronyms.
2. Survey is done on 45.45% females and 54.55% male's response that describe sentiment analysis study.

Result of Facebook Post:

Facebook post whole sentiment is positive having 79% positive, 18% neutral and 3% negative.

Survey Results:

1. Have you ever changed your opinion or decision due to online reviews? 79% user give yes answer and 21% give no answer.
2. Do online reviews play a role in your decision making? 91.92% of people are agreed and 7.07% people does not depend on online reviews and 1.01% sometimes rely on reviews.
3. How useful are the reviews on the Facebook, Twitter or Flip kart? Flip kart is measured as most useful site as 27.16% user has voted it very useful site.
4. Which sites do you visit regularly? 81% public support Facebook followed by Flip kart and Amazon.

4) A survey on Sentiment Analysis on Twitter Data Using Different Techniques

Bholane and Deipali [1] have done a survey on Twitter data using different techniques. They read the previous research papers on the various topics and find out what are the advantages and disadvantages of the particular research paper.

Features:

1. Sentiment analysis is classified into three classes: Positive Sentiments, Negative Sentiments and Neutral Sentiments.
2. Sentiment classification is done based on three levels: Word level, Phrase level and Document level classification.
3. Comparison is of various sentiment classification techniques are done.

Comparative Results:

Table 1:

Comparative results for various classification techniques

Model/Algorithm	**Dataset**	**Accuracy (%)**
SVM	Amazon product review data	89.8
NB	Amazon product review data	89.4
TwitterSentiment	SNAP Twitter dataset	57.2

Naive Bayes classifier is insensitive to unbalanced data which gives more accurate results.

IV. METHODOLOGY

A dictionary is created for negative and positive words. The main approach done in sentiment analysis is various pre-processing tasks that are done on the given data to convert it into the form that is suitable for mining. Fig.1 depicts the flow of the system i.e. the various steps to do with the given data to find the sentiment.

Firstly the tweets are collected from specified sourcethen data pre-processing and cleaning is done after that the tweets are broken into tokens then classification algorithms are applied to get the sentiment.

Today there are many tools available on internet that provides sentiment analysis or opinion mining. Some of them are as follows:

1) Twitrratr- www.twitrratr.com
2) Sentiment 140- http://www.sentiment140.com
3) Tweetfeel- www.tweetfeel.com
4) Opinmind- www.opinmind.com
5) Social Mention-www.socialmention.com

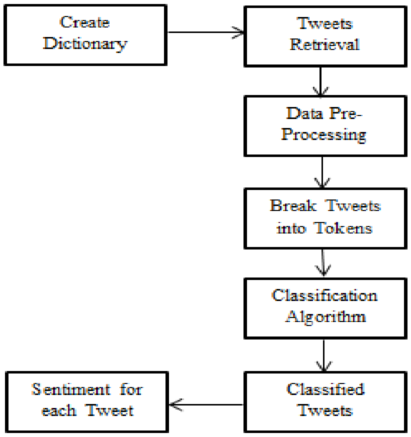

Figure 2: Flowchart of the system

Data Pre-Processing:

Filtering: Filtering supports to deliver the flexibility when user wants to design the data source structure such that a single mining structure can be created for that data. Filter can be used both in training and testing different data. Filters are produced to use only some part of the whole document, which helps to reduce the burden to build a different structure for every subset or part of data. Filters can be used by defining the length of word, content of words and many more.

URLs:Public use twitter for both data sharing and expressing feelings. But sometimes some users give the link of some website or any other thing. These links are useless in the sentiment analysis, so these links are removed and replaced with common word.

Usernames:Tweetsmainly include some usernames that are initiated by @ character these are also replaced by some word.

Duplicate or repeated characters:Public sometimes use some repeated characters in words like goooood which is same as good. Hence goooood is replaced by good.

Twitter slang removal: Twitter tweet is only of 140 characters so public use their own abbreviations like mrngfor morning, tmrw for tomorrow. This slang is replaced by their actual word like morning or tomorrow.

Stop-words removal:Stop-words are the mostly used words in a text document like articles, prepositions. These words occur most common in all the documents, but these are not necessary for text mining uses. These words should be removed.

Stemming:Stemming is a method used for reduction of words into their root form. There are many words in English which can be reduced to their base form like eat, ate, eaten all belong to base form eat.

V. TECHNIQUES FOR SENTIMENT CLASSIFICATION

Sentiment analysis can be done through two types of techniques as below:

1) Sentiment classification using supervised learning: Supervised learning is implemented by making a classifier. It requires two sets of documents for classification one is training set other is testing set. This method is also known as machine learning method. The classifier is trained by the given training dataset. The most frequently used algorithms are Support vector machines (SVM), Naive Bayes classifier and many others.

2) Naive Bayes: It is a supervised learning method that is widely used. It is called naive with the assumption that all data entered is independent of each other which is actually not possible. It is called Bayes as Thomas Bayes is an English scientist which gave Bayes Theorem. This is mainly applied in real difficulties like email spam detection. It is used when we have fewer resources like memory or CPU. It has low execution time.

3)Support Vector Machines (SVM): SVM is other classification method. It creates a hyper plane or set of hyper planes in high dimensional space such that the separation is maximum. It is also called maximum margin classifier. A hyper plane is a function like the equation for a line y=mx+b.

4) K-NN: K-NN stands for k-nearest neighbors. It is a supervised learning as it is provided with training dataset. It is a lazy learner. A lazy learner is that which doesn't do anything during training time only store the data. No classification is done when training data is stored. When unlabeled data is given as input to this algorithm then it classifies the data.

Pang Lee and Vaithyanathan [18] have wrote the first paper to take this method to categorize movie reviews into positive and negative classes.

2) Sentiment classification using unsupervised learning: In the unsupervised classification the text is classified into positive or negative by matching it with certain words or lexicons. The emotion significance for these words or lexicons is earlier well-defined. The document is checked and compared with positive and negative words. If a document or review has more positive words it is called positive if it has more negative words it is classified as negative. Turney [19] has done a great work in unsupervised learning. He used words "poor" and "excellent" as the kernel words then find out thecoordination of words centered on them. He successfully achieved 66% accuracy for the given movie reviews dataset.

VI. APPLICATION AREAS

Following are the various areas where sentiment analysis can be used:

1) E-commerce: Many websites provide summary of their products that allow users to submit their views. These views are beneficial for both other users and the product manufacturers.

2) Voice of customer (VOC): It is the market research technique that defines customer needs and expectations. Hence they define the reliability of the products.

3) Government: Government can view its strengths and weaknesses by the public reviews on the social websites on various social issues.

4) Marketing: Sentiment analysis helps the product manufactures to find out which customers are loyal and which are not and how to make new customers their loyal customers.

5) Politics: Before the actual election result the view of public can be analyzed by their comments or reviews on the social media. A no. of voting applications is available in market to analyze the view of public.

6) Blog analysis: Sentiment analysis can be effectively used to mine contentions in discussions and debate forums. It can be applied to analyze blog posts and perform subjectivity.

7) Stock Market Prediction: Sentiment analysis can be efficiently used to forecast events in stock market.

VII. CHALLENGES IN OPINION MINING

1) The product reviews or feedback can be in different languages like English, Hindi, and Punjabi etc. so it is difficult task to analyze the opinion of every language.

2) Instead of nouns words the verbs and adjectives are also considered as feature words which are not easy to classify.

3) The detection of fake reviews is also a challenging task.

4) The meaning of opinion words may changes according to the condition. Consider the example "He has a big house", here the word big is used in positive nature but in "There occurs a big earthquake" word big describes the negative term. To find the polarity of word according to situation is a challenging task.

5) To find out the synonyms in the comments is also a difficult task. For example if a person comments "she is looking beautiful" or "she is looking gorgeous", here the words beautiful and gorgeous expresses the same meaning.

6) People use a lot of abbreviations in their comments or feedback.For example, tmrw for tomorrow, lol for laugh out loud, sd for sweet dreams, gm for good morning, sry for sorry, b4 for before, plz for please and many more. So it requires a extra work to do with these abbreviations.

7) Mostly people use different style to write any comment or review. Same review may be seen in positive or in negative manner. So it is also a challenging task.

8) People sometimes use the words other than their root term. For example, damaged, damaging, damages are words used for the root term damage. Hence a proper stemming should be done to reduce these words into their root form.

REFERENCES

[1] BholaneSavitaDattu and Prof. Deipali V. Gore, "A Survey on Sentiment Analysis on Twitter Data Using Different Techniques", International Journal of Computer Science and Information Technologies, Vol. 6, 2015.

[2] RushleneKaurBakshi, Ravneetkaur, Navneetkaur and

Gurpreetkaur, "Opinion Mining and Sentiment Analysis", IEEE Third International Conference, 2016.

[3] PragyaTripathi, Santosh Kr Vishwakarma and Ajay Lala,"Sentiment Analysis of English Tweets Using Rapid Miner", IEEE International conference, 2015.

[4]Shubhi Mittal, AshnaGoel and RachnaJain,"Sentiment Analysis of E-commerce and Social Networking Sites",IEEE Third International Conference, 2016.

[5] Shachi H Kumar, "Twitter Sentiment analysis", CMPS 242 Project Report.

[6] TanuVerma, Renuand Deepti Gaur, " Tokenization and Filtering Process in RapidMiner", International Journal of Applied Information Systems (IJAIS), Volume 7– No. 2, April 2014.

[7]Xindong Wu, Vipinkumar, JoydeepGhosh,"Top 10 algorithms in Data Mining", Springer-Verlag London Limited 2007, Dec 2007.

[8]https://rayli.net/blog/data/top-10-data-mining-algorithms-in-plain-english/

[9] GautamiTripathi and Naganna S2, "Opinion Mining: A Review", International Journal of Information & Computation Technology, vol. 4, pp. 1625-1635, 2014.

[10] Nidhi R. Sharma, Prof. Vidya D. Chitre, "Opinion Mining Analysis and its Challenges", International Journal of Innovations & Advancement in Computer Science, vol. 3, Issue 1, April 2014.

[11] HaseenaRahmath P, "Opinion Mining and Sentiment Analysis – Challenges and Applications", International Journal of Application or Innovation in Engineering & Management (IJAIEM), Vol. 3, Issue 5, pp. 401-403, May 2014.

[12] DongSung Kim and Jong Woo Kim, "Public Opinion Mining on Social Media: A Case Study of Twitter Opinion on Nuclear Power", Advanced Science and Technology Letters, Vol.51 (CESCUBE 2014), pp.224-228, 2014.

[13] Rushabh Shah and Bhoomit Patel, "Procedure of Opinion Mining and Sentiment Analysis: A Study", International Journal of Current Engineering and Technology, vol. 4, No. 6, pp.4086-4090, December 2014.

[14] Sakshikalra, Rajesh Sachdeva and Anjali Dhawan, "A Review: Sentiment Analysis and Opinion Mining", International Journal of Research in Engineering and Applied Sciences (IJREAS), Vol. 6 Issue 10, pp. 16-21, October - 2016.

[15] Meghachauhan, "Opinion mining for effective product selection", International Journal of Advance Engineering and Research Development, vol. 3, Issue 4, pp. 429-433, April -2016.

[16] BlessySelvam, S.Abirami, "A Survey on opinion mining framework ", International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 9, pp. 3544-3549, September 2013.

[17] Diana Maynard and Adam Funk, "Automatic Detection of Political Opinions in Tweets", ESWC'11 Proceedings of 8th international conference on The Semantic Web, Pages 88-99.

[18] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," Proceedings of the ACL-02 conference on Empirical methods in natural language processing, vol.10, pp. 79-86, 2002.

[19] P.D. Turney, "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews." Proceedings of the Association for Computational Linguistics (ACL), pp. 417–424, 2002.