Sentiment Analysis of Twitter Data : A Survey

Radhi Desai
M.E Scholar, Computer Engineering Department, Sardar Vallabhbhai Patel Institute of Technology, Vasad, Gujarat, India

ABSTRACT

Sentiment analysis of Twitter data became a research tread the last decade. Among popular social networks portals, Twitter has been the point of attraction to several researcher in important areas like prediction of democratic several events, consumer brands, movie box-office, stock market, popularity of celebrities etc. The term sentiment refers to the feelings or opinion of person towards some particular domain. Analysis of sentiment (opinions) and its classification based on polarity is a challenging task. Other challenges are overwhelming amounts of information on one topic and they all are expressed on different ways. Lot of work has been done on sentiment analysis of Twitter data and lot needs to be done.There are many techniques for sentiment analysis. Supervised, unsupervised and combination of both of them.

Keywords: Sentiment analysis, Twitter, Data Mining

I. INTRODUCTION

Twitter has become very popular and has grown rapidly. An increasing number of people are willing to post their opinions on Twitter, as per the current report 313 million monthly active users per day and 500 million tweets per day that is considered as a valuable online source for opinions. But the main challenging task is extracting and analyzing the useful things from Twitter. The unstructured nature of the content and the natural language used to write these content added up the complexity more and it opened a new area of research called Opinion Mining and Sentiment Analysis. Twitter has been the point of attraction to several researcher in important areas like prediction of democratic several events, consumer brands, movie box-office, stock market, popularity of celebrities etc.[1]

“Sentiment Analysis is defined by the process of computationally identifying and categorizing opinions expressed in a piece of text, especially in order to determine whether the writer's attitude towards a particular topic, product, etc. is positive, negative, or neutral.” Sentiment Analysis is generally carried out in three steps. First, the subject towards which the sentiment is directed is found then, the polarity of the sentiment is calculated and finally the degree of the polarity is assigned with the help of a sentiment score which denotes the intensity of the sentiment. Sentiments can be classified at various levels: Aspects or feature level, sentence level and document level. Aspects or feature level sentiment classification classifies the sentiments based on the sentiments polarity of each aspects or feature about some target object and sentence level sentiment classification on the other hand classifies each sentence based on their sentiment polarity towards some topic. In document level sentiment classification the polarity of whole document is determined. It classifies the entire document into positive or negative or neutral class.[2]

II. SENTIMENT ANALYSIS LEVELS

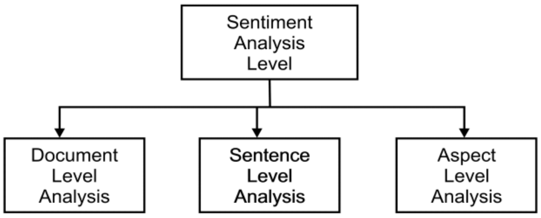

Figure 1. Sentiment Analysis Level

Document level: Sentiment classification the polarity of whole document is determined. It classifies the entire document into positive or negative or neutral class.

Sentence level: classifies each sentence based on their sentiment polarity towards some topic.

Aspect level: classifies the sentiments based on the sentiments polarity of each aspects or feature about some target object.

III. SENTIMENT ANALYSIS TECHNIQUES

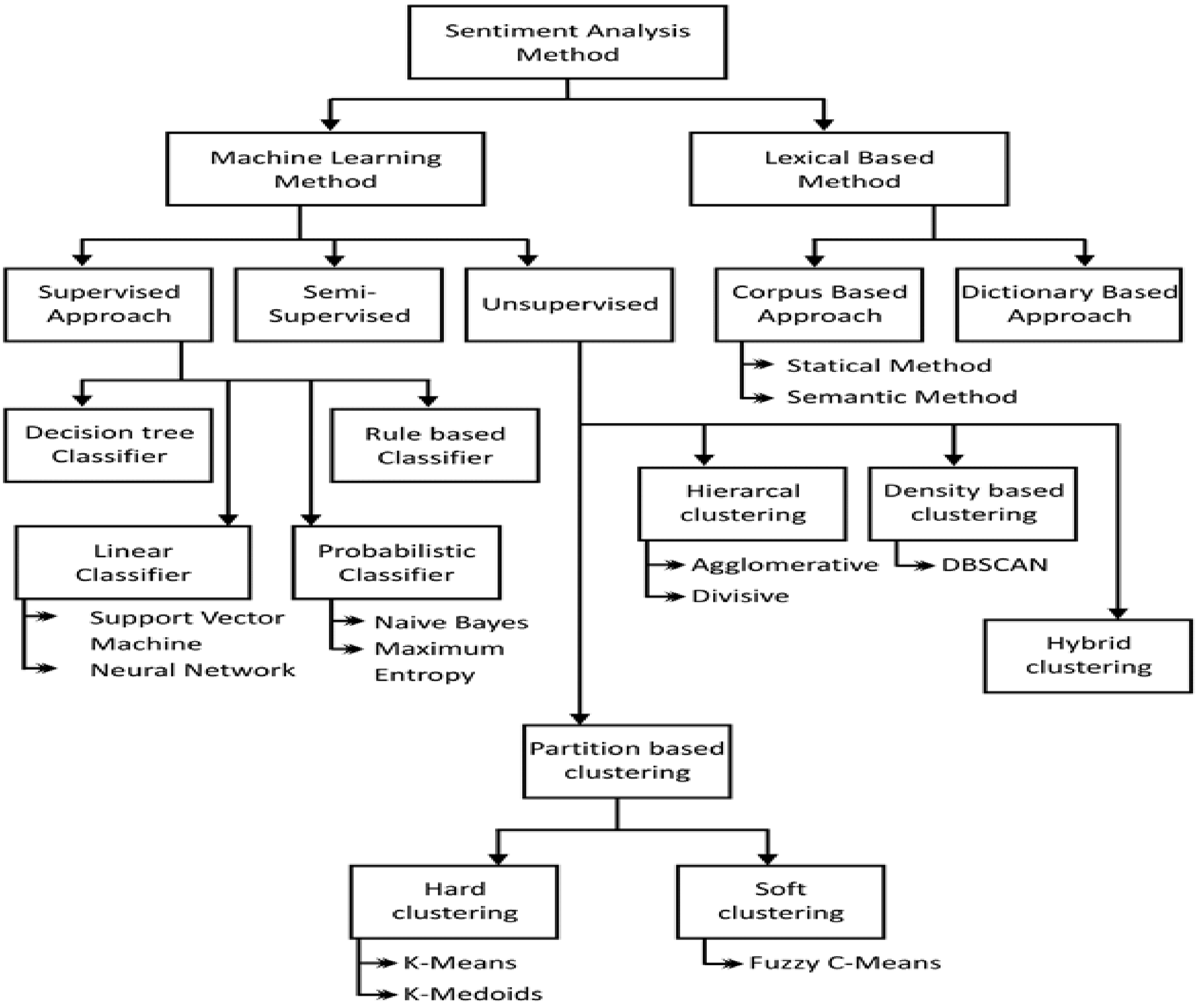

Figure 2. Taxonomy of sentiment analysis techniques

Generally, two techniques are used for opinion mining and sentiment analysis:

1) Lexicon based techniques.
2) Machine learning based techniques

In Lexicon based techniques, a sentiment dictionary with sentiment words are used for sentiment classification. The dictionary contains polarity of each word whether they are positive, negative and objective words. Polarity of the opinion words can be determined by matching those words with dictionary words.

In machine learning based techniques various machine learning algorithms are used for sentiment classification. Both supervised and unsupervised learning algorithm can be used to classify text. Supervised learning generates a model that can map inputs to desired outputs (also called labels), which are labelled by human experts according to some previously selected training samples. The most common supervised learning model is usually formulated as a two-class sentiment classification problem, that is, positive and negative. Since it is a textual classification problem, any supervised learning method can be applied, for example, Naive Bayes classification, and Support Vector Machines.

In contrast, in an unsupervised learning model, the labels are not known during training. As a typical unsupervised learning method, clustering, which tries to find the natural clusters in the data by calculating the distances or similarities from the centres of the clusters, is especially useful for organizing documents to improve information retrieval. For example, clustering methods can be used to automatically group the retrieved documents into a list of meaningful categories. The main advantage of clustering over classification is that it is adaptable to changes and helps single out useful features that distinguish different groups.[3]

IV. RELATED WORK

In current years, a voluminous amount of research has been conducted in the sentiment analysis domain. In supervised approach,In[4] authors presents three main phase of text mining utilized in application is pre-processing, processing and validation, Activities conducted in the pre-processing phase are folding, cleaning, stop-word removal, negation conversion, tokenization to the training data and the test data. In processing phase it performs weighting and classification using Naïve Bayes algorithm on the validated model. The process for measuring the level of accuracy generated by the application using 10-fold cross validation is done in the validation phase. The accuracy achieved by this model is 83% for 105 tweets.In [5] present the approach that combines machine learning and lexicon based approach. It tries to evaluate each emotion hearing word on the basis of it intensity. Instead of TF-IDF, it used AFINN lexicon dataset classify tweets into funny, happy, sad, angry tweets. SVM achieved highest accuracy followed by KNN and NB, but KNN was the most stable among the three classifiers. SVM proved to be superior to KNN and NB. However, the execution time was almost same.

In [6] author used Dictionary based approach to analyse data posted by user. This approach uses a predefined dictionary of positive and Negative words. SentiWord net is a standard dictionary used by most researchers today for sentiment analysis. Task of Polarity classification they mean the reviews collected are classified depending upon the emotions expressed as Positive, Negative and Neutral Various feature extraction methods are mentioned. In [7] authors combined Principal Component Analysis (PCA) with SVM in an attempt to perform dimensionality reduction. This paper shows two different case studies of entirely different social scenarios, US Presidential Elections 2012 and Karnataka Assembly Elections 2013. They conclude the conditions under which Twitter may fail or succeed in predicting the outcome of elections. Experimental results demonstrate that Support Vector Machines outperform all other classifiers with maximum successful prediction accuracy of 88% in case of US Presidential Elections held in November 2012 and maximum prediction accuracy of 58% in case of Karnataka State Assembly Elections held in May 2013. In [8] authors implement two different methods for subjectivity of sentences and then rule based system is used to find feature-opinion pair. System uses SentiWordNet approach and then it uses the method which is based on lexicon consisting list of positive and negative words. They present some rules which is very useful to identify the sentiments and after they apply Naïve Bayes for subjectivity analysis.

In unsupervised approach there are many research done. In [9] authors presents a novel fuzzy clustering model to analyse twitter feeds regarding the sentiment of a particular brand using the real dataset collected over a period of one year. Then a comparative analysis is made with the existing partitioning clustering techniques namely K-Means and Expectation Maximization algorithms based on metrics namely accuracy, precision, recall and execution time. According to the experimental analysis, the proposed approach is tested to be practicable in performing high quality twitter sentiment analysis results. In [10] authors present approach that give the better performance than bag-of-words method. In this approach both unigrams and bigrams are used as features to cluster texts using K-means clustering, which makes texts having similar words clustered in same cluster. Then after Naïve Bayes, a probabilistic classifier is applied. This method reduce the sparsity problem for sentiment analysis and K-Means improve the performance. In [11] The objective of this study is to develop opinion classification system using Maximum entropy (ME)) and K-Means Clustering (KMC). Opinion to be classified was Indonesian textual comments from academic questionnaire. Classification was conducted into two classes, i.e. negative opinion and positive opinion. Data contained of 2000 comments that was sampled as multi domain opinion, represented many objects such as lecturer, class room, etc. Features used for classification was selected from word in the opinion text. The weighting scheme that they used for clustering was TF/IDF. The results show that K-Means Clustering gives better performance as compared with ME in averages about 3% precision. KMC also perform faster than ME about 25 msec using 2000 text opinion. In [12] density based clustering, clusters are considered as dense regions of objects that separated by regions of low density. The tweets are segmented into a number of meaningful segments. The local context and global context along with their stickiness value is considered in segmentation process. DBSCAN algorithm with Jaccard similarity measure is used for clustering of tweets. In [13] authors apply TF-IDF scheme, than to improve the result they suggest voting mechanism. That define as, for each document vector, there are 20 clustering results which are positive or negative. Each of those results regarded as a vote. One document will be determined as positive if it obtains more than 10 positive votes, otherwise it will be regarded as negative. After that they suggest tern scores form WordNet and calculating average value of the score for each document, and partitioning by medium value. It overcomes the challenges of low accuracy and instability of results. In [14] authors propose the system that input is the collection of documents consisting of multi topic segments taken from web. SentiWordNet has been used to calculate the segment score of the segments within the documents. Based upon the segment score segment based clustering is performed on the intra-document level. Once it done with intra document segment based clustering then k-means approach is applied to the entire collection of documents to perform inter-document clustering in which the similar segments of various documents will be clustered under a single cluster. This technique would help in efficient organization of multi topic documents into their corresponding clusters.

Semi-supervised approach is a class of supervised learningtasks and techniques that also make use of unlabeled data for training – typically a small amount of labeled data with a large amount of unlabeled data.. In [15] authors applying sentiment analysis and machine learning methods to study the relationship between the online reviews for a movie and the movies box office revenue performance. The paper shows that a simplified version of the sentiment-aware autoregressive model can produce very good accuracy for predicting the box office sale using online review data. Document level sentiment analysis is used which consists of Term Frequency (TF) and Inverse Document Frequency (IDF) values as features along with Fuzzy Clustering which results in positive and negative sentiments. This lead to the creation of a simpler model which could be more efficient to train and use. In addition, a classification model is created using Support Vector Machine (SVM) Classifier for predicting the trend of the box office revenue from the review sentiment. In [16] authors proposed 'cluster-than-prediction' approach,first cluster the tweets using k-means algorithm and then perform classification using Classification Trees(CART-Classification and Regression Trees). This clustering operation makes the data domain-specific, which results in creation of better predictive models which has led to more accurate classification of sentiments of a recently launched product. The objective of this paper is to effectively perform sentiment analysis of a recently product 'iPhone 6s' developed by the company Apple. The accuracy of classification of the sentiment for proposed approach is highest. It means that using the 'cluster-then-predict' approach, 74.85% of the tweets in test data are correctly predicted as showing the actual sentiment of the users towards the product 'iPhone 6s. This accuracy value is significantly higher than CART, on which the 'cluster-then-predict' approach is based. The comparative study is done with Support Vector Machine, CART, Naïve Bayes and Random Forest.

V. PHASES OF SENTIMENT ANALYSIS

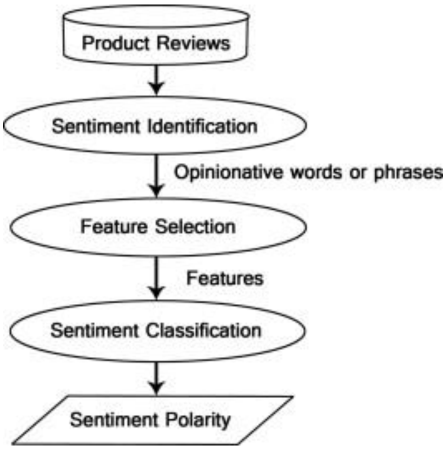

Figure 3. Phases of sentiment analysis

- ✓ First, the system takes the input as reviews.
- ✓ Then pre-processing step will be performed for tokenization, after that stop-word removal is performed- in that words are remove which do not carry meaning like "I", "you", "a', "the".
- ✓ After that POS tagging is applied to reviews. And the adjective and adverbs are retrieved. Then apply the weighting scheme named TF-IDF- that is intended to reflect how important a word is to a document in a collection or corpus.
- ✓ Then after various supervised or unsupervised techniques are apply on the reviews.
- ✓ After that the polarity(positive, negative or neutral) is detected.

VI. CONCLUSION

From the above analysis I concluded that according to different strategies, we can find the interest of the customer, opinions of customer related to particular product. The main challenge is to find out the polarity- positive, negative or neutral reviews of customer according to products. The analysis which has been made from the above techniques will analyze the user feelings, emotions etc.

VII. REFERENCES

[1]. Mitali Desai, Mayuri Mehta, "Techniques for Sentiment Analysis of Twitter Data- A Comprehensive Survey", IEEE, pp.149-154,2016

[2]. Jatinder Kaur, "A Review paper on Twitter Sentiment Analysis Techniques", International Journal for Research in Applied Science & Engineering Technology, vol.4, pp.137-141, October-2016.

[3]. Baojun Ma, Hua Yuan and Ye Wu, "Exploring performance of Clustering methods on Document Sentiment Analysis", Journal of Information Science(JIS), December 9, 2015.

[4]. Liza Mikarsa, SherlyNoviantiThahir, "A Text Mining Application of Emotion Classifications of Twitter's user using Naïve Bayes Method", IEEE, 2015

[5]. Govin Gaikwad, Prof.Deepali J Joshi, "Multiclass Mood Classification on Twitter using Lexicon Dictionary and Machine learning Algorithms", IEEE

[6]. Prerna Mishra, Dr.RanjanaRajsinh, Dr. Pankaj Kumar, "Sentiment Analysis of Twitter Data: Case Study on Digital India",IEEE, 2016

[7]. MalharAnjaria, Ramohana Reddy Guddeti, "Influence Factor based Opinion Mining of Twitter data using Supervised Learning", IEEE, 2014

[8]. PurtataBhoir, ShilpaKolte, "Sentiment Analysis of Movie Reviews using Lexicon Approach", IEEE, 2015

[9]. Hima Suresh, Dr.Gladston Raj. S, "An Unsupervised Fuzzy Clustering Method for Twitter Sentiment Analysis", IEEE,2016.

[10]. Yunchao He, Chin-Sheng Yang et al, "Sentiment Classification of Short Texts based on Semantic Clustering", IEEE,2015.

[11]. Amir Hamzah, NaniekWidyastuti, "Opinion Classification using Maximum Entropy and K-Means Clustering", IEEE,2016

[12]. AnumolBabu, Rose V Pattani, "Efficient Density Based Clustering of Tweets and Sentimental Analysis based on Segmentation", International Journal of Computer Techniques, vol.3, pp.53-57, May-June,2016

[13]. Gang Li, Fei Liu, "A Clustering-based Approach on Sentiment Analysis", IEEE, 2010

[14]. Rupesh Kumar Mishra, Kanika Saini, Sakshi Bagri," Text Document Clustering on basis of Inter passage approach using K-Means", IEEE, 2015

[15]. Nagamma P, Pruthvi H.R et al, "An Improved Sentiment Analysis of Online Movie Reviews based on Clustering for Box-Office Prediction", IEEE, 2015

[16]. RishabhSoni, K. James Mathai, "Effective Sentiment Analysis of a Launched Product using Clustering and Decision Tree", International Journal of Innovative Research in Computer and Communication Engineering, vol.4, pp.884-891, January 2016.