OPEN ACCESS INTERNATIONAL JOURNAL OF SCIENCE & ENGINEERING

A SURVEY ON TWITTER SENTIMENT ANALYSIS

Swarupa Kulkarni[1], Priyanka Kedar[2]

PG Student, Computer Department, DPCOE, Maharashtra, India.[1]

Professor, Computer Department, DPCOE, Maharashtra, India.[2]

Abstract: Millions of people share their thoughts day by day on social networking site like Twitter, in the form of tweet. A tweet is short and basic way of expression. But great insights can be obtained from these short and highly unstructured tweets. Now days various organizations are using Twitter to know public opinion about the services or products which they provide. Sentiment analysis is the process of determining a sentiment of given data, it's one of the area of text data mining and NLP. There are various aspects to perform sentiment analysis of Twitter data. In this survey paper, we have done comparative study of different approaches and techniques to extract sentiment from tweets.

Keywords - *Sentiment Analysis, NLP, Social Networking, Twitter.*

----------∴∴∴----------

I INTRODUCTION

Now day's social networking sites like Facebook, Instagram, and Twitter have become so much popular. Innovations of smart phones and as they are easily available to everyone contributed in success of social networking sites. One of the interesting area is sentiment analysis which is based on text analysis, natural language processing, computational linguistics. It aims to determine the opinions from piece of text. It's being used in many sectors such as politics, finance, sociologyetc.

Most of the data present on social networking sites or we can even say data from the internet is unstructured. It is very difficult task to extract opinion from an unstructured data. Thus to apply sentiment analysis techniques on data, prepossessing of data plays an important role. For example, in case of twitter data stopwords can be removed as they do not contribute in terms of overall opinion of the text. Number of prepossessing steps are carried out before applying any sentiment analysis techniques on data.

Opinions of others matter when a decision is needed to make. As we discussed earlier World Wide Web, social networking sites and smartphones have brought peoples across the world close to each other. People express their views about anything they want on these sites. Here sentiment analysis plays crucial part as it can be used by various sectors. Politicians can use it to know if the people are happy about the latest policy, organizations can use sentiment analysis to know opinion of people about the latest product they have launched. They can use this insights to make changes, in recent time we have seen that many IT companies are using sentiment analysis to know opinion of their employees about organization policies.

II SENTIMENT ANALYSIS

The area where opinions or views of peoples are extracted from text by using various techniques is called as sentiment analysis. There is very minor difference between sentiment analysis and opinion mining, in opinion mining we extract opinion about particular topic or event from given data. While in sentiment analysis we extract the sentiment of given text, it might be positive, negative or in some cases neutral.

Levels of Analysis

Sentiment analysis is categorized into three different levels as below:

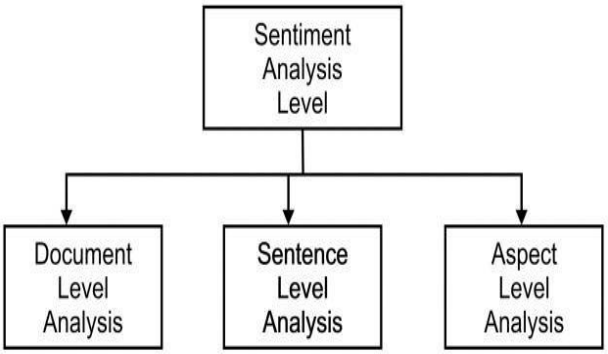

Figure 1. Sentiment Analysis Levels

A. Document Level Analysis: It is simplest form of classification. Whole document is considered as single source of information. It is also assumed that document has information about single object only for which we are doing sentiment analysis such as movie or any product. If document has information about many objects then it reduces the accuracy of classification. If any irrelevant sentences are present then they are eliminated before analysis.

B. Sentence Level Analysis: It is the most fine grained analysis. Here every sentence from data source is considered as separate unit. Polarity of each sentence is calculated and they are classified into groups such as positive, negative or neutral.

C. Entity/Aspect Level Analysis: We cannot get peoples dislikes and likes by using sentence level analysis or document level analysis. Entity/Aspect level analysis gives us detailed analysis. The goal is to find the aggregate of sentiment based of entities present in given data source or some aspect of it.

III SENTIMENT ANALYSIS TECHNIQUES

There are two main techniques used in sentiment analysis:

1. Lexicon based techniques.

2. Machine learning based techniques.

Lexicon based techniques make use of dictionary. Dictionary contains sentiment words along with their polarity (positive, negative or neutral). These sentiment words are used for classification. Words from data source are matched with these sentiment words and polarity of is decided accordingly.

.

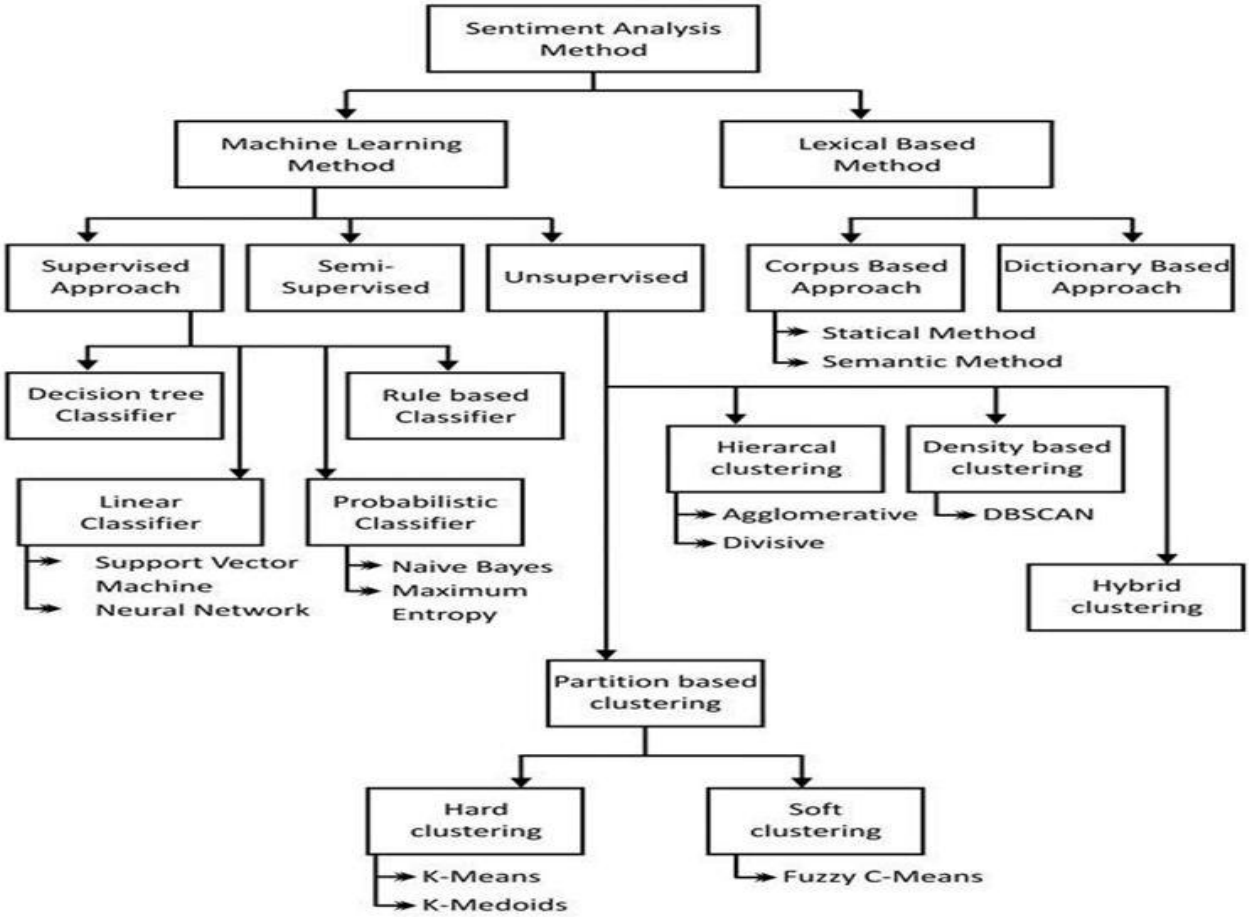

Figure 2. Sentiment Analysis Techniques

In Machine learning based approach different machine learning algorithms are used for classification of text. There are mainly two types of machine learning algorithms, supervised and unsupervised. In supervised algorithms aim is to find out a function which will map input data with its labels. Common problem will be binary classification, it will have only two labels positive and negative. We can use algorithms such as Naïve Bayes or Support Vector Machines. Based on the training set which has both data and corresponding label, machine learning model is trained. Then we can provide actual data to get labels for it.

In contrast to supervised machine learning algorithms, labels are not provided as an input to unsupervised machine learning algorithms. Common algorithm is clustering. It tries to find out clusters among the provided data, by calculating distance among its objects or similarities of object from center of cluster. Main advantage over supervised algorithms is that it can adapt changes in data source, also it helps to find out features that distinguish different clusters

IV RELATED WORK

Xing Fang et. al. [1] worked on the issue of sentiment polarity classification. It is one of the basic problem in sentiment analysis. They used product review data from the amazon website. They used sci kit learn framework, which is an open source framework available in python having implementation of machine learning algorithms. They used algorithms such as Random Forest, Support Vector Machines.

Geetika Gautam et. al. [2] also worked on classifying customers reviews. They used labelled dataset which had tweets from twitter. They used Naïve Bayes, SVM and Max Entropy algorithms for classification. They obtained better results with Naïve Bayes. Later on they applied WordNet from semantic analysis to further improve results. To train machine learning model they used Python and NLTK.

Neethu M S et. al. [3] analyzed twitter data related to different electronic products. They used 8 features such as number of negative hash tags, presence of negation, number of positive keywords, emoticon, pos tag, number of negative keywords, special keyword and number of positive hash tags. They implemented Naïve Bayes and SVM classifiers using MATLAB.

Akshay Amolik et. al. [4] proposed an model for sentiment analysis for upcoming Bollywood and Hollywood movie reviews. They used Feature-Vector along with Naïve Bayes, SVM classifiers. Naïve Bayes produced more accurate results than SVM classifier.

Agarwal et al. [5] proposed an three way model to classify tweets into three categories positive, negative and neutral. They used tree kernel based model, feature based model and unigram model. In tree kernel based model they represented tweets in tree format, for feature based model they used 100 features and in unigram model they used over 10,000 features.

Davidov et al., [6] developed an model which used user defined hash tags from tweets to classify an tweet. Different feature types such as single words, punctuations and patterns are used to create feature vector by combining all these features. Theyused K Nearest Neighbor algorithm to classify tweets at the end.

Po-Wei Liang et.al. [7] used streaming Twitter API to extract the live tweets from Twitter. They classified training data into three categories such as camera, mobile and movie. They used unigram Naïve Bayes to classify tweets. They also removed useless features by using Chi square feature extraction method and Mutual Information.

Pablo et. al. [8] implemented Naïve Bayes algorithm in two different ways. Baseline which classified tweets into three categories such as positive, negative and neutral. Binary where neutral tweets were neglected and classification is done only in two categories positive and negative. Features which were used by classifiers are Polarity Lexicons, Multiword, Lemmas (verbs, nouns, adverbs and adjectives).

Turney et al [9] developed an model using bag of words approach. Relationship between different words from a document is not considered. Sentiment of every word is determined and later on some aggregation functions were used to determine the sentiment of whole document.

Kamps et al. [10] proposed an model by using WordNet lexical database. Emotional content of word is determined along different dimensions. Distance metric is developed on WordNet and sematic polarity of adjectives iscalculated.

Xia et al. [11] developed an model using ensemble approach. Wordrelations and part of speech is used as feature sets. They used SVM, Naïve Bayes and Max Entropy classifiers. Meta classifier combination, weighted combination and fixed combination ensemble approaches are used to get better results.

ZhunchenLuo et. al. [12] discussed various challenges in opinion mining from Twitter. Spam and different languages are discussed. Effect of multiple languages on opinion mining is also highlighted.

V CONCLUSION

Analysis of twitter data is done with different approaches to extract opinion or sentiment from tweet. In order to conduct sentiment analysis one must understand twitter data and its structure. Preprocessing of data plays important part in development of machine learning model as tweets are highly unstructured. In this paper we discussed various approaches to carry out sentiment analysis of tweets. Accuracy of machine

learning models can be increased by using ensemble approaches. We mainly focused on analysis of tweets in English language, analyzing tweets from different languages pose its own challenges.

REFERENCES

[1] Fang, Xing, and Justin Zhan. "Sentiment analysis using product review data." Journal of Big Data, 2015.

[2] Gautam, Geetika, and Divakar Yadav. "Sentiment analysis of twitter data using machine learning approaches and semantic analysis." Contemporary computing (IC3), 2014 seventh international conference on. IEEE, 2014.

[3] Neethu, M. S., and R. Rajasree. "Sentiment analysis in twitter using machine learning techniques." Computing, Communications and Networking Technologies (ICCCNT), 2013 Fourth International Conference on. IEEE, 2013.

[4] Amolik, Akshay, et al. "Twitter sentiment analysis of movie reviews using machine learning techniques." International Journal of Engineering and Technology, 2016.

[5] Agarwal, B. Xie, I. Vovsha, O. Rambow, R. Passonneau, "Sentiment Analysis of Twitter Data", In Proceedings of the ACL 2011Workshop on Languages in Social Media, 2011.

[6] Dmitry Davidov, Ari Rappoport." Enhanced Sentiment Learning Using Twitter Hashtags and Smileys", Coling, 2010.

[7] Po-Wei Liang, Bi-Ru Dai, "Opinion Mining on Social MediaData", IEEE 14th International Conference on Mobile Data Management,Milan, Italy, 2013.

[8] Pablo Gamallo, Marcos Garcia, "Citius: A Naive-Bayes Strategyfor Sentiment Analysis on English Tweets", 8th InternationalWorkshop on Semantic Evaluation (SemEval 2014), Dublin, Ireland, 2014.

[9] P. D. Turney, "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews," 40th annual meeting on association for computational linguistics, pp. 417–424, Association for Computational Linguistics, 2002.

[10] J. Kamps, M. Marx, R. J. Mokken, and M. De Rijke, "Using wordnet to measure semantic orientations of adjectives," Proceedings of the Fourth International Conference on Language Resources and Evaluation, 2004.

[11] R. Xia, C. Zong, and S. Li, "Ensemble of feature sets and classification algorithms for sentiment classification," Information Sciences: an International Journal, 2011.

[12] ZhunchenLuo, Miles Osborne, TingWang, An effective approachto tweets opinion retrieval", Springer Journal, 2013.