Opinion Mining Using Machine Learning Approaches

Prasanna Kumar Lakineni
Assoc.Prof & HoD
Dadi Institute of Engineering & Technology, Visakhapatnam
N.Viswanadha Reddy
Assistant Prof,
Dadi Institute of Engineering & Technology, Visakhapatnam
Sampathirao Suneetha
Research Schola
Andhra University

Abstract:

Opinion mining and Sentiment analysis is highly demanding field on social media. This attracts a large community of researchers to extract mindset of people which varies from time to time. Sentiments of end users that are expressed on massive amount of user data generated from Social platforms like E-commerce, social media sites, micro blogging sites has great influence on the readers. There is a large requirement of new techniques and algorithms to analyze the unstructured data from the social media. For this the sentiment analysis and machine learning techniques have been merged because of machine learning models are effective due to their automatic learning capability. The main aim of this paper is to provide a brief introduction to sentiment analysis area with machine learning techniques.

Keywords: Machine Learning, Sentiment Analysis.

I.INTRODUCTION

World Wide Web is a virtual environment for people to share their experiences through electronic communication media. Sometimes it is also called as 'Electronic word of mouth' [EWOM][1]. People or users expressing their feelings, emotions, moods, attitudes on social platforms, which can be extracted and analyzed to know their opinions from different perspective. It is interesting and estimated that around 3.4 billion internet users are there out of which 2.3 billion of them are active users of social media.

Opinion mining and sentiment analysis can be used interchangeably even then some researchers identified and proved slight differences in between those fields. Natural language processing can be classified into text mining and opining mining. The Process of Sentiment Analysis is shown in Figure 1.

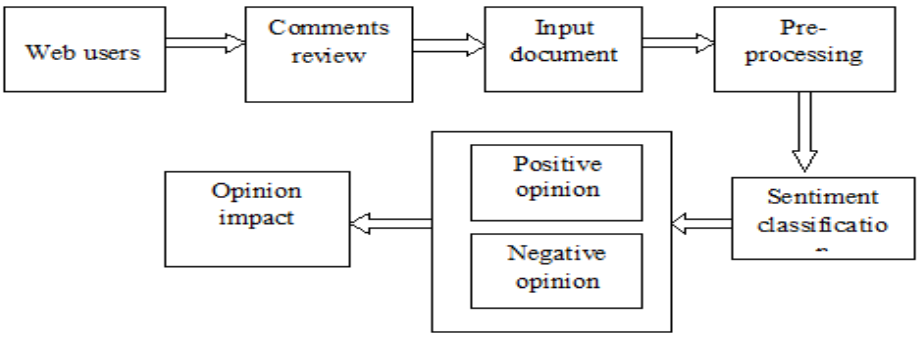

Figure 1:

Why sentiment analysis?

Social media have a large influence on society as well as business. The social impact can be both behavioral any psychological. The process of applying various methods to know sense of the social data is called social data analytics.

In this paper we have survived various types of techniques and algorithms available for sentiment analysis. This can be classified in two categories as lexicon orientation and machine based approaches.

There are several approaches for sentiment analysis. In which machine learning approach uses machine learning algorithms for classifying the data. Lexicon-based approach used to determine the sentiment polarity using a dictionary of positive and negative words.

Lexicon based or corpus based techniques

These techniques are based on decision trees such as k-Nearest Neighbors (k-NN), Conditional Random Field (CRF), Hidden Markov Model (HMM), Single Dimensional Classification (SDC) and Sequential Minimal Optimization (SMO), related to methodologies of sentiment classification

Machine learning based techniques

This type of techniques is implemented by extracting the sentences and aspect levels. The features consist of Parts of Speech (POS) tags, n-grams, bi-grams, uni-grams and bag-of-words. Machine learning contains three flavors at sentence and aspect, i.e., Nave Bayes, Support Vector Machine (SVM) and Maximum Entropy.

II.LITERATURE REVIEW

A. Levels of sentiment analysis:

Analysis can be done in four levels, which are at document level, aspect level, sentence level and at concept level as shown in Figure. 2.

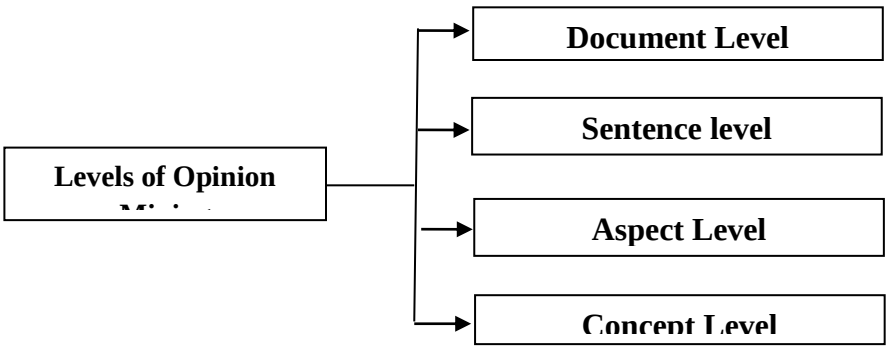

Figure 2.

1. **Document level:**

   At document level, the whole document can be considered as one entity and different sentiment analysis; tasks are studied on this entity by applying algorithms. Most of the techniques used in this level are based on the supervised learning approaches.

2. **Sentence level:**

   At sentence level, every sentence will be taken as a standalone piece on which process is carried out at each sentence to find out the sentence is subjective or an objective. Polarity classification can be used to identify user's sentiment expression as positive or negative. The sentence level of sentiment classification assumes one sentence expresses a single opinion from a single opinion holder.

3. **Aspect level:**

   Document level or sentence level opinion classification is useful but they are insufficient to provide the required details for the application.

   The aspect level aims on opinion itself instead of considering at the constructs of the documents. Aspect level analysis can be categorized into two sub tasks. One is aspect extraction and other is aspect classification.

   An example of this aspect-level sentiment analysis is by sentence: '*I love star trek but I hate star wars*'. Two sentiments are here, i.e. Love and hate, and two aspects, i.e. star trek and star wars.

4. **Concept Level:**

   Concept level opinion mining is based on deep level learning where natural language (NL) understanding of text is important. Concept level opinion is a new method through we can estimate emotions based on the inference of semantic information about sentiments and emotions associated with NL.

B. Machine Learning Approaches:

Machine learning is one of the emerging technologies, which attracts large community of researchers, due to its accuracy and adaptability. Machine learning can be applied on sentiment feature extension as well as text classification.

Machine learning approaches can be classified into supervised and unsupervised. Various machine learning approaches show in Figure 3.

Supervised learning approach can further be classified into decision tree classifier, linear classifier, and rule-based probabilistic classifier. Further linear classifier can be classified into support vector machines and neural networks. Probabilistic classifier can be classified into naïve Bayes, Bayesian network, and maximum entropy. Naïve bayes classifier is simplest algorithm among others which gives best accuracy in its results. This can be used at document level

classification. Kang and Yoo et al. proposed an improved version of this approach [2], through this approach positive sentiment accuracy can be increased by 10% to that of negative sentiment.

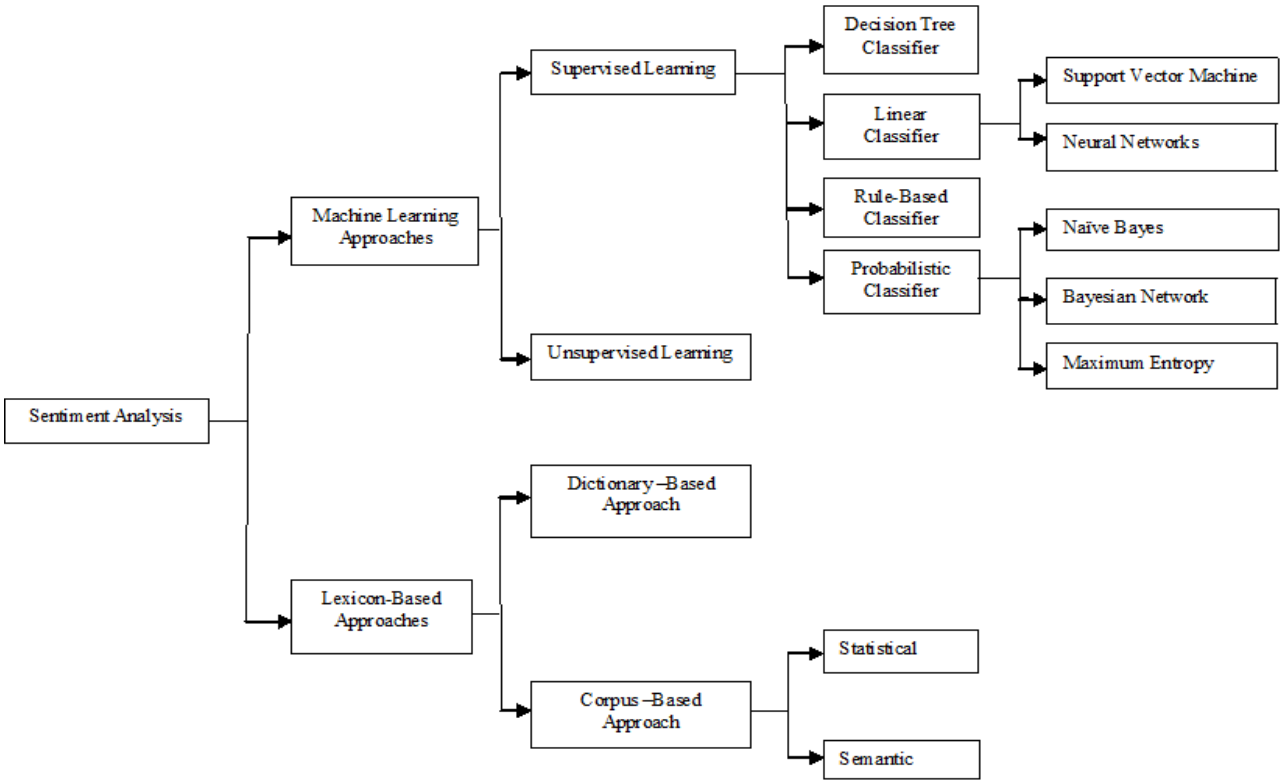

Figure 3

Another supervised approach is maximum entropy classifier, which works on the probability that a given entity belongs to a certain class basing on given context which maximizes the entropy of the total classification model. Linear based classifier such as support vector machine (SVM) which works on the principal that how to know linear separators in this space that best splits data into classes. Usually text data are best suitable for SVM classification due to sparse nature of text, where some features are irrelevant, but using SVM this data can be organized into linearly separable categories. Various experiments were carried out by Zhang et al [2]. to find the effect of size and representations of a feature on performance of classification.

Another one more important approach of classification is neural network. In this a neuron is a fundamental unit of neural network. The inputs providing to these units of are described by a vector where vector shows the frequencies of word in the i^{th} document over a line x_i, which can be shown in figure3.

The advantage of neural network is that they are data driven self adaptive methods. In which they can adjust themselves to the features without any functional specification. Neural network is a best machine learning approach.

III. STEPS INVOLVED IN MACHINE LEARNING TECHNIQUES

Due to more popularity of social media and social media text in particular small posts and comments in face book and twitter throwing more challenges which require new techniques and measures.

The important steps involved in machine learning approach can be explained as follows.

Step1-Data Collection:

Data can be collected from various social media sites like blogs, web 2.0, social networking sites basing on the area of application.

Step2-Preprocessing:

The collected data is cleaned and prepare for sending into the classifier. Cleaning means extraction of keywords and symbols.

Step3-Training Data:

A class label will be given a collection of data. This class label is a trained model. This data will be feed to the algorithm for learning purpose.

Step4-Classification:

This is the main step of the entire process based on the requirement of application, SVM or Naïve Bayes will be used for analysis. After computing the training this classifier is to be deployed to the real time data for extraction of sentiment involved in texts.

Step5-Results:

The final results can be plotted in the form of pictorial representation using charts, graphs, etc. performance tuning is to be done before the release of the algorithm.

III. TOOLS USED IN TEXT MINING

They are number of tools available for text mining. Text mining is the initial stem for Sentiment analysis. Following table.1 will show us different text mining tools available in this domain.

S.No	Name of the text mining tool	Purpose	Web link
1	Rapidminer	TF-IDF score of words	www.rapidminer.com
2	GNU Aspell	Spell checker	http://aspell.net/
3	Lancaster stemming algorithm	Stemmer	http://textminingonline.com/tag/lancaster-stemmer
4	POS tagger	Twitter POS tagger	www.cs.cmu.edu/~ark/TweetNLP/
5	Snowball	English stemmer	https://pypi.python.org/pypi/snowballstemmer

6	Stanford Log-linear Part-Of-Speech Tagger TweeboParser	POS tagger Tweet dependency parser	http://nlp.stanford.edu/software/tagger.shtml https://github.com/ikekonglp/TweeboParser
7	TweetMotif	Tokenization of tweet	https://github.com/brendano/tweetmotif
8	TweetNLP	Twitter natural language processing	www.cs.cmu.edu/~ark/TweetNLP/

IV. PERFORMANCE MEASURES

After preparation of the classifier for sentiment analysis, the trained classifier needs to be validated with cross fold validation [6]. The performance of the classification model can be determined with the following measures.

a) Accuracy:

It is a measure of the correct predictions over the total number of predictions. Usually accepted accuracy is in the range of 70% to 90%.

b) Precession:

This can be used to check how accurately the model makes predictions against each class. Precision is measured by number of total correct predictions over true positive and true negative examples.

c) Recall:

This measure shows the completeness of the model for each class. It is measured by the fraction of number of correct predictions by total number of true positives and false negatives.

d) F-Score:

F-Score is a combination of precision and recall. Its range is 0.0 to 1.0. 1.0 would be the perfect. F-Score is very useful. The formula for calculating F-Score [4]is:

$$F = \frac{2 * precision * recall}{Precision + Recall}$$

V. APPLICATIONS OF SENTIMENT ANALYSIS

Opinion mining is used to express and evaluate the expressed sentiments of the users of the internet on the social websites. These can be many issues in the social websites regarding political agendas, Product reviews, movies reviews, cultural affairs etc [6]. Following are the major areas of applicability of SA.

a) Opinion mining in the areas of Commercial Products of a company.
b) Opining mining of the people of the country on political parties.

c) Opining mining in the area of Stock market and Stock market forecasting.
d) Opining mining in the area of Social reviews on different contexts.

VI. MODELS OF TEXT DATA REPRESENTATION IN VECTOR FORM

Text data representation in vector form lies at the core of Machine Leaning Techiniques, assigning a mathematical object to each and every word in the collected text is often a vector of real numbers[8].Many researchers have represented text as vectors, tested and compared to identify the worthiness of different models for solving specific problems related to the text processing for sentiment analysis.

Lets us review several text representation models existed.

1) Bag-of-Words
2) Bag-of-N-Grams
3) Part-of-speech tagging

VII. CONCLUSION

In this paper we have studied all aspects of Sentiment Analysis and opinion mining with machine learning techniques. The main aim of this paper is to study the well know methods of machine learning techniques in the field of SA and their challenges. Sentiment Analysis on social media is vast area of research which requires necessary background that we have covered in this paper. Largest number of researches is working in this area of data analytics which has large potential in the social media market. Many issues are related with text based data and unstructured data. Supervised machine Learning methods are mostly accepted and adopted for sentiment analysis by large community of researchers. In future we are going to work on real time data with supervised machine learning algorithms in order to get fast response on both textual data and unstructured data on the social media.

VII. REFERENCES

[1] Vidisha M. Pradhan ,Jay Vala, Prem Balani,"A survey on Sentimental analysis for Opinion Mining", International Journal of Computer Applications (0975 – 8887) Volume 133 – No.9, January 2016

[2] Z. Zhang, Q. Ye, Z. Zhang, Y. Li, "Sentiment classification of Internet restaurant reviews written in Cantonese", Expert Syst. Appl. 38 (2011)7674–7682

[3] Isidro Peñalver-Martinez, Francisco Garcia-Sanchez , Rafael Valencia-Garcia," Feature-based opinion mining through ontologies", Expert Systems with Applications-2014.

[4]. F. Bobillo, U. Straccia,"Fuzzy ontology representation using OWL 2",Approx. Reason., 52 - 2011.

[5] H. Kang, S. Yoo, D. Han," Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews", Expert Systems with Applications, 39 -2012

[6] Joseph Turian, Lev Ratinov, and Yoshua Bengio. Word representations: a simple and general method for semi-supervised learning. In Proceedings of the 48th annual meeting of the

association for computational linguistics, pages 384–394. Association for Computational Linguistics, 2010
[7] G.Vinodhini, RM.Chandrasekaran, "Performance Evaluation of Machine Learning Classifiers in Sentiment Mining", International Journal of Computer Trends and Technology (IJCTT) , Volume 4, Issue 6,June 2013, Page 1783-1786.
[8] B. Pang and L. Lee. A sentimental education:Sentiment analysis using subjectivity summarization based on minimum cuts. In Proc. of the ACL,pages 271–278. ACL, 2004.