Sentiment Analysis and its Challenges

Chandni
M.tech (CSE),
Banasthali Vidyapith,
Rajasthan, India

Nav Chandra
B.tech (CSE)
Kurukshetra University,
Haryana, India

Sarishty Gupta
M.tech (CSE),
NIT, Jalandhar
Punjab, India

Renuka Pahade
MSc[Eng] (Software Engineering),
The University of Sheffield,
United Kingdom

***Abstract:* An essential part of gathering of information has always been completed through the opinions of other people means what they think. And with the increasing popularity of opinion resources like personal blogs and review sites people can easily gather information and analyze the opinions of other people. Sentiment Analysis is a current ongoing field of research. Sentiment Analysis deals with the computational study of human thoughts or opinions, emotions and attitudes toward an object. This paper deals with recently proposed algorithms, approaches and its applications briefly. The main goal of this paper is to provide a complete description of Sentiment Analysis or Opinion Mining technique and the challenges faced by it. We also include the social issue regarding manipulation, privacy and economic impact faced during the development of information seeking services.**

Keywords:* *Sentiment analysis, lexicon, Sentiment classification technique, Feature extraction, Naïve Baye's, Maximum Entropy.

I. INTRODUCTION

Sentiments are feelings, thoughts; emotions of an individual for a particular event or topic e.g. love/hate, bad/good. Sentiment Analysis is defined as a computational study of human's thoughts or opinions, emotions and attitudes toward an object. An object can be an individual, topic or an event. Sentiment Analysis can also be called as Opinion Mining, Sentiment Mining, Opinion Extraction, and Subjectivity Analysis. Opinion Mining and Sentiment Analysis are replaceable because researchers argue that Opinion Mining and Sentiment Analysis are somewhat different to each other [1].The main goal of Sentiment Mining is to analyze the sentence and examines the sentiments shown in sentence and finally determines the polarity of the sentences and is shown in figure 1. Sentiment Mining can be examined as a systematization process as shown in figure 1. Three systematization levels in Opinion Mining (*OP*) are: Document level, Sentence level and Aspect level. Document level OP goals to organize a thought document as conveying a negative or positive or neutral point of view. Here, the complete document is believed as a primary knowledge (discussing relating to a topic) [2].

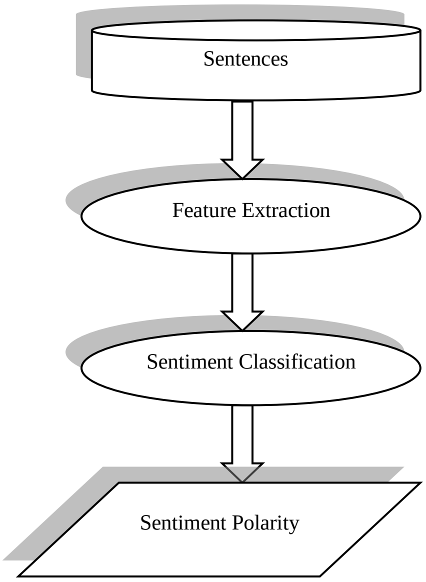

Fig. 1.Process of Sentiment Analysis

Sentence Level OP goals to organize sentiment conveyed in sentences. Firstly, it is recognized that either the sentence is objective or subjective.

If the recognized sentence is subjective then sentence level OP will decide that the sentence conveys negative or positive thought. Also, it is not essential that a sentence should be subjective. No elementary distinction exists between sentence level and document level classifications because sentences are small records or documents [3].

II. FEATURE EXTRACTION IN SENTIMENT CLASSIFICATION

The very first step is to extract features from the analyzed sentence. The types of features are:

Part of speech: It includes adjectives which are important for opinion or thought.

*Presence of Terms and their Frequencies:*These features are type of word i.e. individual word or N-gram words and their relative count of frequencies. Frequency count is used to show the relative value of features [4].

Words and Phrases for opinion: words and phrases that are commonly used to the opinions like *love or hate*, *high or low*.

Negation: as a negative word before any word may change the meaning of that word or opinion e.g. *not love* is similar to *hate.*

III. TECHNIQUES OF SENTIMENT CLASSIFICATION:

The techniques of Sentiment classification is divided into: lexicon based approach, machine learning approach and hybrid approach [5].

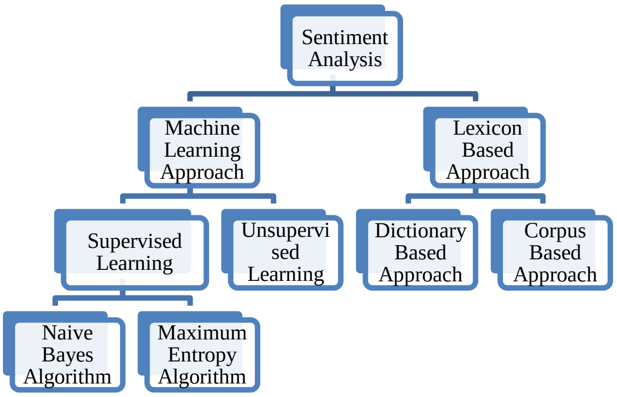

Fig. 2.Techniques of Sentiment Classification

In lexicon based approach, it is important to find lexicon which further use to examine the sentence. It is further divided into corpus based approach and dictionary based approach. Corpus based approach makes the use of methods to find the polarity of sentiments that is either positive or negative sentiments e.g. positive- wonderful, good, nice etc and negative- bad, very costly, arrogant, etc. Dictionary based approach begins with a set of words of opinions and it uses hierarchies to gather the opinions.

The machine learning approach uses the machine learning algorithms. By making the use of Machine Learning (*ML*) approach, text classification is of two types: supervised learning and unsupervised learning. When there is in bulk amount of training document then supervised method is used. And when it is difficult to attain this bulk amount of training document then unsupervised method is used. Whereas the last approach for sentiment classification is hybrid approach, is generated by combining both lexicon based approach and machine learning approach.

A. *Lexicon Based Approach*

Opinion lexicons are evaluated from text in sentiment classification technique. Opinion lexicons are of two types: positive and negative opinions. Positive opinion words are used to show the desired stage and negative opinion are used to show the undesired stage. Opinion lexicons are also known as Opinion phrases or idioms. Two approaches are:

1) *Dictionary Based Approach*

And the orientation of words is lexicon to us. Then this small collection of words is grown slowly by searching words in Corpora or thesaurus for their antonyms and synonyms [6].And this iteration stops, when no further words are left. But there is a disadvantage with dictionary-based approach that it is not capable to extract opinions with domain specific orientation.

2) *Corpus-Based Approach*

It comes into consideration to resolve the problem of dictionary-based approach. It has the capability of extracting opinion with domain specific orientations whereas corpus-based approach is not as efficient as dictionary-based approach because there is a need to make a large corpus for covering English words and which is a very difficult task. In spite of this, it is used commonly because of the big advantage is to provide the opinion words with domain specific orientations.

B. *Machine Learning Approach*

Machine Learning Approach totally based on algorithms of ML and used to solve the sentence classification problem and that further makes the rule of syntactic features. Two approaches are used here: supervised learning approach and unsupervised learning approach. In Supervised learning, when there is bulk of labeled training document then it is called supervised learning and is mainly of two types: Naïve Baye’s algorithm and Maximum Entropy Classifier.

1)*Naïve Baye’s Algorithm*

It is a supervised machine learning approach. It is totally based on Baye’s Theorem. Naive Baye’s provides good result in spite of having low Naïve Baye's Classification probability.

$$c* = argmac_c P_{NB}(c|d)$$

$$P_{NB}(c|d) := \frac{\left(P(c)\sum_{i=1}^{m} P(f|c)^{n_i(d)}\right)}{P(d)}$$

Where ‘f’ denotes feature, (d) denotes the no. of features found in text, m is maximum no. of features, P(c|d) denotes the conditional probability of class ‘c’ given the document ‘d’, P(c) is the probability of class ‘c’ and P(f|c) is the conditional probability of ‘f’ given class ‘c’. It has dependent assumption for its features. In order to classify the new sentence, we have to compute the product of the probability of every word of the sentence given a particular class, multiplied by the probability of the particular class [7].

Naïve Baye’s Classifier is well used in the real problems such as email Spam detection, Sentiment analysis, and Sexual content detection. It is highly recommended when we have lack of resources in terms of memory and CPU. It needs low processing memory and less execution time.

2) *Maximum Entropy Algorithm*

It provides a machine learning technique for prediction which is widely applied in the field of computer vision. And this model is known as multinomial logic model. A major advantage of maximum entropy model is that they are very much flexible and allows extra semantic, syntactic features [8].This model requires large amount of data to estimate the parameters of model. Parameter estimation of model is very expensive and also leads to round-off errors. Therefore, in order to get the estimation of parameters error free very accurate, efficient methods are needed.

$$P_{ME}(c|d,\lambda) = \frac{\exp[\Sigma_i \lambda_i f_i(c,d)]}{\Sigma_{c'} \exp[\Sigma_i \lambda_i f_i(c,d)]}$$

C is class, d is sentence, and λ is weight vector whereas weight vector calculates the importance of a feature. More weight denotes that a particular is strong enough in the class.

IV. CHALLENGES AND PROBLEMS FACED

A. Language Problem: In OP, English language is very well used because of its resource availability means lexicons, dictionaries and corpora but researchers get attracted by using OP with language other than English (Arabic, Chinese, German, etc). Therefore, researchers face a challenge for building resources i.e. lexicons, dictionaries and corpora for these languages.

B. Natural Language Processing (NLP): Using NLP in the OP process needs more enhancements because it attracts the researches. And NLP provide better OP results and provides good language understanding. There is a need to pay more attention in the research of domain-dependent opinion mining or context- based opinion mining because domain specific OP gives good result than domain independent corpus. And domain specific OP is difficult or more complicated to build.

*C. Fake Opinion:*It is also called Fake review and refers to bogus or fake reviews which misguide the readers or customers by providing them untruthful negative or positive opinions related to any object and in order to lower the reputation of any object. These spams make sentiment opinion useless in various application areas. This is a social challenge faced by the opinion mining and in spite of this challenge, OP made progress.

V. REFERENCES

[1] TsytsarauMikalai, Palpanas Themis. Survey on mining subjective data on the web. Data Mining Knowledge Discovery.

[2] Wilson T, Wiebe J, Hoffman P. Recognizing contextual polarity in phrase-level sentiment analysis. In: Proceedings of HLT/EMNLP; 2005.

[3] Liu B. Sentiment analysis and opinion mining. Synth Lecture Human Language Technology 2012.

[4] Yelena Mejova, Padmini Srinivasan. Exploring feature definitionand selection for sentiment classifiers. In: Proceedings of the fifthinternational AAAI conference on weblogs and social media;2011.

[5] Diana Maynard, Adam Funk. Automatic detection of political opinions in tweets. In: Proceedings of the 8th international conference on the semantic web, ESWC’11; 2011.

[6] Hu Minging, Liu Bing. Mining and summarizing customerreviews. In: Proceedings of ACM SIGKDD international conference on Knowledge Discovery and Data Mining (KDD’04);2004.

[7](2015) Machine Learning with Naïve Bayes Classifier [Online].Available:http://blog.datumbox.com/machine-learning-tutorial-the naive-bayes-text-classifier/

[8] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.