A Survey of Lexicon and Machine Learning based Sentiment Analysis

Vikas Kodag
Computer Engineering Department, PICT Pune
Dhankawadi, behind Bharti Vidyapeeth Pune. 411043.
vikaskodag2@gmail.com

***Abstract*—Sentimental Analysis is reference to the task of Natural Language Processing to determine whether a text contains subjective information and what information it expresses i.e., whether the attitude behind the text is positive, negative or neutral. It is also known as emotion extraction or opinion mining. This is a very popular field of research in text mining. The basic idea is to find the polarity of the text and classify it into positive, negative or neutral. It helps in human decision making. This paper presents various approaches used to classify the sentiment.**

Key words: sentiment analysis, Classifiers, Supervised learning, Unsupervised learning.

I. INTRODUCTION

Sentiment analysis (also referred as opinion mining) is the study of affective states and subjective information in the customer data (such as reviews and survey responses, online and social media) by using natural language processing and data mining techniques[1]. Sentiment analysis aims to determine the attitude of a subject with respect to some topic or the overall contextual polarity or emotional reaction to some object, such as a document, interaction, or event. The attitude may be a judgment or evaluation, affective state, or the intended emotional communication. Sentiment analysis is widely used in many applications that range from marketing to customer service to clinical medicine.

Sentiment level- sentiment analysis can be performed at various levels

- Document Level- In it the whole document is given a single polarity positive, negative or objective.
- Sentence Level – In it document is classified at sentence level. Each sentence is analyzed separately and classified as negative, positive or objective. Thus overall document has a number of sentences where each sentence has its own polarity.
- Phrase Level- It involves much deeper analysis of text and deals with identification of the phrases or aspects in a sentence and analyzing the phrases and classify them as positive, negative or objective. It is also called aspect based analysis.

II. CLASSIFICATION TECHNIQUES

Generally, there are two approaches in sentimental analysis. One is by considering symbolic methods and other one by machine learning method. In symbolic learning technique, which is categorized according to some learning strategies. In machine learning technique it uses unsupervised learning, weakly supervised learning and supervised learning. Along with lexicon based and linguistic method, machine learning will be considered as one of the mainly used approach in sentiment classification. The Fig.1 shows the sentiment classification techniques in detail.

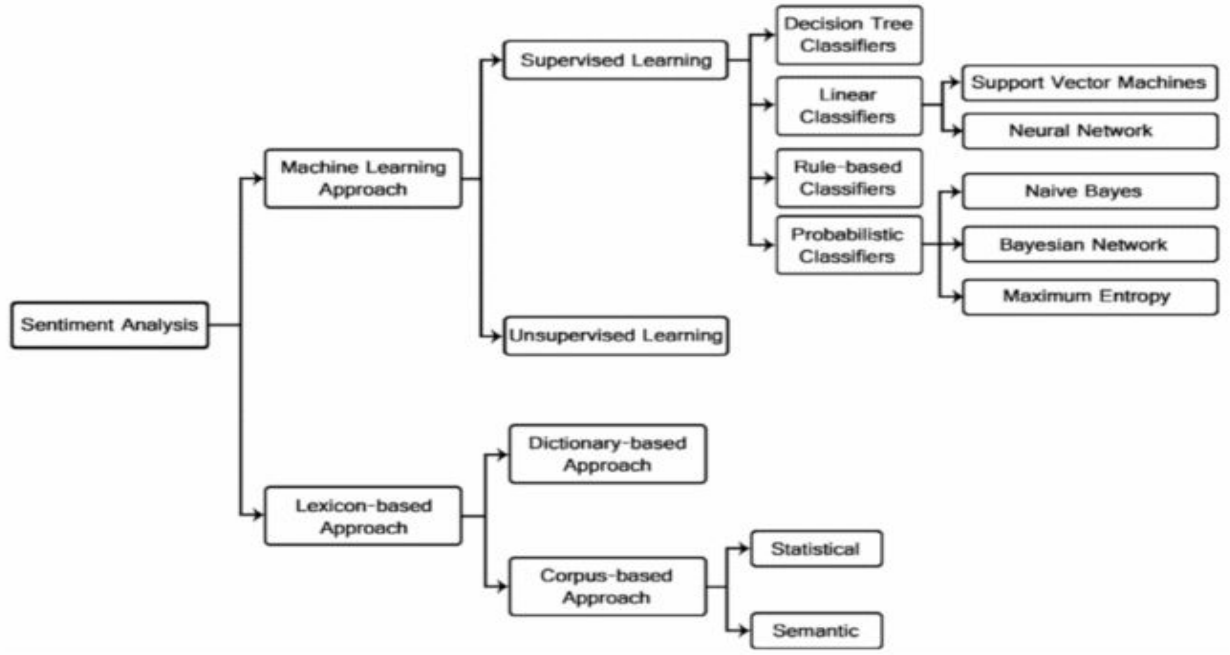

Figure 1: Sentiment Classification Techniques

2.1 Lexicon Based Approach

Subjective lexicons are collection of words where each word has a score indicating the positive, negative, neutral and objective nature of text. In this approach, for a given piece of text, aggregation of scores of subjective words is performed i.e. positive, negative, neutral and objective word scores are summed up separately. In the end there are four scores. Highest score gives the overall polarity of the text.

2.1.1 Dictionary Based Approach

In this method, manually annotated words are collected and a seed list is prepared. The synonyms and antonyms of text is searched in dictionaries and thesaurus. This newly found synonym is added to the seed list and the process continues until no new words are found.

2.1.2 Corpus Based Approach

Corpus is collection of writings, often on a specific topic. A seed list is prepared which is expanded with the help of corpus text, thus resolving the problem of limited domain oriented text.

2.2 Machine Learning Approach

This is an automatic classification technique. Classification is performed using text features. Features are extracted from text. This method includes two approaches. Supervised Learning where the system is trained using labelled training examples and Unsupervised Learning which do not require pre tagged data. Methods used in machine learning approach are:

2.2.1 Decision Tree Classifier

In Decision Tree classifier, the interior nodes were marked with features and edges that are leaving the node were named as trial on the data set weight. Leaves in the tree are named by categorization. This categories whole document by starting at the root of the tree and moving successfully down through its branches till a leaf node is reached. Learning in decision tree adopts a decision tree classifier as an anticipated model in which it maps information of an item to conclusions of that item's expected value.

2.2.2 Linear Classifier

In linear classifier, linear decision margins is used for classifying input vectors to classes. There are many types of linear classifiers. Support vector machine is one of them. This classifier provides a good linear scatters between various classes.

2.2.2.1 Neural Network

Neural network includes numerous neurons as its elemental unit. Multilayer neural network were used with non-linear margins. The results of the neurons in the previous layer will be given as input for the next layer. In this type of classifier training of data set is more complicated, because the faults must be back-propagated for various layers.

2.2.2.2 Support Vector Machine

Support Vector Machine (SVM) is known as the best classifier that provides the most accurate results in speech classification problems. They achieved by creating a hyperplane with maximal Euclidean distance for the nearest trained examples. Support Vector Machine hyperplane are completely resolved by a comparatively minute subset of the trained data sets which are treated as support vectors. The remaining training data sets have no access on the qualified classifier.

2.2.3 Rule Based Classifier

In this method, data set is designed along with a group of rules. In rules left hand side indicates the condition of aspect set and right hand indicates the class label.

2.2.4 Probabilistic Classifier

These classifiers use various forms for categorization. This variety of forms takes each and every class as part of that mixture. All various elements are the productive model in which it gives the probability of inspecting a distinct word for that element.

2.2.4.1 Naïve Bayes Classifier

It is frequently used for classification technique. It is named as naive since it considers every pair of features being classified is independent of each other. Calculation of whole document feasibility would be the substance in aggregation of all the feasibility report of single word in the file. These Naïve Bayesian classifiers were frequently applied in sentiment categorization since they are having lower computing power when comparing to the other approach.

2.2.4.2 Bayesian Network

This Bayesian network is directed non-cyclic graph where nodes correspond to variables and those edges are correspond to conditional independency. It is not usually used in text classification since it is expensive in computation.

2.4.3 Maximum Entropy

Maximum Entropy classifier is parameterized by a weight set that are used to associate with the joint-future, accomplished by a trained data set by encoding it. This Maximum Entropy classifier appear with the group of classifiers such as log-linear and exponential classifier, as its job is done by deriving some data sets against the input binding them directly and the result will be treated as its exponent.

III. APPLICATIONS

A) Support in decision making: Decision making is a very important field of our life. Opinions extracted from reviews helps us in making various decisions like "which books to buy", "which hotel to go", "which movie to watch" etc.
B) Business application: In today's world of competition, every company wants to satisfy its

customers requirements by creating new innovative products. Assessments of individuals are an essential angle today with the goal that organizations can get an input from clients and can roll out sought improvements in their item. Google Product Search is one illustration.

C) Predictions and trend analysis: sentiment analysis enables one to predict market trends by tracking views of public. It is also helpful in elections where candidates wants to know the expectations of people from them.

IV. RESULTS

In paper [2], they have taken online movie reviews for analyzing sentiments. For classification they used three supervised learning approaches such as Naïve Bayes, SVM and kNN. Experimental results show that SVM method beat the kNN and Naïve Bayes approaches.

The accuracy obtained by using three algorithms are shown in table 1. They have done 10 experiments for each approach. Result shows that even data is either small or large SVM provides higher accuracy than NB and kNN.

Experi-ment	Reviews	SVM (%)	Naïve Bayes (%)	kNN (%)
1	50	60.07	56.03	64.02
2	100	61.53	55.01	53.97
3	150	67.00	56.00	58.00
4	200	70.50	61.27	57.77
5	400	77.50	65.63	62.12
6	550	77.73	67.82	62.36
7	650	79.93	64.86	65.46
8	800	81.71	68.80	65.44
9	900	81.61	71.33	67.44
10	1000	81.45	75.55	68.70

Table 1: Accuracy obtained after testing data set

V. CONCLUSION

This paper presents a survey of sentiment analysis and classification algorithms. From the survey we can conclude that supervised learning methods like Naive Bayesian and Support Vector Machine are considered as standard learning method. Support Vector Machine provides excellent accuracy as compared to many other classifiers. Lexical based approaches are ideally aggressive because it requires manual work on document. Sentiment analysis of tweets is very popular. Datasets from sites like Amazon, IMDB, flipkart are widely used for sentiment analysis. Deeper analysis is required in case of social networking sites. In many cases, context consideration is very important. Therefore more research is required in this field.

VI. REFERENCES

1. D. Jurafsky, “Speech and language processing: An introduction to natural language processing,” Computational linguistics, and speech recognition, 2000.

2. P.Kalaivani, Dr. K.L.Shunmuganathan , “ Sentiment Classification Of Movie Reviews By Supervised Machine Learning Approaches”, ISSN : 0976-5166 Vol. 4 No.4 Aug-Sep 2013.

3. H. Kaur, V. Mangat and Nidhi, "A survey of sentiment analysis techniques," *2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, Palladam, 2017, pp. 921-925. doi: 10.1109/I-SMAC.2017.8058315

4. B. K. Bhavitha, A. P. Rodrigues and N. N. Chiplunkar, "Comparative study of machine learning techniques in sentimental analysis," 2017 International Conference on Inventive Communication and Computational Technologies (ICICCT), Coimbatore, 2017, pp. 216-221. doi: 10.1109/ICICCT.2017.797