Scientific Journal Impact Factor (SJIF): 1.711
e-ISSN: 2349-9745
p-ISSN: 2393-8161
International Journal of Modern Trends in Engineering and Research
www.ijmter.com

A Survey on Sentiment Categorization of Movie Reviews

Hardik Patel[1], Asst.Prof. Shafin Vahora[2]
[1]*Department of Computer Engineering, Ipcowala Institute of Engineering & Technology, Dharmaj, Anand, Gujarat, India- 388430*
[2]*Department of Computer Engineering, Ipcowala Institute of Engineering & Technology, Dharmaj, Anand, Gujarat, India- 388430*

Abstract – Sentiment categorization is a process of mining user generated text content and determine the sentiment of the users towards that particular thing. It is the approach of detecting the sentiment of the author in regard to some topics. It also known as sentiment detection, sentiment analysis and opinion mining. It is very useful for movie production companies that interested in knowing how users feel about their movies. For example word "excellent" indicates that the review gives positive emotion about particular movie. The same applies to movies, songs, cars, holiday destinations, Political parties, social network sites, web blogs, discussion forum and so on. Sentiment categorization can be carried out by using three approaches. First, Supervised machine learning based text classifier on Naïve Bayes, Maximum Entropy, SVM, kNN classifier, hidden marcov model. Second, Unsupervised Semantic Orientation scheme of extracting relevant N-grams of the text and then labelling. Third, SentiWordNet based publicly available library.

Keywords – Sentiment categorization, Naïve Bayes, Maximum Entropy, Support Vector Machine, k-nearest neighbour classifier, HMM, n-grams, SentiWordNet.

I. INTRODUCTION

Sentiment categorization refer as to extract contextual information by finding relation between words in unstructured text data review. It is a language processing task using computational approach to identify opinion and categorize it into positive or negative content. There are mainly three types of approaches for sentiment categorization of texts. First one is using machine learning based on text classifier such as Naïve Bayes, Maximum Entropy, Support vector machine, k-NN classifier, HMM etc. with suitable feature selection scheme.

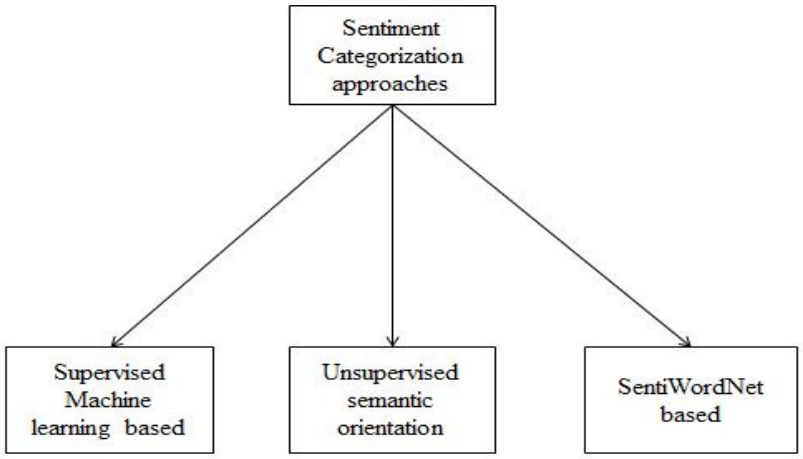

Fig 1. Sentiment categorization approaches

Second one is sentiment classification using unsupervised semantic orientation scheme of extracting relevant N-grams of the text and then labelling either as positive or negative or neutral and

consequentially the document from the user. And third one is classification using SentiWordNet based publicly available library that provides positive, negative and neutral scores for word [6].

Machine learning approach is the one which generally belongs to supervised classification technique, it is also known as sentiment analysis using supervised approach. In supervised approach there are two types of documents are used to carry out the sentiment analysis, first one is known as Training dataset and second one is known as testing dataset. Lexicon based approach is also known as the dictionary based approach or semantic based approach, in this approach there is no need for separate training and testing dataset but instead of that list of words or dictionary of words will be used to classify the text data in form of sentence or document. Much of the research based on lexicon approach make use of available lexical resources such as dictionary of positive and negative words which are going to be used to classify the sentence or document. As if there are some positive words in the sentence then it means that sentence represents positive polarity, and if there are negative words then it represents negative polarity of a sentence or document.

The machine learning approach applied to this problem mostly belongs to supervised classification in general and text classification techniques in particular for opinion mining. This type of technique tends to be more accurate because each of the classifiers is trained on a collection of representative data known as corpus. Thus, it is called "supervised learning". In contrast, using semantic orientation approach to opinion mining is "unsupervised learning" because it does not require prior training in order to mine the data. Instead, it measures how far a word is inclined towards positive and negative [2].
Sentiment categorization is perform at three level such as document level, sentence level and feature/aspects level.
(1) Document level – Classify the documents text data into positive or negative of words. It attempts to categorize the entire document into positive or negative. Document level determines the overall sentiment of a given review without considering the individual aspects. The entire process is combination of two steps: (a) Extracting the subjective features from the training data and converting them as feature vectors. (b) Training the classifier on the feature vectors and classifying its subject.
(2) Sentence level – Classify the sentences text data into positive or negative of words. It is just a short document, which targets the sentences and categories it as objective sentence or no opinion and subjective sentence or with opinion. The result is summarized to provide the overall result of the document. It is also known as Clause level analysis.

In this, the polarity of each sentence is calculated. The same document level classification methods can be applied to the sentence level classification problematic also but Objective and subjective sentences must be found out. The subjective sentences contain opinion words which help in determining the sentiment about the entity. After which the polarity classification is done into positive and negative classes.
(3) Aspects/feature level – It produces more focused and accurate sentiment summary. Multiple reviews on different aspects or domain specific evaluation.

II. MOVIE REVIEW OPINION MINING

Special challenges are associated with movie review mining. As it has been pointed out elsewhere, movie review mining is very domain specific and word semantics in a particular review could contradict with overall semantic direction (good or bad) of that review. For example, an "unpredictable" camera gives negative meaning to that camera model, whereas a movie with "unpredictable" plot sounds

positive to moviegoers. Therefore, we need to train the machine learning classifiers with movie review dataset as well as adapt the semantic orientation approach to movie review domain [2].

III. SUPERVISED MACHINE LEARNING APPROACH

The machine learning based text classifiers learn the set of rules (the decision criterion of classification) automatically from the training data. This clearly indicates that machine learning based text classification is a kind of supervised machine learning paradigm, where the classifier needs to be trained on some labeled training data before it can be applied to actual classification task. Usually the training data is an extracted portion of the original data hand labeled manually. Once the algorithm is trained to correctly classify the documents in the training set, it can be applied to the unseen data. If the learning method is statistical, the classifier is called a statistical text classifier. Naïve Bayes (NB) is one such classifier example. These text classifiers usually employ some kind of feature selection scheme, which decides what attributes of the text documents are evaluated while making a classification decision. Vector space model based classifiers are another popular category of text classifiers that represent the text documents as high dimensional vectors.

The process of machine learning as the following five steps:
(1) Extraction of key words as feature items
(2) Calculate the weights of each feature word
(3) Training samples
(4) Sentiment categorization
(5) Evaluation of performance

A. Naïve Bayes Algorithm

The Naive Bayes classifier is a probabilistic model based on the Bayes' theorem, which calculates the probability of a tweet belonging to a specific class such as neutral, positive or negative. This assumes that all the features are conditionally independent. Even though Naive Bayes classifier has yielded better results. It did not show superior results compared to some other classifiers. Naive Bayes classifier is a simple probabilistic classifier based on applying Bayes‟ Theorem with strong independence assumptions. The more expressive term for the underlying probability model would be independent feature model. This independence hypothesis of features make the features order is irrelevant and as a result that the presence of one feature does not affect other features in classification tasks which makes the computation of Bayesian classification approach more efficient. Naive Bayes classifiers can be trained powerfully by requiring a small amount of training data to estimate the parameters necessary for classification. It provides a simple method for text classification. It makes use of prior probability and feature distribution to determine the group that each text belongs to. Vector d(w1w2…wn) denotes text A and w_i is the feature item. If A belongs to group C_k then,

$$C_k = \arg\max P(c|d) \ , P(c|d) = \frac{P(c)P(d|c)}{P(d)}$$

B. Maximum Entropy

Its main task is to find the most suitable result which can meet all requirements in current situation. ME does not have to rely on the assumption of independent feature items, and it can treat all possibilities equally. We use exponential,

$$P_{\mathrm{ME}}(c \mid d) := \frac{1}{Z(d)} \exp\left(\sum_i \lambda_{i,c} F_{i,c}(d,c)\right)$$

where Z(d) is a normalization function. F_i,c is a feature/class function for feature fi and class c, defined as follows,

$$F_{i,c}(d,c') := \begin{cases} 1, & n_i(d) > 0 \text{ and } c' = c \\ 0 & \text{otherwise} \end{cases}$$

C. Support Vector Machine

SVM is a useful technique for data classification. In the former work, SVM has been shown to be highly effective at traditional text categorization. In the two-category case, the basic idea behind the training procedure is to find a hyper plane, that not only separates the document vectors in one class from those in the other, but for which the separation, or margin, is as large as possible.

SVM constructs a hyper plane or set of hyper planes in a high or infinite-dimensional space, which can be used for classification, regression, or other tasks. When the sets to discriminate are not linearly separable, the original finite dimensional space is mapped into a much higher dimensional space. Then we get the optimal classification plane which cannot only classify data into two groups but also ensure minimum errors. Classification function is shown in following:

$$dx = \sum_{i=1}^{m} a_i y_i K(x_i, x) + b$$

Where a_i and b can be obtain by SVM algorithm, K (x_i,x) is kernel function. When the value of a_i is not 0, samples become "support vector". For the line linear classification, we only need to know the inner product; for the non-linear classification, we need to convert it to a linear problem in high dimensional space.

D. k-Nearest Neighbour classifier

k-Nearest Neighbors algorithm for reducing the features extracted in text classification. The k-nearest neighbor algorithm (k-NN) is used to test the degree of similarity between documents and k training data. This method is an instant-based learning algorithm that categorized items based on closest feature space in the training set. The key element of this method is the availability of a similarity measure for identifying neighbors of a particular document.

This method is non parametric, effective and easy for implementation. One of the various classifier, 'KNN classifier' is a case based learning algorithm which is based on a distance or similarity function for various pairs of observation such as the Euclidean distance function. It is tried for many applications because of its effectiveness, non-parametric & easy to implementation properties. However, under this method, the classification time is very long & it is difficult to find optimal value of K. Generally, the best alternative of k to be chosen depends on the data. Also, the effect of noise on the classification is reduced by the larger values of k but make boundaries between classes less distinct. By using various heuristic techniques, a good 'k' can be selected. In order to overcome the above said drawback, modify traditional KNN with different K values for different classes rather than fixed value for all classes.

E. Hidden Markov Model

HMM is a probabilistic model for modelling time series data. It extends the concept of Markov Random Process to include the case where the observation is a probabilistic function of the states. Its hidden states are not directly visible and each state can emit observable output symbols determined by its own probability distribution. This extension makes HMM applicable to many fields of interest such as Natural Languages Processing (NLP), where the amount of observable events, i.e. Words, is often as big as hundreds of thousands.

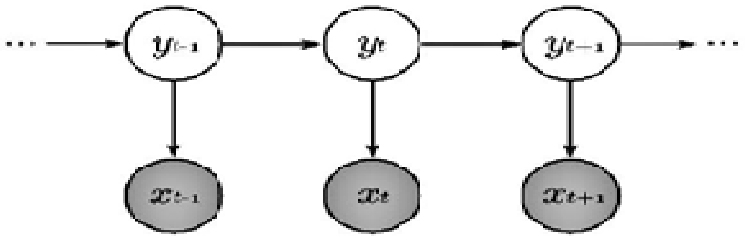

Fig 2.Graphical Representation of Hidden markov model[6]

It is a generative probabilistic model consisting of a states and an observations at each time stamp. There are two dependency assumptions that define this model. Represented with directed arrows in the fig 2. The current state at time t, namely yt, depends only on the previous state yt-1(Marcov assumption). The observation state at time t, namely xt, depends only on the current state yt at that time slice. Using these assumptions we can specify an HMM using three probability distributions: the distribution over initial states p (y1), the transition probability p (yt|yt−1) representing the probability of going from one state to the next, and the observation distribution p (xt|yt) indicating the probability that observation xt was generated by the state yt.We can factorize the joint distribution in terms of three distributions described above as follows:

$$P(x_i|y_i) = P(y_0)P(x_0|y_0)\prod_{i=1}^{n} P(y_i|y_{i-1})\ P(x_i|y_i)$$

There are several advantages of hidden markov model like, HMM are dynamically assembled according to the class sequences, Model will consider relative word of sequence in sentence from the dataset, In HMM every states are directly visible to observer [6].

IV. UNSUPERVISED SEMANTIC ORIENTATION APPROACH

A. N-gram Classifier

Semantic orientation from a word could be positive (i.e. praise) or negative (i.e. criticism). It indicates the direction that the word is in relative to the average. There are several dimensions we could consider regarding semantic orientation: direction and intensity. Direction indicates whether a word has positive or negative meaning. In opinion mining application, a word could indicate praise or criticism. Intensity designates how strong the word is. In opinion mining, a review could be found negatively milder than some other negative reviews. Another related work using semantic orientation included conjunctive words (i.e. and, but) to improve training a supervised learning algorithm, because we can understand the tone of the sentence from its conjunctions. “And” indicates that both adjectives have the same semantic orientation, whereas “but” indicates adjectives with opposite semantic orientations.

Semantic orientation consisted of three steps. First, a part-of-speech tagger extracted two-word phrases containing at least one adjective or one adverb from the review. The adjective or adverb carries semantic orientation, while the other word in the phrase provides context. Second, a technique called SO-PMI (Semantic Orientation using Pointwise Mutual Information) was used to calculate semantic orientation for the selected phrases. The extracted phrases will be judged in terms of how inclined they are towards positive or negative edges. The overall semantic orientation of each review is determined by averaging the SO-PMI values of all the phrases in it. Finally, the entire piece of review is identified as either positive or negative by comparing the overall semantic orientation and a baseline value [2].

When we use combination of more than two words for the feature vector that model will be referred as N-grams model. Which refers to combination of more words together to generate the feature vector and use that feature vector for classifying new or testing data. For the purpose of sentiment analysis is unigram model is considered to be best as far as the results are considered. All the experiments and evaluation provided in this report make use of Unigram as a feature selection model and which also provides some good results compared to other model like bigram.

V. SENTIWORDNET BASED

SentiWordNet based scheme for both document-level and aspect-level sentiment classification. The document-level classification involves use of different linguistic features (ranging from Adverb + Adjective combination to Adverb + Adjective + Verb combination). We have also devised a new domain specific heuristic for aspect-level sentiment classification of movie reviews. This scheme locates the opinionated text around the desired aspect/ feature in a review and computes its sentiment orientation. For a movie, this is done for all the reviews. The sentiment scores on a particular aspect from all the reviews are then aggregated.

The SentiWordNet approach involves obtaining sentiment score for each selected opinion containing term of the text by a lookup in its library. In this lexical resource each term *t* occurring in WordNet is associated to three numerical scores obj(t), pos(t) and neg(t), describing the objective, positive and negative polarities of the term, respectively. These three scores are computed by combining the results produced by eight ternary classifiers. To make use of SentiWordNet we need to first extract relevant opinionated terms and then lookup for their scores in the SentiWordNet. Use of SentiWordNet requires a lot of decisions to be taken regarding the linguistic features to be used, deciding how much weight is to be given to each linguistic feature, and the aggregation method for consolidating sentiment scores [5].

VI. CONCLUSION

Movie review mining is a challenging sentiment categorization problem. In this paper, various sentiment categorization approaches such as supervised machine learning, unsupervised semantic orientation, SentiWordNet approach uses for movie review opinion mining surveyed. In future, novel approach conditional random field(CRF) can be use for categorize movie review in terms of star cast, direction and music.

REFERENCES

[1] BoPang, Lillian Lee,Shivakumar Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques" In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10, pp. 79-86, 200

[2] Pimwadee Chaovalit, Lina Zhou,"Movie Review Mining: a Comparison between Supervised and Unsupervised Classification Approaches" IEEE Proceedings of the 38th Hawaii International Conference on System Sciences, 2005.

[3] P.Waila,V.K.Singh and M.K.Singh, "Evaluating Machine Learning and Unsupervised Semantic Orientation Approaches for Sentiment Analysis of Textual Reviews," Automation, Computing, Communication, Control and Compressed Sensing (iMac4s),pp.712-717,2012.

[4] B Agarwal, N Mittal, "Sentiment Classification using Rough Set based Hybrid Feature Selection", WASSA 2013: 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis.

[5] V.K. Singh, R. Piryani, A. Uddin, P. Waila, "Sentiment Analysis of Movie Reviews" IEEE 978-1-4673-5090-7,2013.

[6] Rohankumar Prajapati, Mukesh Goswami, "Sentiment classification on movie reviews using probabilistic graphical model" IEEE 2014.

