Machine Learning Technique for Sentiment Classification

Pramod Jangir[1], Navneet Kumar[2], Ompal Jangir[3]
Assistant Professor
Department of Computer Application
Shekhawati Institute of Technology, Sikar,
, ,

Abstract: Large amount of information are available online on web.The discussion forum, review sites, blogs are some of the opinion rich resources where review or posted articles is their sentiment, or overall opinion towards the subject matter. The opinions obtained from those can be classified in to positive or negative which can be used by customer to make product choice and by businessmen for finding customer satisfaction .This paper studies online movie reviews using sentiment analysis approaches. In this study, sentiment classification techniques were applied to movie reviews. Specifically, we compared two supervised machine learning approaches SVM, Naive Bayes for Sentiment Classification of Reviews. Results states that Naïve Bayes approach outperformed the SVM. If the training dataset had a large number of reviews, Naive bayes approach reached high accuraciesas compare to other.

Keywords: *Sentimental Analysis, supervised Algorithm, Naive Bayes, Support vector machine.*

*****

1. Introduction

Opinions are important to almost all human activities because they are key influencers of our behaviors. At whatever point we have to settle on a choice, we need to know other‟s opinion. In this present reality, organizations and associations dependably need to discover buyer or general feelings about their Items and administrations. Singular buyers additionally need to know the sentiments of existing clients of an item before buying it, and others' feelings about political competitors before settling on a voting choice in a political decision. Before, when an individual required assessments, he asked loved ones. At the point when an association or a business required shopper opinion, it led studies, assessment surveys, and center gatherings. Securing open and buyer suppositions have for some time been an immense business itself for promoting, advertising, and political crusade organizations. With the hazardous development of online networking (e.g., audits, gathering dialogs, sites, smaller scale websites, Twitter, remarks, and postings in interpersonal organization destinations) on the Web, people and associations are progressively utilizing the substance as a part of these media for choice making. These days, if one needs to purchase a customer item, one is no more restricted to approaching one's loved ones for conclusions in light of the fact that there are numerous client audits and examinations in broad daylight gatherings on the Web about the item. For an association, it might never again be important to direct studies, conclusion surveys, and center gatherings with a specific end goal to accumulate popular assessments on the grounds that there is a wealth of such data openly accessible

Our goal is to calculate the polarity of sentences that we extract from the text of reviews. We will find the sentiment of this review and find whether the movie is successful or not. So that we can find whether movie is positive or negative. We examine the effectiveness of applying machine learning techniques to the sentiment classification problem. Our analysis helps concerned organizations to find opinions of people about movies from their reviews, if it is positive or negative. One can in turn formulate a public opinion about a movie.

The challenging aspect in sentiment analysis is an opinion word which is considered as a positive in one situation may be considered as negative in another situation. The traditional text processing considers that a little change in two bits of content has no change in the significance or meaning [1]. But in sentiment analysis a little change in two bits of content has change in the significance or meaning, consider Example "story is good" is different from "the story is not good". The system process it by analyzing one by one sentence at a time [3]. However, blogs and twitter contains more informal sentences which user can understand and but system cannot understand it. Consider example, "that movie story was as good as its previous movie" is dependent on previous movie whose details is not available.

Another challenging aspect of this problem that seems to distinguish it from traditional topic-based classification is that while topics are often identified by keywords alone, sentiment can be expressed in a more subtle manner [2]. For example, the sentence "How could anyone watch this Drama?" contains no single word that is obviously negative. Thus topic-based classification can easily understandable then sentiment. So, apart from presenting our results obtained via machine learning techniques, we also understand the problem to gain a better understanding of how difficult it is. Consider another example visual effect of movie were good but storyline was terrible this convey both positive and negative meaning respectively.

Thus review can be helpful by providing usefull information to customer as well as businessmen. For customer it provide useful information that which product is good by examining the rating that come with it. Opinions or sentiment, can also provide researchers, businessmen, and policy-maker with valuable information ranging from rates of customer satisfaction to public opinion trends.

2. Related Work

The concept of sentiment analysis and opinion mining were first introduced in the year 2003. Several techniques were used for opinion mining in history. The following few works are related to this technique. Pang‟s work in paper[1] from 2002 on using supervised machine learning techniques to perform sentiment classification. They used the machine learning methods such as Naive Bayes, maximum entropy classification, and support vector machines .This methods commonly used for topic classification.

The objective of this [2] paper is to determine the positivity or negativity of the movie reviews at document level .The system generate the results generated which are summarized and helpful. Experimental result indicate that the „Document based Sentiment Orientation System‟ perform well as compared to „AIRC Sentiment Analyzer with respect to movie domain.

In this paper[3], compares three supervised machine learning algorithms of SVM, Naive Bayes and KNN for sentiment classification of the movie reviews that contains 1000 positive review and 1000 negative reviews. The results show that the SVM approach outperformed than the Naive Bayes and k-NN approaches and the training dataset had a large number of reviews, the SVM approach reached accuracies of more than 80%.

The proposed paper [4] work presents an approach for sentiment analysis by comparing the different classification methods in combination with various feature selection schemes. It successfully analyzed the different feature selection schemes and their effect on sentiment analysis. The classification clearly shows that Linear SVM gives more accurate result than Naive Bayes classifier. Although many other previous works have also shown SVM as a better method for sentiment analysis but work differs from previous works in terms of the comparative study of the classification approaches with different feature selection schemes.

This paper [5] shows that using emoticons as noisy labels for training data is an effective way to perform different supervised learning .Machine learning algorithms can achieve high accuracy for classifying sentiment by using this method. Although Twitter messages have unique properties compared to other machine learning algorithms classify tweet sentiment with same performance.

This paper [6] introduce new approach called combined approach to classify text reviews based on sentiment present in that reviews. With the help of two classifier and classifier combination rules it is possible to improve expected classification results. It also propose way of handling slang words and smiley for overall causes of good sentiment classification with higher accuracy.

3. Methodology

3.1 Data Collection

This paper uses the Internet Movies Database (IMDB) movie review dataset. This data consists t of unprocessed, unlabeled file from the IMDB archive at http://reviews.imdb.com/Reviews. In The dataset we have 1400 processed text files. These files are divided in two types with respect to their classification as "pos" and "neg", indicating the true classification (sentiment) of the component files

3.2 Text Preprocessing

This stage includes getting the actual text for all the data we have and trying to separate the individual reviews by considering each review is a single line of the file. As a result, this method will turn into just splitting the content of the file by the end of the line character.

Other part of this stage is to convert the resulted reviews into lower case, so in that case we can get matches with the AFINN data that we used. Also to avoid mismatch cases we omitted punctuations, numbers and control characters to get better matches.

3.3 Classification algorithm

There are different levels of Sentiment analysis. The document level, sentence level or the attribute level. Here we use document level sentimental analysis.In this study, we applied two supervised machine learning models for sentiment classification for the selected movie reviews. These models are Naive Bayes (NB), and support vector machines

To implement these machine learning algorithms on our document data, we used the following standard features. Let f1,f2,.. ,fm be a predefined set of m features that can appear in a document . Let ni(d) denote number of times features fi occurs in document d. Then, each document d can be
represented by the document vector d
:= (n1(d), n2(d),…. , nm(d)).

3.3.1 Naive bayes

This is a simple probabilistic classifier that is based on the Bayesian probability. The Naive Bayes classifier is based the assumption that feature probabilities are independent of one another. This classification technique assumes that the any feature in the document is independent of other feature. Naive Bayes classifier considers a document as collection of words and assumes that the probability of a word in the document is independent of its position in the document and the presence of other word .We derive the Naive Bayes (NB) classifier by Bayes' rule,

$$\mathrm{p(c/d)}=\frac{p(c)p(d/c)}{p(d)}$$

Where P(d) plays no role in selecting c. But its conditional independence assumption clearly does not exist in real-world situations, Naive Bayes-based text classification still tends to perform well.

3.3.2Support vector machines

Svm have been the efficient way for document classification. These are large margin classifiers. The basic idea behind SVM classification is to find hyper-plane with maximum margin that separates the document vector in one class from the other with maximum margin. They are large-margin, rather than probabilistic, classifiers, in contrast to Naïve Bayes. This search corresponds to a constrained optimization problem; let the classcj $\{1,-1\}$(consider as positive and negative) be the correct class of document denoted by dj , the solution can be given by vector W

$$w{:=}\propto jcjdj, \propto j \geq 0$$
$$j$$

Where the αj 's can be obtained by solving a problem of dual optimization. Those document dj such that αj is greater than zero are called support vectors, because αj are the only document vectors contributing to vector w. Classification of instances consists of finding which side of w's hyper plane they fall on

4. Experiments

4.1Naive Bayes

Naive Bayes classifier work on the principle of probabilities and the Bayes rule given by: p (c/d) = (p(c)p(d/c))/p(d). Where P (c|d) is the probability of a given document (text) belongs to class c, which is the classification part which we are interested in. Below is the confusion matrix for the naive bayes classifier in our project. The classifier has obtain accuracy of 65.57%.

Table 1: Confusion Matrix of Naive bayes

	Actual		
predicted		Positive	Negative
	positive	434	216
	Negative	266	484

4.2 Support Vector Machine

Below is the confusion matrix of the performance of the support vector machine. We can see that this classifier has misclassified more number of data points as compared to naïve bayes. The accuracy of this model comes out to be 45.71% which is lower than that for naïve bayes.

Table 2: Confusion Matrix of support vector machine

	Actual		
predicted		Positive	Negative
	positive	267	257
	Negative	36	373

5. Conclusion

In this paper we propose approach to classify text reviews based on sentiment present in that reviews. We learned that the traditional machine learning classification algorithms do not work very well with sentiment analysis of text as compared to their performance with topic based classification. We also learned that out of the two algorithms we used for the baseline Naive bayes performed the best by giving high accuracy. Following are result obtain after applying supervise classification algorithms.

Table 3: Result

Method	Accuracy
Naive Bayes	65.57 %
Support Vector Machine	45.71 %

6. Future Work

We will make feature selection using unigrams, bigrams and trigrams of the data and using these n-grams as features to train a model we built a logistic regression classifier to test if n-grams help for a better classification.

References

[1] Bo Pang and Lillian Lee and ShivakumarVaithyanathan "Thumbs up? Sentiment Classification using Machine Learning Techniques", Language Processing (EMNLP), Philadelphia, July 2002, pp. 79-86.

[2] Richa Sharma, Shweta Nigam and Rekha Jain "Opinion mining of movie review at document leve"l, International Journal on Information Theory (IJIT), Vol.3, No.3, July 2014.

[3] P.Kalaivani, Dr.K.L.Shunmuganathan, "Sentiment classification of movie review by supervise machine learning approach", Indian Journal of Computer Science and Engineering (IJCSE) Vol. 4 No.4 Aug-Sep 2013

[4] GautamiTripathi and Naganna S, "Feature Selection and classification approcha for Sentiment Analysis", Machine Learning and Applications: An International Journal (MLAIJ) Vol.2, No.2, June 2015