[[1]](#footnote-1)

Classifying the polarity of reviews using Sentiment Analysis based on various Classification Techniques

Pratik Walawalkar

***Abstract*—A review describes and analyzes different aspects about a product and the experience of user with respect to those aspects. It is nothing more than an assessment of the product. For movies, reviews is based on the quality of movie, generally assessing the actors performance, the story of the movie, direction, screenplay of movie etc. The review generally consists of short text ranging from few words to few lines. The dataset I am using currently consists of 294 movie reviews. The dataset is consisting of positive movie reviews as well as negative movie reviews. In this project I propose to build a classifier using various classification techniques to analyze the movie reviews and classify the polarity of those movie reviews. Also my aim is to determine which classification technique is more useful for sentiment Analysis for review of movies. This will help in analyzing what general opinion; the users have about a particular movie.**

# INTRODUCTION

With huge advent of databases and its usage to store large quantity of data, a new concept was required to convert the vast data into meaningful information. This lead to the research of data mining. But there is a catch. It is not necessary that the information be present only in structured format. It may also be available in text format. So different techniques and methods are required to properly extract the useful information from these texts. So that the information obtained can be analyzed properly for future usage. The text format type of information may be present in journal, documents, web, manuals, e-books etc. [1].

To take advantage of this information present in documents, journals, research papers, manual, web etc. we use Text Mining. Text Mining involves extracting and analyzing high quality information from data present in texts. [2]. Text Mining is a vast field and has many sub-fields like Text Clustering, Sentiment Analysis, Information retrieval, Pattern Recognition, Lexical Analysis, Information etc. [3][4][5].

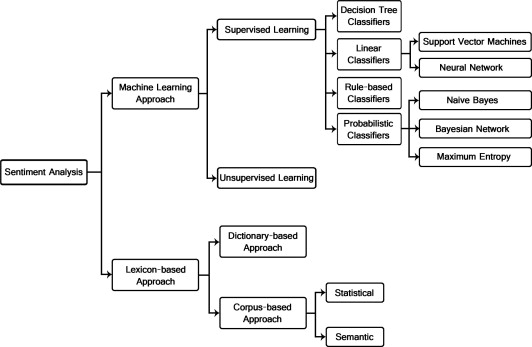
Sentiment Analysis is one of the sub fields of Text Mining. It is used to find out the views or feeling of the user from the text he has written. In commercial world, one of the most important part of information gathering is to gauge the interest

and opinion of people. Now the availability of social media, blogs and review sites allows the users to share their opinion.

This has resulted in more interest in sentiment analysis as competition to evaluate the texts of users and determine its

polarity increases. Now one of the main areas where the people’s opinion matters the most is in movies. Here the attitude of the viewers towards the movie determines its outcome at the box-office. With few exceptions, generally, movies with bad reviews tend to perform poorly at box-office. These movies then determine the fate of not only which

genre of films but also, which actors, directors would be in most demand and which would be less so. So determining the people’s attitude towards the movie is very important. In this project the main aim is to classify the polarity of the reviews which customers have posted on the amazon/IMDb/rotten tomato website and also analyze the various sentiment analysis techniques that can be used here. For this a dataset will be used which consists of movie reviews. On this dataset, Sentiments Analysis technique will be performed to determine the polarity of reviews and then we will analyze the various sentiment analysis techniques used to determine its efficiency. The image shown below describes the various sentiment analysis techniques which are available.



Sentiment Analysis Techniques [6][7]

# Description

## Problem Description

Movie reviews play a major role in Film Industry. They have a major impact at determining the box office success of the movies. Good reviews, naturally tend to attracts more movie goers to the theatres any a poorly reviewed movie. Earlier the movie critics would review the movies and played a major role in success of the movie. Their reviews in newspapers, film magazines would play a substantial role in determining the success of its box office collections. But now a day with the advent of social media even general audience are able to review and critic the movie. Sites like IMDb are solely designed with the aim to provide the general movie goers to review the movie. What’s more is that, the social media has a wider scope of fan base and audience and is more popular than the traditional newspapers. Also, the average movie goer is able to connect more with the reviews posted by a layman like him than a professional movie critic. The general audience’s reviews are more closely reflect the preferences of

movie goers in general. A good review by average movie goers can attract more audience to the movie while a poor review can play an important role in keeping the audience away from the theatres. Observing these reviews helps determining the kind of products the audience expects form movies. Various researches have been carried out to determine whether is there really a co-relation between box office performance of a movie and reviews it got. Almost all the research carries out like Basuroy, Chatterjee, & Ravid, 2003; Desai, & Basuroy, 2005; Gemser & Gerda, 2007 have given a positive co relation between reviews and performance of a movie [8]. Also in 2006, Gino and Ferrain carried out a research which also proved the impact social media reviews play in determining the success of the movie. [9]

So Movie reviews are really important for the film makers, and actors to gauge and understand the taste and preference of audience. It helps them to understand where they came up short and improve on their previous work. Also movie reviews are important to large movie studios as it helps them understand better the new trends among movie audience.

So it is necessary to analyze the movie reviews and analyze the important information from it. But reading and analyzing millions and millions of movie reviews is not an ideal way as it would take a lot of time and achieve nothing. So one of the best way to analyze the movie reviews is to analyze its polarity. This will help in get a general understanding of audience’s attitude towards a particular film. Once the general attitude towards the film is known, it will help film makers in predicting the box office success as well as information required for future movie endeavors.

## Solution

The polarity of the movies can be identified by performing sentiment analysis on it. Sentiment analysis is one of the sub field of Text Mining. Sentiment analysis identifies the polarity and then categorizes it into positive, negative. Thus it is easy to identify the audience’s general opinion of the film. The earliest work was done by Pang Bo, Lillian Lee[9] in using sentiment analysis to classify the polarity of movie reviews in their work “Thumbs up? Sentiment Classification using machine learning techniques.”.[10]

Now there are various techniques for sentimental classification. The Sentimental Analysis classification can be classified into two types Machine Learning Approach and Lexicon based approach. The Machine learning approach is further categorized into supervised learning and unsupervised learning approach. The Lexicon based approach is classified into Dictionary based approach and corpus based approach. Now in this project I have implemented Machine Learning based approach to implement sentimental analysis to classify the polarity of reviews. The two categories of machine learning based approach, supervised learning and unsupervised learning both have its advantages and disadvantages. In supervised learning, the output categories are already known. In supervised learning, before the raw data or test data that is to be classified is provided, a feature set called training data set is also provided. The data in this dataset consists of labelled data. So the algorithm firsts learns/supervises how the data is labelled to the output categories. Once the algorithm learns how to map the data, it then performs classification on the test dataset /raw dataset and correctly maps to the output categories. In unsupervised learning, no such training data is provided. The algorithm has to adapt itself gradually and find correlations to classify the data. The algorithm then gradually finds the correlation and maps the data to output categories it has discovered. Generally this types of algorithms are better suited for clustering, as it beneficial to find the correlation between data by itself rather than providing it our self.

## Algorithms Used

In this project, I have implemented supervised learning algorithms to implement the classifier because, the output categories are already known. We have to find the polarity of review, so the output category will be positive and negative. So the test data/raw data is to be mapped to these two categories. So when the sentiment analysis is finally implemented, the reviews will be identified either as positive or negative. For this approach, a training data consisting of known positive and negative polarity data is provided. I have provided 80 positive and 80 negative reviews as a training dataset. The supervised algorithms will first analyze this data set and learn how to label the reviews as positive or negative. Then test dataset of 294 reviews will also be provided to implement sentimental analysis and classify the polarity of of those reviews. Once sentimental analysis is performed, I will analyze all the algorithms implemented to find which algorithm provides better result. In order to analyze the results, the test dataset that is provided by me , already has the data labelled, to identify whether the polarity is identified correctly. Bout when the dataset is passed to the algorithm, its polarity label data will not be provided to the algorithm, so it can identify the polarity itself based on the training data set.

The machine learning algorithm implemented by me in this project are naïve Bayes, SVM (support vector machine), Random Forest, Decision Tree, MAXENT (maximum entropy) and GLMNET (generalized linear model).NNET(Neural Network), SLDA(Supervised Latent Dirichlet Allocation), Bagging and Boosting.

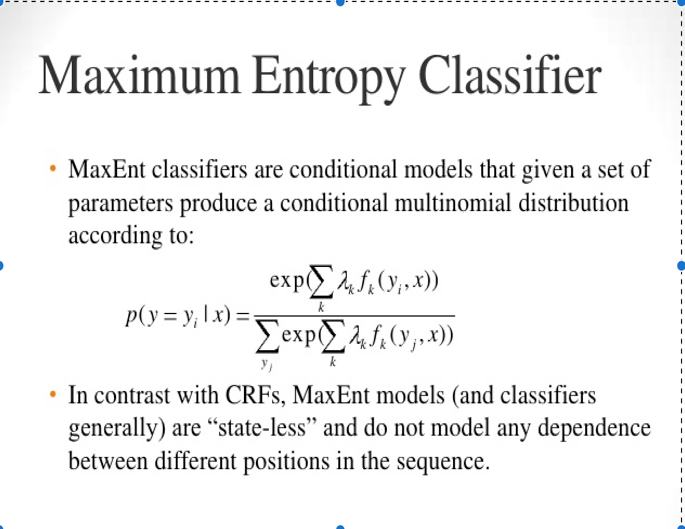
Of these naïve Bayes is the most common and widely used algorithm in supervised learning. I will compare the results obtained by other method with this and analyze which method provides better results.

Of all the algorithms, I am implementing, naïve Bayes and Maximum Entropy are probabilistic classifiers. Probabilistic classifiers are those classifiers which predict the probability distribution over the output categories based on the input data. Naïve Bayes is a probabilistic classifier in the sense that it uses Bayes theorem on likelihood possibilities to classify the input to output labels.



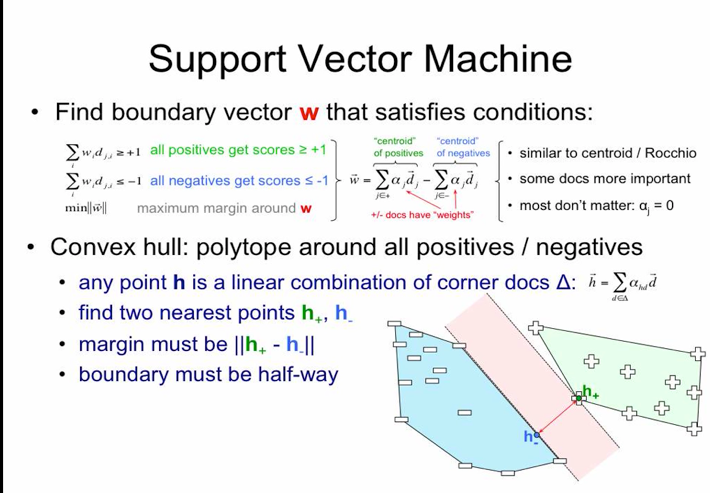
naïve Bayes probability [10]

Maximum entropy model is a classifier which recognizes uniform distribution over the input and then calculates their likelihood probability and label them to output categories.

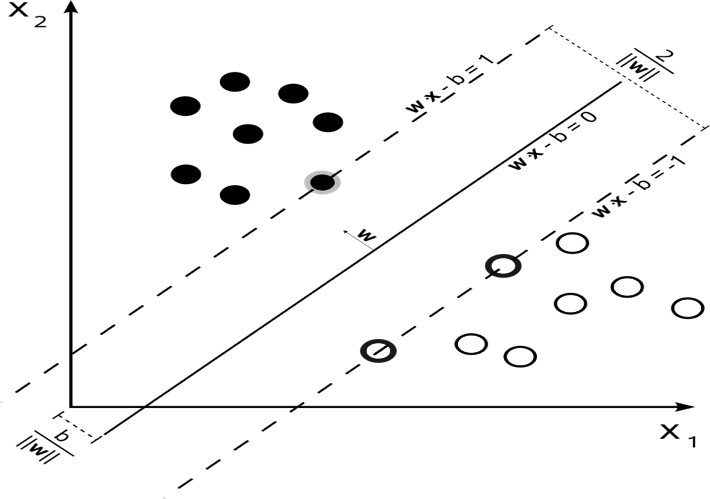


MAXENT probability [11]

SVM (support vector machine) is a linear classifier which categorizes the data from the output data set into two output categories. It designs a hyper plane which categorizes the input data into output labels.

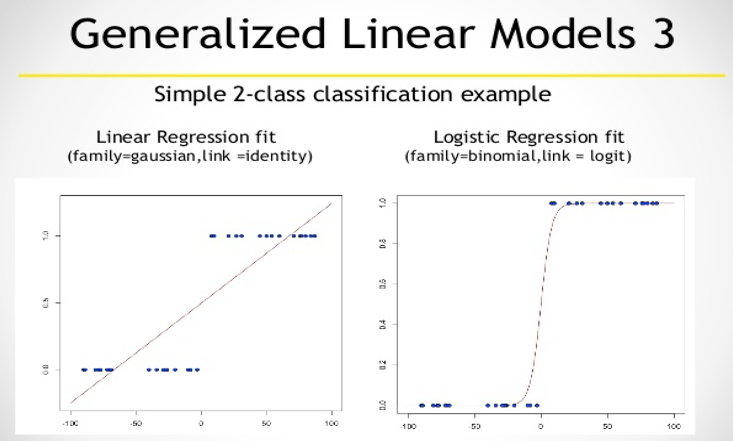


SVM algorithm [12]



SVM classification [12]

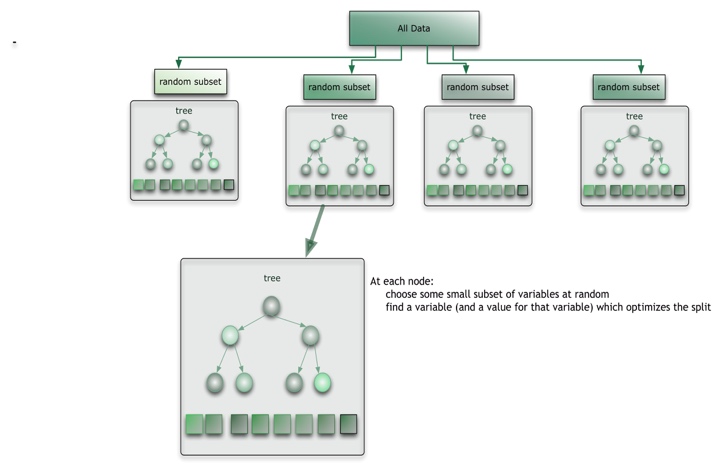
GLMNET is a type of linear regression model which predicts the expected value from the input data set by making use of the predictor.



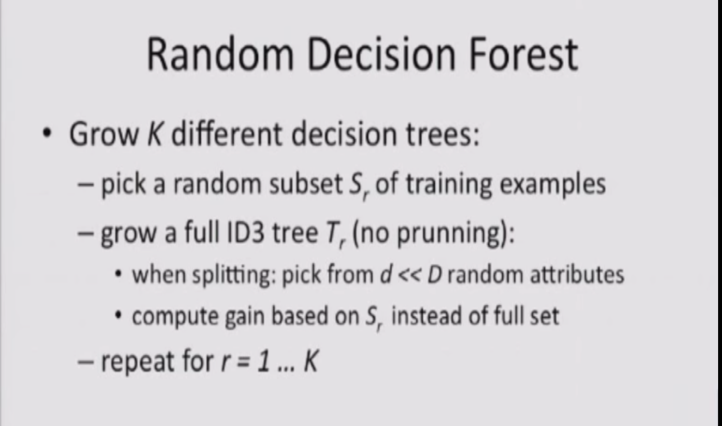
GLMNET [13]

Random Forest and Decision trees are Decision tree classifiers. Decision trees are famous for ignoring noise in the data set which helps them achieve good result in data sets involving interference and noise. Both these use decision tree to observes the input and then properly maps it with output categories to categorize the out put. But the main difference between decision trees and random forest is that, while in Decision tree classifier only 1 decision trees is generated, in random forests, k different trees are generated. It is achieved by selecting random subsets from training data and then implementing decision trees on it. They are fast and compact which gives them an advantage over decision trees which implements a single tree on large training data set. Both Random forest and decision tree use ID3 algorithm. They implement top down approach to build the tree. They start from the root node and continues to partition until eventually

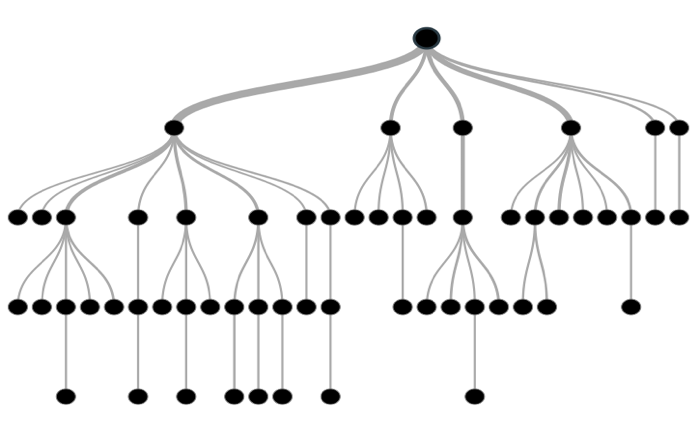
the subsets of input data is properly categorized into output labels.



Random Forest [14]

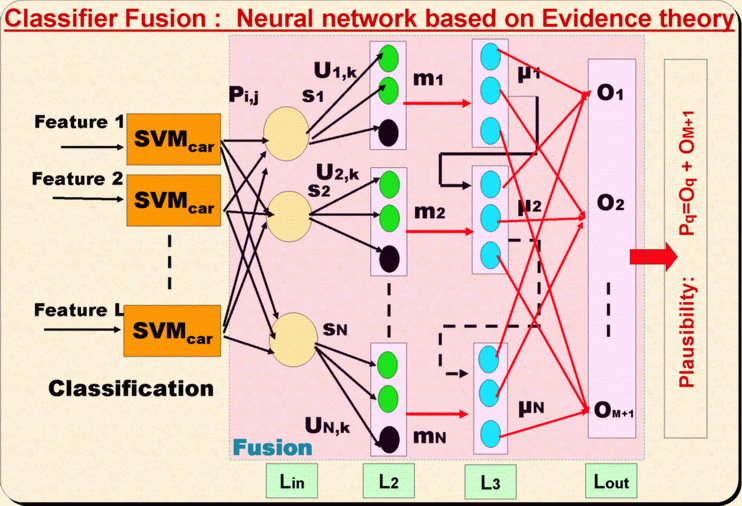


Random Forest Algorithm [14]



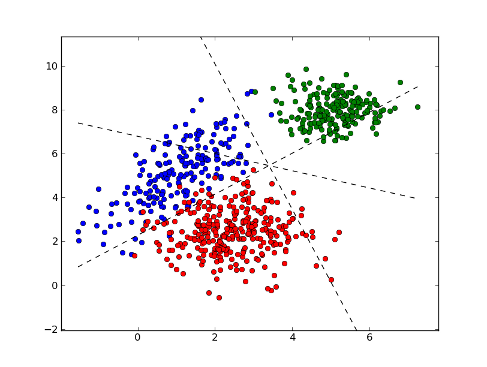
Decision Tree [15]

The next one to be implemented is NNET. It is also a linear classifier. It forms an artificial network. It categorizes the input data into classes and then label the data to represent the output label.



NNET algorithm [16]

SLDA is a type of Supervised Latent Dirichlet Allocation classifier. But the main difference being while LDA is unsupervised, SLDA is supervised learning. This algorithm was specifically developed to determine the movie ratings and movie reviews polarity. It was proposed by David M. Blei and Jon Mcauliffe in their paper Supervised topic models. They proposed it as a statistical model which will label documents appropriately.



SLDA classifying the data [17]

## Approach

I have based my approach on the concept proposed by A. Amolik in his paper “Twitter Sentimental Analysis of movie review using Machine Learning” [24]. In my project, I have used the same concept where I am providing tweets obtained from twitter database as training database and providing movie reviews as testing dataset. I am going to analyze the efficiency of classifiers implemented using this concept. For this , I needed to obtain training dataset of tweets with labelled polarity.

The data required for the training is obtained from the training data set in “Twitter Sentimental Analysis” project by “victorneo” from github. While the testing dataset is taken from “Sentimental Analysis” project by “abromberg” from github.

The project is implemented in R language and the code is based on the code published by Cheng-Jun Wang in datascienceplus.com website. In this project, first all the classifiers are built and trained are by providing them with training data, and then they are implemented to analyze the polarity of the review.

Following is the approach of how I completed the project.

|  |  |
| --- | --- |
| Task | Approach |
| Understanding and gathering information on various Pattern Recognition Topics | Read various content of Pattern recognition on web and once again go through the lectures |
| Deciding on Topic | Determine which topics look the most interesting |
| Literature Study to gather information on Text Mining | Read various papers on Text Mining and watch tutorials on text mining |
| Literature Study to gather information on Sentiment Analysis | Read various papers on Text Mining and watch tutorials on Sentiment Analysis |
| Preparing the Project Proposal | Write Project Proposal in IEEE format |
| Gather More information on Sentiment Analysis | Read various papers on Text Mining and watch tutorials on Sentiment Analysis |
| Literature Study to gather information on various classification techniques | Read various papers on Text Mining and watch tutorials on Classification Techniques |
| Learn R | Watch tutorials and practice from rtutor and code school |
| Programming Sentiment Analysis | Code in R and implement the classifiers based on Machine learning supervised algorithms. |
| Classify the polarity of reviews | Analyze the outputs |
| Analyze the results obtained from various classification techniques | Analyze the output obtained from coding in r and comparing them with each other |

# Evaluation

As noted above, the supervised machine learning algorithm were implemented and then tested on the test dataset of movie reviews. The output label generated were of following four categories.

True Negative(TN): This is the outcome when the output category of a negative movie review is correctly labelled as having negative polarity.

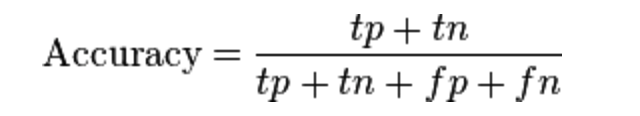
True Positive(TP): This is the outcome when the output category of a positive movie review is correctly labelled as having positive polarity.

False Negative(FN): This is the outcome when the output category of a positive movie review is incorrectly labelled as having negative polarity.

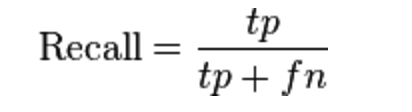
False Positive(FP): This is the outcome when the output category of a negative movie review is incorrectly labelled as having positive polarity.

Also the following were calculated to analyze the efficiency of supervised algorithms.

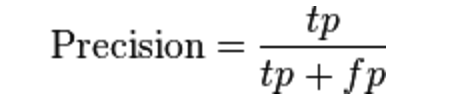
Accuracy: Accuracy is total number of correct classification divided by total number of cases. It gives the percentage about how much data is correctly labeled from the available data. It is calculated as the ratio of true positive and true negative upon the sum of true positive and true negative and false positive and false negative.



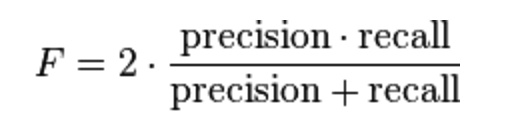
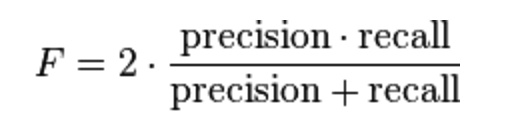
Recall: Recall is the number of correct identified positive labels divided by actual positive labels. It determines how much correct percent of positive polarity was calculated. It helps to identify as many relevant reviews as possible which are of positive polarity. It is nothing more than the percentage of labelled input data that is correctly identified. It is also known as sensitivity. It is calculated as the ratio of true positive upon the sum of true positive and false negative.



Precision: Precision is the number correct identified positive labels divided by total positive labels identified. It helps to identify how much percentage of particular polarity identified was correct to that particular polarity. Here the polarity identified is positive. It is nothing more than the percentage of correctly labelled input data. It is also known as sensitivity. It is calculated as the ratio of true positive upon the sum of true positive and false positive.



F-score: F-Score or F-measure is the harmonic mean of recall and precision.



After implementing all the classifiers the following results were obtained:

Naïve Bayes:

The results obtained from implementing naïve Bayes is

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | 90 | 62 |
| Negative | 91 | 51 |

Accuracy: 0.4796

Recall: 0.4972

Precision: 0.5921

F – Measure :0.54

All Supervised Learning Methods:

The results obtained for implementing all the rest Supervised learning methods are:

SVM:

The observations are as follows:

|  |  |
| --- | --- |
| Precision | 0.550 |
| Recall | 0.510 |
| F-Score | 0.385 |
| Accuracy | 0.53 |

SLDA:

The observations are as follows:

|  |  |
| --- | --- |
| Precision | 0.520 |
| Recall | 0.505 |
| F-Score | 0.405 |
| Accuracy | 0.520 |

BOOSTING:

The observations are as follows:

|  |  |
| --- | --- |
| Precision | 0.760 |
| Recall | 0.505 |
| F-Score | 0.350 |
| Accuracy | 0.523 |

BAGGING:

The observations are as follows:

|  |  |
| --- | --- |
| Precision | 0.760 |
| Recall | 0.505 |
| F-Score | 0.350 |
| Accuracy | 0.523 |

Random Forest:

The observations are as follows:

|  |  |
| --- | --- |
| Precision | 0.495 |
| Recall | 0.495 |
| F-Score | 0.495 |
| Accuracy | 0.513 |

GLMNET:

The observations are as follows:

|  |  |
| --- | --- |
| Precision | 0.760 |
| Recall | 0.505 |
| F-Score | 0.350 |
| Accuracy | 0.523 |

Decision Tree:

The observations are as follows:

|  |  |
| --- | --- |
| Precision | 0.760 |
| Recall | 0.505 |
| F-Score | 0.350 |
| Accuracy | 0.523 |

MAXENT:

The observations are as follows:

|  |  |
| --- | --- |
| Precision | 0.465 |
| Recall | 0.470 |
| F-Score | 0.465 |
| Accuracy | 0.477 |

Also cross validation was performed to observe how they stack up against independent data set. For this K fold cross validation was performed, which will give validation on 3 subsets of training data sets. It is generally used to get the accuracy on independent data. It first partitions the data into K subsets and then performs validation on it such that out of k subsets, 1 is training data and the rest are training data. It is generally used to get good accuracy of the classifiers which can be then used for their analysis. It performs validation for all k subsets. The results observed were as follows:

For K=3 for all the below algorithms,

K- Fold validation for SVM:

|  |  |
| --- | --- |
| Fold 1 | 0.72 |
| Fold 2 | 0.72 |
| Fold 3 | 0.72 |

K- Fold validation for GLMNET:

|  |  |
| --- | --- |
| Fold 1 | 1 |
| Fold 2 | 1.13 |
| Fold 3 | 0.79 |

K- Fold validation for MAXENT:

|  |  |
| --- | --- |
| Fold 1 | 0.64 |
| Fold 2 | 0.65 |
| Fold 3 | 0.67 |

K- Fold validation for TREE:

|  |  |
| --- | --- |
| Fold 1 | 0.63 |
| Fold 2 | 0.65 |
| Fold 3 | 0.64 |

From the above comparison it is clear that, the other supervised learning model are more efficient in accuracy than naïve Bayes which gives a bit less accuracy as compared to others.

So from here on we will analyze the rest of complex supervised algorithms as they are more efficient than simple naïve Bayes.

Also, in sentimental analysis of movie reviews not just the accuracy but all measures are important for analysis the performance of algorithms to know whether they give correct polarity. In order to this we will observe the precision, recall and F Measure of these algorithms and find out which method tends to perform better than the other, in finding the polarity of movie reviews. Of these F Score is more important as it gives the combined score of precision and recall. It balances the performance of both precision and recall and then gives the score which is more all round than any other score.

From the obtained observations each of Bagging, Boos GLMNET, and Decision tree has a low F Score and thus can be safely eliminated as F-Score gives the whole picture in terms of performance of algorithms.

Of the above observed results it seems that, though when these models are used as a container they provide better results than naive Bayes in terms of accuracy, recall, precision and f measure but when implemented individually are not an improvement on naïve Bayes. In fact considering each algorithms advantage and disadvantage only MAXENT, SVM, SLDA and Random Forest can be used as a substitute for naïve Bayes to classify the polarity of data. But out of that too, though Random Forest provides a good measure of F Score ,they are not space optimal, and when combined it with respect to these observations, it seems costly to implement random forest given there well known problem with space optimization and over fitting. In addition to it they are also complex once data gets large. So these disadvantages of over fitting, space and complexity overshadow these observations which don’t provide much improvement over already used methods. So it can be observed from this that Random Forest and Decision trees can be eliminated for implementing sentimental analysis. SVM certainly provides consistent result over a small dataset, but as the data set increases, they get overly complex also, their accuracy goes on decreasing and in some cases go below the accuracy of other algorithms. So it can be observed that for small datasets, SVM can be used but should be avoided for larger datasets. Another of SLDA as can be observed from the results , provides consistent results and also is not that complex. Also the fact that it was specially developed for sentimental analysis of reviews provide it with a distinct advantage over the other methods. For small datasets though there is not much significant gain in calculating precision and recall as the data set goes on increasing, it is believed, it will provide a good measure. But it is still a work in progress, as more and more research is going on that classifier. Also another thing which falls in its favor is that, it is adapted form an unsupervised technique of LDA, so it has advantage of working on large dataset as well. It is expected that in the future it would be able combine the advantage of supervised and unsupervised learning algorithms and be more efficient.

But for present conditions it can be observed that probabilistic classifiers as MAXENT and naïve Bayes are best to use to classify the polarity of the data. The probabilistic classifiers have an inherent advantage that they are simple to implement and work on datasets irrespective of the size. Combining it with their performance, these algorithms provide excellent classifiers which are simple and efficient compared to others. Though these classifiers provide good result as can be observed from the above table, It is better to pre process the data first and then implement algorithms on it. The raw data sometimes consist of noise and interference and performing processing on them can improve their results more. Using negation handling and using sequencing of words rather than individual can help improve their result more.

According to me. I will suggest that naïve Bayes currently is better of all algorithms as it is simple to understand and implement. Also it provides good result on small as well as large dataset. Intact from the results above, it provided better accuracy, precision, recall and F measure than all of SVM, SLDA, Bagging, Boosting, Decision Tree, Random Forest, GLMNET and MAXENT individually. Also due to its simple implementation it uses less computing power than any other classifier. These properties provide naïve Bayes with a clear advantage over other classifier and, thus I recommend naïve Bayes Classifier for sentimental analysis.

# Related Work

Due to the importance movie review many research work is going on in sentimental analysis. In fact as previously mentioned, SLDA technique was specifically designed for classifying the polarity of movie reviews. In their paper Supervised topic models David M. Blei and Jon Mcauliffe proposed the SLDA technique based on the unsupervised learning. This is an interesting method proposed by them. It also looks that this method will be used on large scale in future. But for now, as it is in its development stage, it is not that effective. In addition to it a competition was held on Kaggle, to classify the polarity of movie reviews from rotten tomato database. Here too large entries were provided, each displaying there own work on Movie review polarity classification. Each person had its own approach. Also, Socher et al conducted research and went into details of sentimental analysis in his paper Sentimental Tree Banks. The first dataset of movies was obtained by Lee and Pang for their research on sentimental Analysis. They published it in their paper “Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales”. Bo Pang and Lillian Lee carried further research and published their work in Opinion Mining and Sentimental Analysis. A. Amolik further researched on movies sentimental analysis and proposed a method of using methods of Twitter Sentimental analysis on movie reviews to identify the polarity of the reviews. As mentioned previously, my method of classifying the polarity of reviews was based on this approach.

# Summary

A review describes and analyzes different aspects about a product and the experience of user with respect to those aspects. Now a days, with the advent of social media even Identifying the polarity of movie reviews has vast commercial base. To take advantage of this information present. we use Sentimental Analysis. Sentimental Analysis involves classifying the polarity of the reviews. There are many ways to implement sentimental analysis. The Sentimental Analysis classification can be classified into two types Machine Learning Approach and Lexicon based approach. The Machine learning approach is further categorized into supervised learning and unsupervised learning approach. The Lexicon based approach is classified into Dictionary based approach and corpus based approach. Now in this project I have implemented supervised learning Machine Learning based approach to implement sentimental analysis to classify the polarity of reviews. The machine learning algorithm implemented by me in this project are naïve Bayes, SVM, Random Forest, Decision Tree, MAXENT (maximum entropy) and GLMNET (generalized linear model).NNET(Neural Network), SLDA(Supervised Latent Dirichlet Allocation), Bagging and Boosting.

In my project, I used the concept of providing tweets as training data set and movie reviews as a testing dataset.

After implementing these algorithms, it was found that the best classifiers while implementing the approach were naïve Bayes and Maximum Entropy. Between the two, I would say, naïve Bayes has more advantage due to its simple implementation and requiring less computing power.

# Conclusion

In this project I discussed, how important it is in todays world, for sentimental analysis. I also discussed the importance of classifying the polarity of movie reviews. I then implemented various classifiers to classify the polarity of movie reviews. I then found that, with my approach, naïve Bayes proved to be the best algorithm to implement sentimental analysis on movie review data set.

References

1. https://www.ukessays.com/essays/film-studies/the-importance-of-online-reviews-film-studies-essay.php
2. http://www.rogerebert.com/balder-and-dash/why-is-film-criticism
3. “An Approach to Text Mining using Information Extraction “Haralampos Karanikas, Christos Tjortjis and Babis Theodoulidis
4. “Text Mining: The state of the art and the challenges” Ah-Hwee Tan
5. Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P. (1996). From data mining to knowledge discovery: An Overview. In Advances in Knowledge Discovery and Data Mining, U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, eds., MIT Press, Cambridge, Mass., 1-36.
6. Text Mining Scientific Papers: a Survey on FCA-based Information Retrieval Research Jonas Poelmans1,5, Dmitry I. Ignatov4 , Stijn Viaene1, 2, Guido Dedene1, 3, Sergei О. Kuznetsov
7. http://www.sciencedirect.com/science/article/pii/S2090447914000550
8. Sentiment Analysis Algorithms and Applications : A survey. Walaa Medhat, Ahmed Hassan
9. “Sentiment Analysis of Short Informal Texts” Svetlana Kiritchenko,Xiaodin Zhu, Saif Mohammed
10. “Sentiment Analysis a combined approach” Ruby Prabawo , Mike Thellwall
11. “Nearest Neighbor Norms: NN Pattern classification Technique” Belur V Dasarathy
12. Shakhnarovish, Darrell, and Indyk, ed. (2005). Nearest-Neighbor Methods in Learning and Vision
13. Mozina, M.; Demsar, J.; Kattan, M.; Zupan, B. (2004). Nomograms for Visualization of Naive Bayesian Classifier
14. Maron, M. E. (1961). "Automatic Indexing: An Experimental Inquiry"
15. Steinwart, Ingo; and Christmann, Andreas; Support Vector Machines, Springer-Verlag, New York, 2008.
16. Theodoridis, Sergios; and Koutroumbas, Konstantinos; "Pattern Recognition", 4th Edition, Academic Press, 2009
17. “Thumbs up? Sentiment classification using Machine Learning Technique.” By Bo Pang Lillian Lee
18. http://www.saedsayad.com/naive\_bayesian.htm
19. http://www.slideshare.net/JhihMing/adaptive-web-page-content-identification-6558704
20. http://www.slideshare.net/0xdata/generalized-linear-models-with-h2o
21. http://cse-wiki.unl.edu/wiki/index.php/Bagging\_and\_Boosting
22. “Twitter Sentimental Analysis of movie review using Machine Learning” by A Amolik
23. “Feature Selection and Classification Approach “ by Gautam Tripathi
24. Franco Salvetti, Stephen Lewis, and Christoph Reichenbach. “Impact of Lexical Filtering on Overall Opinion Polarity Identification.”
25. Philip Beineke, Trevor Hastie, “The Sentimental Factor: Improving Review Classification via Human-Provided Information.”
26. “Finding topic-specific strings in text categorization and opinion mining contexts” by Remi Lavalley, Clhoe Clavel,

1. [↑](#footnote-ref-1)