Indian Journal of Science and Technology, *Vol 10(19), DOI: 10.17485/ijst/2017/v10i19/112756, May 2017* *ISSN (Print) : 0974-6846* *ISSN (Online) : 0974-5645*

Opinion Mining and Analysis of Movie Reviews

Vibhor Singh, Priyansh Saxena, Siddharth Singh and S. Rajendran

Department of Information Technology, SRM University, Kattankulathur, Chennai – 603203, Tamil Nadu, India;
vibhorsingh19495@gmail.com, priyansh.saxena05@gmail.com,
its.siddharth2308@gmail.com, rajendran.s@ktr.srmuniv.ac.in

Abstract

Background/Objective: Customer reviews are important for various fields (e.g. Movies,Products, Services). Moviereviews plays vital role in describing its success and failure. People have now become very specific on what movies to watch and what not to watch. Hence people don't want to waste time on a movie that has bad reviews. Nowadaysonline reviews are important for personal recommendation. **Methods/Statistical Analysis:** There are various works done on opinion mining and text mining. This paper focuses on analyzing sentiment from semantic orientation of words that occur in a text by manually defining dictionary for positive, negative and intensifier words stored in different text file. In this method we extract data from three different sources. Opinions that has to be analyzed is preprocessed and stored in text file. The data are then compared with our bag of words in order to find the number of positive and negative sentiments in the reviews of that particular movie. To predict movie rating three machine learning algorithms have been used to create three different classifier using a trained data set. **Findings:** After rating prediction, the accuracy of each classifier is calculated on the data set containing predicted rating of various movies given by each classifier. Out of all three machine learning algorithms used, Naive Bayes is found to be more accurate than Decision Tree and K Nearest Neighbour algorithm. **Improvements:** There can be more many ways where we understand how the users write reviews as the reviews only support English language. Hence we can also bring in multiple languages, so that we do not even miss out on a single review. Most of the time,a single user may not provide review for all the movies. So, our database will resemble a sparse matrix. A prediction algorithm may be designed to mathematically guess the rating for movies that are not originally reviewed by the user.

Keywords: Machine Learning, Opinion Mining, Sentiment Analysis, Text Mining

1. Introduction

There are lots of topics where analysis could play a major role in its development. With the advancement in the technological field and having internet access many people are giving their opinions in various internet sites and blogs. The option of the people plays a major role in the development as well as improvement in the upcoming applications. For example, any proposal enforced by the government can be first reviewed by the people and then registered as a rule to avoid disagreement. There has been a lot of research work that is being done on opinion mining and sentiment analysis. In order to make a proper sentimental analysis tool first we need to learn how to extract data from different resources available on internet and then applying a proper algorithm to analyze the polarity(positive or negative) of the extracted data.The analyzed data can be used for improvements or as guide to other people. This paper focuses on how we can extract the movie reviews from different sources and then applying a dictionary based algorithm to find whether the movie is good at the box office or not. The reviews of the people are taken in study and then the analysis is done on it and manifested in the form of various forms like pie charts, graphs and gifs.

**Author for correspondence*

2. Existing Work

There have been many research and algorithm made on opinion mining and sentiment analysis on the reviews given by people on various topics for social topics.There are various data sets available on net and different methods are available to extract these data and perform the analysis . Tools like R Studio, Weka, Rapidminer are also available for various data mining and analysis.Variousmethods toperform sentiment analysis have been described[1] using supervised and unsupervisedway of document level sentiment analysis with brief of evaluating sentimentClassification. Different source of data extraction,basic criteria used in sentiment analysis[2] and their application is discussed. The process of data extraction itself consist of many procedure of making data suitable for analysis such that more accuracy will be there in final outcome. An example of collaborated opinion mining[3] is discussed for a student from various teacher, these papers summarize how to implement the technique to detect various sentiment patterns available as word, sentence or in document format.

In this paper, a database of sentiment words has been used for analysis of opinion and every sentiment word in the database has been given a value. When a sentiment word is detected in a sentence the value saved in the database is used for evaluating the cumulative opinion value.Various works have been done to extract movie review datasets from movie review websites like IMDB4,5 which focuses on how to handle large amount of movie review data sets that can be used for opinion mining with feature extraction and pre-processing followed by building recommendation system that recommends a movie to user based on his previous reviews. Three methods for extraction of meaningful features(Bag of Words,N-Gram Modelling,TF-IDF Modelling) from the review text which could be used for training purposes using different machine learning algorithm.The general order of performance for the model was Logistic Regression> NaÃ´rveBayes > SGDClassifier>RandomForestClassifier >kNNClassifier.

Social networking sites like Twitter6 which uses trained data to build a classifier and perform sentiment analysis on tweets that includes modeling of feature vector and classification of tweets using naive bayes classifier and support vector machine.According to the7 BOW model, the document gets a representation as a vector of words in Euclidean space where each word is independent from others. This bag of individual words is commonly called a collection of unigrams.The independence of uni grams means that the appearance of one uni gram in the text will not influence the appearance of any other uni gram. Ark Tweet NLP library which was developed by the team of re-searchers from Carnegie Mellon University and was specially designed for working with twitter messages. Ark Tweet NLP recognizes specific to Twitter symbols, such as hash tags, at-mentions, re tweets, emoticons, commonly used abbreviations, and treats them as separate tokens.In twitter users usually use hash tags to mark topics8 and perform sentiment analysis of the reviews extracted that includes emoticons in form of text, apart from that different types of model like uni-gram ,senti-features etc have been explained and implemented and their accuracy is compared.

To build a Recommendation/prediction system different types of machine learning algorithms are there for it. Using a collaborative approach9 consider the user data when processing information for recommendation. For instance, by accessing user profiles in an on-line music store, the RS has access to all the user data, such as the age, country, city, and songs purchased. Another example is to build a restaurant recommendation system10 using the Yelp Data set that uses different machine learning techniques for recommendation.In a recommendation system, there are two classes of entities, which are users and items.Users have preferences for certain items. A recommendation system main task is to extract the preference information from the user data. The data itself is represented as a utility matrix. Given each user-item pair, a value in the utility matrix represents what is known about the degree of preference of that user for that item. Values may come from an ordered set, e.g., integers 1-5 representing the number of stars that the user gave as a rating for that item. This utility matrix is typically sparse, as most users do not rate most items.Most recommended systems take three basic approaches: collaborative filtering, content based filtering and hybrid method.

The use of various linear classifiers11 in a model-based approach to the recommendation task and in a recommending system, they are also interested in the likelihood that a customer will accept a recommendation.To understand different machine learning techniques12and methods to evaluate and compare them have been described which can be used for choose best suited algorithm.These machine learning algorithms can be used for rating prediction and recommending a movie based on his previous choice. The popularity13of a movie is a very important factor to find, which gives impact of it towards viewers and also effects movie rating.It is implemented using Machine learning to classify whether a movie is popular or not after extracting and analyzing data from IMDB. It includes various process in which machine learning algorithm uses feature selection to train data and then classification of test data. Movie rating prediction is a very complex process which is done by using hybrid methods14 of combining the singular value decomposition method (SVD) and the K nearest neighbors algorithm (KNN) to predict the empty rating for each user.Good ratings prediction requires the correct interpretation of the available data about both the user and the item.

3. Proposed Methodology

The objective of the project is to analyze the sentiments of the movie reviews from various sources available on net and to predict the rating of the movie based on various criteria using machine learning. Figure 1 shows the flow of various process for analysis

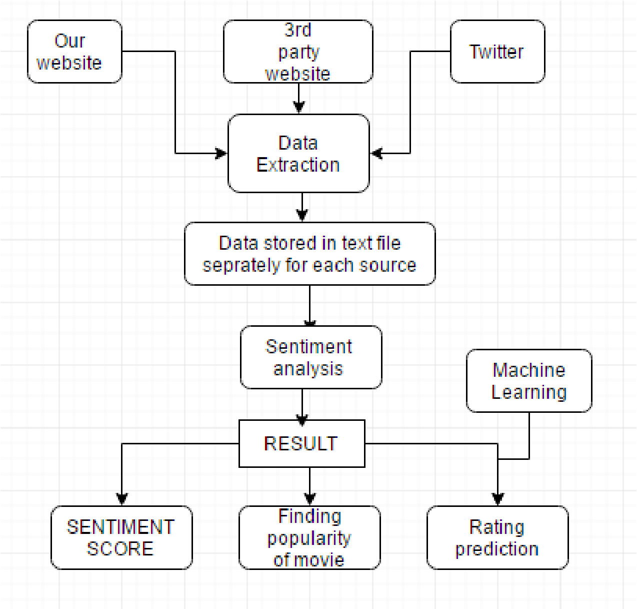


Figure 1. Flow Diagram.

3.1 Data Sources

The different data sources considered during extraction include our own website, third party website like Flixster and the most important source for analytics that is twitter. The data is extracted from our website and using PHP we store the data into a text file and then the text file is downloaded and attached to the application to perform the analysis. The data from twitter is extracted using Twitter4J. The user has to type in the movie with the hash tag and all the tweets related to the movie will be extracted. The data from the third party website is extracted using HTML tags using jsoup .

3.2 Dictionary Creation

The data extraction and analysis is based on dictionary based algorithm. Manually created dictionary is constructed for positive, negative and intensified words in lower case. The different words are stored in a text file named positiveword.txt, negativeword.txt and intense.txt. For example words like horrible, bad, worse, boring, and ordinary, brutal are stored in the negative.txt file. The positive words like good, excellent, brilliant, amazing, charming, hilarious etc. comes under positive.txt file. There is a separate dictionary for intensifiers like very, extremely, highly, too etc. which are saved in intensifier.txt. All these intensifiers are added up with the favorable and unfavorable words and increase the polarity of the word. For example if a user writes a review “The movie was good”, the positive sentiment score will be +1 and same polarity for negative sentiments. But if the user writes “The movie was very good” the sentiment score will be +2 . The algorithm also takes into account the smiles like “ :D” , “ :)” , “ :(” . The algorithm handles negation words like “not at all good”, “not bad ” which can totally change the review meaning.The review of movies is stored in text file format. Table 1 few positive, negative and intensifier words are shown.

Table 1. Dictionary Content

Positive	NEGATIVE	INTENSIFIER
Best	boring	Very
Absorbing	bad	Extremely
Enjoyable	disappointing	Highly
:)	:(	Far

3.3 Data Preprocessing

After the extraction of data from different sources, the data is converted in lower case since all the words in the dictionary are in lowercase. The data processing from twitter includes only English language. The extraction of data via "Flixster" includes only reviews and parameters consisting like and dislikes percentage.

3.4 Sentiment Analysis

After the conversion the data stored in the text file for that movie is compared with bag of words from the dictionary. The dictionary consists of various positive and negative words, negation words like not , so , neither etc and intensifiers like very, so etc. The algorithm compares the next word after the negation or intensifier word in order to find out the sentiment of the user. The algorithm also counts the type of smiley's and adds up to the count of the positive and negative words. There is also an option to search about a specific keyword like acting, cast , direction etc in order get an in depth view about the movie. This helps the viewer to get a better idea whether they should watch the movie or not based on the keywords they are interested in. Hence the algorithm works mainly on the bag of dictionary and keywords for which the user wants to get an analysis forTable 2 shows the polarity assignment of different sentiments used for sentiment analysis.

Table 2. Polarity assignment

	POSITIVE	NEGATIVE	INTENSIFIER
POLARITY	+1	+1	+2

3.5 Machine Learning

After sentiment analysis the rating prediction comes into play which is done with help of naive bayes, decision tree and K nearest neighbors. It uses trained data and predicts the movie rating of test data based on different attribute values of a movie.

4. Results and Discussion

For easy use an UI is designed which makes the understanding and analysis simpler as shown in Figure 2. It contains of four variety of tools which are the different sources from which we are extracting the data. The first data source is the extraction of data from our own website. All the registered users will write the reviews on our website and the data will be downloaded from the database and then the algorithm is applied on the downloaded data. The second source allows us to give the URL of a website where people have already given their reviews. The data is extracted using web scraper and then the data is stored in a text file. The file containing data is attached to the tool and algorithm is applied on it. The third module uses twitter. Twitter is the also the source for analysis since many people write all their views on twitter. Twitter has grown exponentially in terms of number of users and its usage for analysis. Thus twitter analysis will give us an in-depth analysis about the movie. Since many critics write their reviews on twitter it is a very important platform which needs to touched for analysis. During analysis through twitter just write the movie name followed by a hashtag and you will get all the reviews related to the movie. We use Twitter4j for extracting the data from twitter. The last module is an in depth analysis about the movie. Mostly people want to know how their superstar acted in the movie or how good was the action , romance , comedy etc of the movie. Nowadays not only movie reviews but specific reviews about each and every term related to the movie play a vital role in determining how the movie is at the box office. The movie file is taken by tool and then write the category for which you want to find the review like acting, comedy, direction etc and it will display the result suitably.

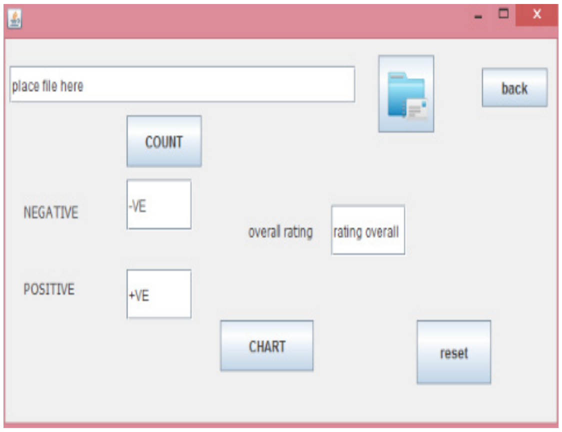


Figure 2. Sample user Interface.

The algorithm evaluation displayed in a pie chart to calculate the total number of positive and negative reviews percentageas shown in Figure 3.

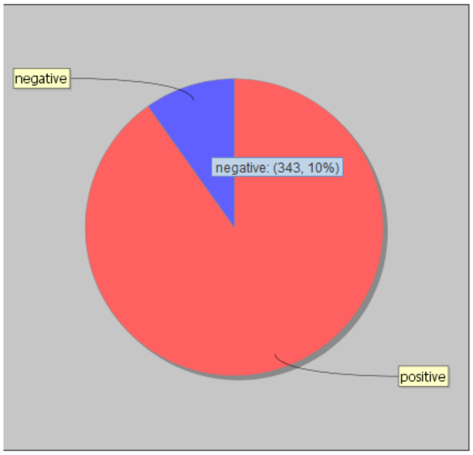
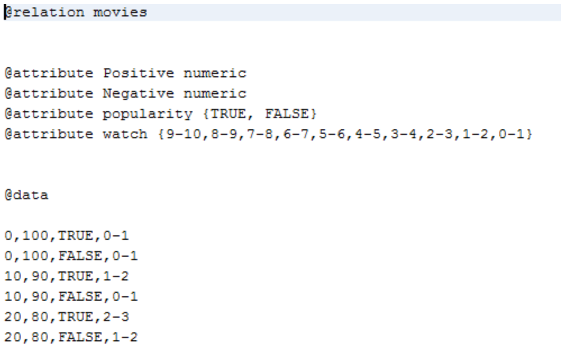


Figure 3. Pie Chart Displaying Negative and Positive Reviews in Percentage.

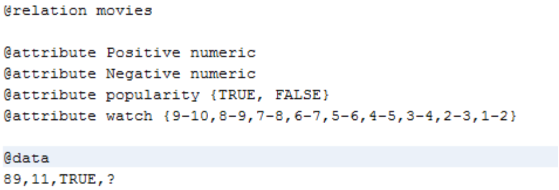
4.1 Training data for Machine Learning Algorithm

To give the rating of the movie , machine learning algorithms are used with the help of trained data as shown in Figure 4. The data is trained based on movies popularity and positive and negative comments percentage according to the which the rating of the movie is given. Hence the paper has implementation of both sentiment analysis on the reviews and machine learning classifiers for predictionon test data as shown in Figure 5.



```
@relation movies


@attribute Positive numeric
@attribute Negative numeric
@attribute popularity {TRUE, FALSE}
@attribute watch {9-10,8-9,7-8,6-7,5-6,4-5,3-4,2-3,1-2,0-1}


@data

0,100,TRUE,0-1
0,100,FALSE,0-1
10,90,TRUE,1-2
10,90,FALSE,0-1
20,80,TRUE,2-3
20,80,FALSE,1-2
```

Figure 4. Training Data.

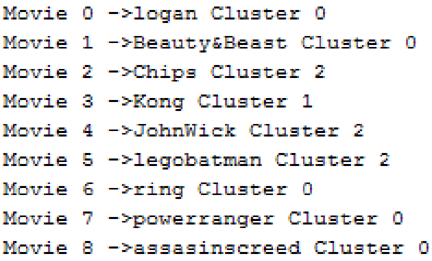


```
@relation movies

@attribute Positive numeric
@attribute Negative numeric
@attribute popularity {TRUE, FALSE}
@attribute watch {9-10,8-9,7-8,6-7,5-6,4-5,3-4,2-3,1-2}

@data
89,11,TRUE,?
```

Figure 5. Applying Naive Bayes, KNN, Decision Tree on Test Data.

4.2 K Means Clustering

After prediction of movie ratings ,cluster is formed using K Means clustering according to rating range. Figure 6 as is an exampleshowing some movies are cluster.



```
Movie 0 ->logan Cluster 0
Movie 1 ->Beauty&Beast Cluster 0
Movie 2 ->Chips Cluster 2
Movie 3 ->Kong Cluster 1
Movie 4 ->JohnWick Cluster 2
Movie 5 ->legobatman Cluster 2
Movie 6 ->ring Cluster 0
Movie 7 ->powerranger Cluster 0
Movie 8 ->assasinscreed Cluster 0
```

Figure 6. Clustering Based on Ratings Predicted.

5. Conclusion

Thus this paper uses dictionary based algorithm for sentiment analysis. The polarity of the reviews are a key factor in determining the overall sentiment. The study not only concentrates on the sentiment of reviews but also predicts the rating of the movie using Machine learning class. The data is trained based on features like popularity and positive and negative polarity percentage then the rating is given for a particular movie. We have used different machine learning classifiers result to compare their accuracy. Out of all these algorithms, Naive Bayes gives the best result. The accuracy of different classifiers tells about which classifier is the best. Let's take a look at the following measures mentioned in Table 3.

Table 3. Performance of different classifiers

MEASURE \ MODEL	NAIVE BAYES	DECISION TREE	KNN
ACCURACY	54.10%	44.26%	50.28%
TP	0.66	0.56	0.46
TN	0.95	0.95	0.94
FP	0.05	0.05	0.06
FN	0.34	0.44	0.54
PRECISION	0.73	0.46	0.56
RECALL	0.66	0.56	0.46
F-MEASURE	0.65	0.48	0.47

True Positive weighted(TP): rating predicted positive that are actually positive
True Negative weighted (TN) :rating predicted negative that are actually negative

False positive weighted(FP) :rating predicted positive that are actually negative
False Negative weighted(FN) : rating predicted negative that are actually positive
Accuracy : Closeness to a standard value=(TP+TN)/(TP+TN+FP+FN).
Precision:fraction of true positive upon total predicted positive=TP/(TP+FP)
Recall: fraction of true positive upon actual positive=TP/(TP+FN)
F-measure: the harmonic mean of precision and recall= 2x(Precision x Recall)/(Precision +Recall)

6. References

1. Hajmohammadi MS, Ibrahim R, Ali Othman Z. Opinion mining and sentiment analysis survey. International Journal of Computers and Technology. 2012; 2(3):171–8.
2. Vinodhini G, Chandrasekaran RM. Sentiment analysis and opinion mining. International Journal of Advanced Research in Computer Scienceand Software Engineering. 2012; 2 (6): 282–92.
3. Virmani D, Malhotra V, Tyagi R. Sentiment analysis using collaborated opinion mining. Cornell University Library; 2014.
4. Goyal A, Parulekar A. Sentiment analysis or movie reviews. Semantic Scholar. 2015.
5. Andrew LM, Raymoi ED, Peter TP, Huang D, Ng AY, Potts C. Learning word vectors for sentiment analysis. ACM Publications. 2011; 1: 142–50.
6. Amolik A, Jivane N, Bhandari M, Venkatesan M. Twitter sentiment analysis of movie reviews using machine learning techniques. International Journal of Engineering and Technology. 2016; 7 (6): 1–7.
7. Kolchanya O, Souza TPT. Twitter sentiment analysis lexicon method machine learning method and their combination. Cornell University Library. 2015; 1: 1–32.
8. Agarwal A, iXie B, Vovsha, Rambow O, Passonneau R. Sentiment analysis of Twitter data. ACM Publications; 2011.p. 30–8.
9. Portugal I, Alencar P, Cowan D. The use of machine learning algorithms in recommender systems. A systematic review. Cornell UniversityLibrary. 2015; 1: 1–16.
10. Gao M. Application of machine learning techniques to recommendation system. eScholarship University of California; 2015.
11. Zang T, Iyenagar SV. Recommender systems using linear classifiers. Journal of Machine Learning Research. 2002; 2: 313–34.
12. Khairnar J, Kinikar M. Machine learning algorithms for opinion mining and sentiment classification. International Journal of Scientific and Research Publications. 2013; 3(6): 274–609.
13. Latif MH, Afzal H. Prediction of movies popularity using machine learning techniques. International Journal of Computer Science and Network Security. 2016; 16(8): 1–5.
14. Mladen MaroviÂťc, Marko Mihokovi. Automatic movie ratings prediction using machine learning. IEEE Conference Publications. 2011.p. 1640–45.