International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT239013

Sentiment Analysis to Predict Movies Success Rate Based on NLTK Movie Review Corpora Using Machine Learning

Muhammad Luqman*1, Amir Yaqoob2, Majid Bashir Ahmad2, Kanza Majid3
*1Department of Computer Science, University of Gujrat, Gujrat, Pakistan
2Department of computer science, University of Lahore, Lahore, Punjab, Pakistan
2Department of computer science, University of Lahore, Lahore, Punjab, Pakistan
3Department of computer science, National College of Business Administration & Economics, Lahore, Punjab, Pakistan

Article Info

Publication Issue :
Volume 9, Issue 1
January-February-2023

Page Number : 46-54

Article History
Accepted: 01 Jan 2023
Published: 19 Jan 2023

ABSTRACT

With the proliferation of social networks, peoples express their opinions about different things or issues on social media without any hesitation. The rapid growth of textual data on social media are required to develop algorithms and techniques for recognizing people's opinions towards specific subject. These opinions are helpful in business plans development, marketing trends, political parties' popularity. The film industry can be an important revenue generating industry of any country. Peoples express their opinion on movie trailer using social media. The effective sentiment analysis of opinions on social media such as Twitter can be helpful to predict movie ratings. This research work focuses on developing a technique to predict movie success rate on the basis of tweets data. We have collected tweets about different movies after their trailer released by using hash tag method. We applied Sentiment analysis approach using Machine learning. In this study we utilized four key algorithms (Naïve Bayes, SVM, Neural Networks, decision tree) on NLTK Movie review corpora.

Keywords : Social networks, Sentiment Analysis, Machine learning

I. INTRODUCTION

This In our research, we are going to predict the popularity of a movie through sentiment analysis of social media and also proposed a movie rating system. People on social media allowed freely to express their opinions and suggestions about any topic. So the social media can be the source to collect the user's reviews on a particular topic [1].

User-generated data on the social media is increasing day by day. In 2013 SINTEF published a research report in which tells that 90% data on the web or internet was produced in the just past two years and this speed of increasing data still continue. The World Wide Web (www) is a prove of large data because it has approximately 4.55 billion pages on the web. In this large data, it can be difficult to find our required information [2]. Critical part of collecting information about a new thing, we should find the information that what other people think about it. It is the human nature deciding they must talk and collecting recommendations from other peoples. for example, if we want to buy a new car, we must ask our relatives or friends that which car is best for us. Large amount of reviews data on social media makes it easier to find out the related opinions of those peoples we have never met before [3]. The fastest growth of information on social media this makes easy for buyer's decision making. And on the other hand, sellers can improve their products adds on the bases of this information [4]. Social media plays an important role in decision making (e.g., forum, discussion, comments, blogs, twitter, microblogs and posting on social network websites). For example, if someone wants to buy a product one should need some opinions about this product so now this one is not bound to asking one's family or friends for opinions about the product because there are lots of reviews/comments and suggestions on social network sites about the product. But there are some issues for average readers because they can feel difficulty to identify those sites and extract and summarize the useful opinions from a large number of opinions. So the automated sentiment analysis systems are used to solve this problem [5].Sentiment and websites to post the feelings of people have been used by the people in recent years. Sharing opinions or current issues/topic is their main purpose. Therefore, among the Internet users, a networking tool has become called the micro-blogging service. From Alta Plana's Text Analytics 2014 research study [6], there is a graph in which based on 962 sources of textual information 216 respondents were surveyed.

Our research is about success rate prediction on the bases of reviews. For this purpose, we use Twitter as a data source. Because twitter now a day is a very popular source for expressing people's opinions about any topic. When a movie trailer is released, twitter becomes a platform on which peoples can express their thoughts or opinions about the movie after watching the trailer. They can judge the movie on the basis of cast, cast popularity, music, graphics, or story highlights which is given in a movie trailer. After watching trailer, they express their likes and dislike about the movie in the form of textual form which is tweets. We have collected these tweets on the trailer about the particular movie before its release. This research investigates how to pre-process tweet data and applying machine algorithms and compare their results and also check that whether these results are useful or not for achieving our goal of success rate prediction of a movie.

Sentiment analysis

Sentiment analysis focuses on the attitude but on the other hand traditional text mining concentrate on fact analysis. Sentiment analysis involves sentiment classification of features and characterization of opinions. Sentiment classification concerns with classifying all the documents which are relevant to a particular object. But if we talk about the sentiment classification of features this will concern with the object features. In opinion characterization, only those opinions are mined which are given by the users of a **Sentiment analysis for reviews :**

targeted object or topic. Chinese and English languages are studying mostly in the past few years some of the researchers doing work on sentiment analysis for "Thai", "Arabic" and "Italian" languages. Sentiment analysis was an emerging field in 1990, but in 2000 this becomes a great emerging subfield of data management [7].To evaluate the movie rating the reviews about the movie are the most important. We can know from the textual reviews of a movie that the movie meets the reviewer's expectations or not. Mostly the movie reviews posted on different sites are unstructured and informal and not follow any grammar pattern. For solving this issue, we should use sentimental analysis on textual data.

Sentiment analysis in reviews is the way of investigating item surveys on the web to decide the general sentiment or feelings about an item. Reviews speak to the so-called user-generated content, and this is of developing consideration and a rich asset for showcasing groups, sociologists, and analysts and other people who may be worried about assessments, sees the open state of mind and general or individual states of mind [8].The immense number of audits on the web speaks to the present type of client's input. Choosing about the slant of sentiment in a survey is a testing issue because of a few elements. One issue is the subjectivity in the writings and the need to recognize supposition bearing from non-stubborn sentences. Another issue that makes it difficult to characterize surveys is the alleged "foiled desire" which implies that the essayist composes many sentences one way which can be comprehended as positive, and after that closes with one negative sentence that turns around the importance of the whole content [9]. This is the reason there ought to be better techniques for highlight choice. Turnery [10] infers that motion picture audits are difficult to arrange because the general conclusion spoke to in the survey is not really the whole of all of the conclusions said in the content. Many investigations have investigated different techniques in examining item audit assumption, some of which utilize machine learning approaches [9][11], and some utilization lexical techniques [10] [12].

The way that there is an expansive number of online surveys and that some of these contain just a little division of sentences that express a conclusion makes it harder for clients' and organizations to know the general feeling about an item and to set up a conclusion about it. Thus, assessment condensing is proposed. Feeling condensing could be led by revising the first content and centering on the primary point. It additionally can utilize angle level notion examination, and select the primary elements in the writings, their angles, and the relating conclusions to create an "element-based survey outline". Cases can be found in [13] [14].In this theory, film audits are examined, utilizing a standout among the most well-known information sets in sentiment analysis. The issue is moved toward utilizing machine learning approach, and the feeling holder point of view.

Machine learning Approach:

Machine learning could be considered as a part of artificial intelligence. Supervised Learning and Unsupervised Learning [15] are the main two categories that are included in machine learning. Usually, on concluding information about the characteristics of sets of data machine learning algorithms work well. For the process of sentiment analysis, natural language processing (NLP) techniques from which sentiment features selected and extracted and also the success of machine learning relies on them.

Machine Learning Approach for Tweet Sentiment Analysis:

From the machine learning approach, the majority of research work was focused on product reviews that are highly subjective English normal texts. On the other side, additionally, there has been also some multilingualism, short and messy sentiment analysis of tweets. With distant supervision sentiment of Twitter messages that are automatically classified with the help of machine learning algorithms, was presented by Go at al. [16]. Noisy labels of Twitter messages were used by them in order to show that pre-processing steps are essential. For making distant supervised learning feasible, after training three machine learning algorithms they found using tweets with emotions. Twitter API for sentiment analysis that automatically collects twitter corpus, was used by Pak and Paroubek [5]. Based on the multiracial Naïve Bayes Classifier they built a classifier with the help of with a document could be annotated in negative, positive and neutral sentiments. Using social relationships the performance of user-level sentiment analysis could be improved, it is founded by Ten et al. [17] . For example, to refer other users for reasons users of Twitter often use the symbol “@”. In this manner referring to the other users automatically alert them. According to Ten et .al that people may hold similar opinions who use “@” symbol to connect to each other. Therefore, in their experiments "@" symbol became an important feature. Their outcome demonstrated that incorporating link information from Twitter could enhance the performance of sentiment classification essentially in light of Support Vector Machine. A method based on the context and syntactic relationships for target-dependent twitter sentiment classification was introduced by Dong et.al [18] called Adaptive Recursive Neutral Network (AdaRNN). AdaRNN contained various composition functions. In the experiment, an interested target was created with the given tweets. With a specific end goal to get more attention to tweets of different languages (not English), a few researchers intended to continue Twitter sentiment analysis on another domain Kaya et.al [19].Naïve Bayes, Support Vector Machine and Maximum Entropy are three machine learning classifiers with them Kaya et.al [19] did sentiment analysis of Turkish Political columns. Their inspiration was deciding positive and negative opinions from entire documents without considering subjects. From unlabeled tweets data to labelled political columns, to improve classifications performance Transfer Learning was used. In an unsupervised way extracting features from unlabelled data and accuracy of sentiment classification can be improved with the help of transfer learning method. A first ranking algorithm with bi-grams and uni-grams and second skip-grams to the speech processing are two approaches provided by Fernandez at.al [20] for sentiment analysis of Spanish tweets. Positive, strongly positive, none, strong negative and negative are five levels in which the polarity of Spanish Tweets was divided. As data set in the paper [21], movie reviews were used. In the paper, standard machine learning strategies were used which out-perform human-produced baselines. As compared to traditional topic-based categorization the three machines learning techniques do not perform efficiently. From the performance point of view, SVMs performs worst. In a paper [22] the documents are classified as positive, negative and neutral with the help of a system called Document Level Opinion Mining System. Negation also handled by the proposed systems from IMDB movie dataset used by experimental. POS tagging used in a movie review; Document Based Sentiment Orientation System performs better than AIRC Sentiment Analyzer. Movie reviews dataset that contain 1000 positive and 1000 negative reviews are classified and in the paper [17] for this classification, three supervised machine learning algorithms such as KNN, SVM, and Naïve Bayes were compared. It is recognized in this approach that the training dataset had a larger number of reviews and a large number of reviews and as compared to Naïve Bayes the SVM approach performed well. In classifying data correctly more than 80% accuracies acquired by SVM approach, for all three algorithms about movie reviews in sentiment classification when a huge amount of training dataset containing 800 to 1000 reviews will perform better.

In this paper [23] using a combination of natural language processing and machine learning approach, sentiment analysis of movie reviews was proposed. Firstly, the dataset data pre-processing was done. Secondly, to obtain the results for sentiment analysis different feature selection schemes in the combination with two classifiers, Naïve Bayes and SVM were used. Thirdly, for obtaining the results for higher order n-grams, the model of sentiment analysis was extended. As compare to Naïve Bayes, the linear SVM classifier gives more accuracy shown in the classification of a movie review.

In this research [24] using emoticons for training data to perform distant supervised learning is an effective way. High accuracy can be achieved by using different machine learning algorithms in this approach. Tweet sentiment classified by machine learning algorithms with the same performance because twitter messages have unique characteristics.

Previous study [18] perform sentiment analysis on a movie review, a new approach was used that is called a Combined Approach. Two separate classifiers such as Support Vector machine and Hidden Markov

Model (HMM) were combined in this approach and the results of these classifiers also combined. By using the combination of these classifiers there is a possibility to improve the results of classification. slang words and smiley handled by the classifier. With higher accuracy, a good sentiment classification achieved in this approach.

In a study [25], Random Forests Classifier by changing the values of different hyperparameters supervised learning technique called Random Forests for classification results. The comparison of some supervised learning technique like BN, C4.5, and ID3 with Random Forests Classifiers was focused in this paper. And this comparison is focused with respect to ROC Area and incorrectly classified instances. In terms of incorrectly and correctly classified instances and ROC Area, Random Forests outperforms all the three classifiers.

This research [20] investigates parametrization rule for the Forest-RI algorithm should be considered in order to focus on the radiance that is an important property. Influence of Hyperparameter on Random Forest Accuracy is presented in this paper.

By tuning of Hyperparameters in Random Forest, to perform sentiment analysis of movie review dataset on using the Random Forest was focused on the paper [26]. On movie review dataset Random Forest performed well on the basis of experimental results. The result in a high accuracy of 87.85% was of dataset V1.0 and the result with an accuracy of 91.0% was provided by the dataset V2.0. For the sentiment classification, most of the past work has focused on Maximum Entropy, Naïve Bayes and Support Vector Machine. But in this paper, the experiments that are carried out shows that if Hyperparameters are fine-tuned then Random Forest can give better outcomes.

II. METHODS AND MATERIAL

This section describes detailed comparison results of machine learning sentiment analysis.

1) Comparison of Supervised Machine Learning Algorithms:

Comparison of Supervised Machine Learning Algorithms to perform sentiment analysis of movie reviews.

Dataset source = NLTK Movie review corpora

We have NLTK movie review corpora dataset for different machine learning algorithms and extract binary classification results. We have chosen 4 different machine learning algorithms for this purpose. Following are the selected algorithms.

1) Naïve Bayes
2) Support Vector Machine
3) Decision Tree
4) Neural Networks

We have used the same dataset (NLTK Movie review corpora) for all four Machine learning algorithms and generate a comparison chart on the base of extracted results. The comparison was designed on the base of the following three parameters.

1) Precision
2) Recall
3) F1-Score

In order to extract precision, recall, and f1-score confusion matrix was created. Moreover, we have also determined the accuracy of all 4 algorithms on the same training and testing dataset.

Confusion Matrix

	Correct	Not Correct
Selected	Tp	Fp
Not Selected	Fn	Tn

Confusion matrix was designed in order to summarize the predicted results by machine learning algorithms. Confusion matrix classifies result in 4 different classes' True positive, false positive, false negative and True Negative.

Precision:

Precision is the fraction in order to analyse either the retrieved documents are related to the selected query or not. Following is the formula used in order to extract precision by using the confusion matrix.

$presision = \frac{tp}{tp+fp}$ ………… (Eq.1.1)

Recall:

The recall is the fraction of relevant results that are extracted. Following is the formula to extract recall using confusion matrix.

$recall = \frac{tp}{tp+fn}$ ……………… (Eq. 1.2)

F1- Score:

F1-Score is the harmonic mean of precision and recall. Following is the formula to extract f1-score using precision and recall.

$F = \frac{2PR}{(P+R)}$ …………….. (Eq. 1.3)

III. RESULTS AND DISCUSSION

Following are the scores for those factors that are obtained for all 4 selected algorithms.

Naïve Bayes:

Confusion Matrix:

Table 1.1 : Naïve Bayes Confusion Matrix

176	30
41	153

Classification Report:

Table1.2: Naïve Bayes Classification Report

	Precision	recall	f1-score	Support
Positive	85	81	83	217
Negative	79	84	81	183
avg / total	82	82	82	400

Following is the Accuracy Score for Naïve Bayes algorithms:

Accuracy Score = 82.25 %

Support Vector Machine:

Confusion Matrix:

Table 1.3: Support Vector Machine Confusion Matrix

166	40
25	169

Classification Report:

Table 1.4: Support Vector Machine Classification Report

	Precision	recall	f1-score	Support
Positive	81	87	84	192
Negative	87	82	84	209
avg / total	84	84	84	400

Following is the Accuracy Score of support vector machine:

Accuracy Score = 83.75 %

Decision Tree:

Confusion Matrix:

Table 1.5: Decision Tree Confusion Matrix

135	71
68	126

Classification Report:

Table 1.6: Decision Tree Classification Report

	Precision	recall	f1-score	Support
Positive	67	66	67	209
Negative	63	64	64	191
avg / total	65	65	65	400

Following is the Accuracy Score for Decision Tree:

Accuracy Score = 65.25%

Neural Network:

Confusion Matrix:

Table 1.7: Neural Network Confusion Matrix

165	41
28	168

Classification Report:

Table 1.8: Neural Network Classification Report

	Precision	recall	f1-score	Support
Positive	80	85	83	193
Negative	86	80	83	207
avg / total	83	83	83	400

Following is the Accuracy Score Neural Network.

Accuracy Score = 82.75

Comparison:

The given table 1.9 shows performance comparison of 4 key algorithms. In this research doing this comparison of algorithms for finding the best suitability binary classification machine learning methods to predict movie success ratings.

Table 1.9: Comparison of Algorithms

	Naïve Bayes			Support Vector Machine			Decision Tree			Neural Network		
	P	R	F	P	R	F1	P	R	F1	P	R	F1
Positive	85	81	83	81	87	84	67	66	67	80	85	83
Negative	79	84	81	87	81	84	63	64	64	86	80	83
avg / total	82	82	82	84	84	84	65	65	65	83	83	83

Comparison Graph:

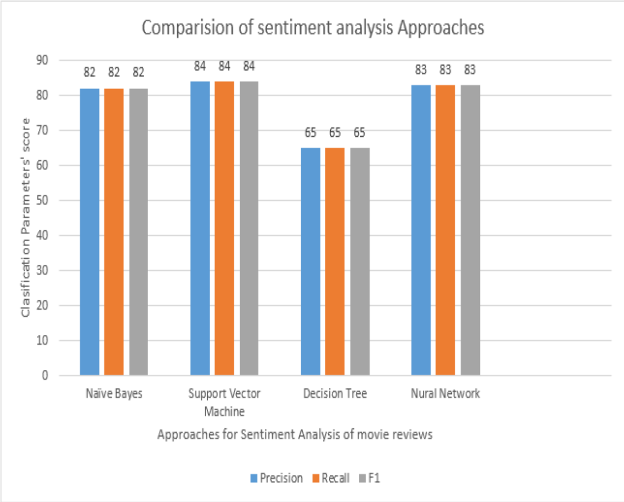

Figure 1: Comparison of Algorithms

Accuracy:

According to the accuracy in results our study finds that SVM provides the high accuracy as compared to other machine learning algorithms as shown in following table 1.10.

Table 1.10: Accuracy of Algorithms

Algorithms	Accuracy
1. Naïve Bayes	82.25%
2. Support Vector Machine	83.75%
3. Decision Tree	65.25%
4. Neural Network	82.75%

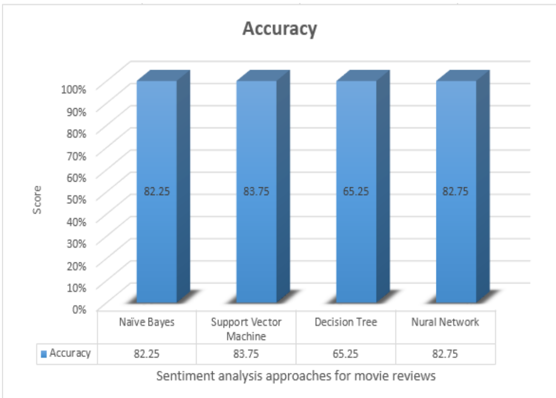

Figure 2 : Accuracy of Algorithms

IV.CONCLUSION AND FUTURE WORK

In this research, we have used the machine learning based approach in order to predict movie success rate. We have used 4 different algorithms (Navie Bayes, Neural Network, Decision Tree, SVM) and find SVM provides better Accuracy in results as compared to the other algorithms. We have analysed challenges we have faced in machine learning approaches like machine learning provide results in a positive and negative format and we were unable to find suitable data set for predicting movie success rate in Star rating format.

In future we can work on creating a suitable dataset for movie star rating for the machine learning method. In future, we can work on the classification of movie reviews using machine learning techniques for multi-class sentiment analysis. We can extract movie reviews from other social media applications like Facebook and IMDB.

V. REFERENCES

[1]. Banovic, N., et al. Modeling and understanding human routine behavior. in Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. 2016.

[2]. Feldman, R., Techniques and applications for sentiment analysis. Communications of the ACM, 2013. 56(4): p. 82-89.

[3]. Di Fabbrizio, G., A. Aker, and R. Gaizauskas. Starlet: multi-document summarization of service and product reviews with balanced rating distributions. in 2011 IEEE 11th International Conference on Data Mining Workshops. 2011. IEEE.

[4]. Asad, K.I., T. Ahmed, and M.S. Rahman. Movie popularity classification based on inherent movie attributes using C4. 5, PART and correlation coefficient. in 2012 International Conference on Informatics, Electronics & Vision (ICIEV). 2012. IEEE.

[5]. Pak, A. and P. Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. in Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). 2010.

[6]. Grimes, S., Text Analytics 2014: User Perspectives on. 2014.

[7]. Asur, S. and B.A. Huberman. Predicting the future with social media. in 2010 IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology. 2010. IEEE.

[8]. Tang, H., S. Tan, and X. Cheng, A survey on sentiment detection of reviews. Expert Systems with Applications, 2009. 36(7): p. 10760-10773.

[9]. Daiyan, M., et al., A literature review on opinion mining and sentiment analysis. International Journal of Emerging Technology and Advanced Engineering, 2015. 5(4): p. 262-280.

[10]. Turney, P.D., Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. arXiv preprint cs/0212032, 2002.

[11]. Kim, S.-M. and E. Hovy. Automatic identification of pro and con reasons in online reviews. in Proceedings of the COLING/ACL 2006 main conference poster sessions. 2006.

[12]. Popescu, A.-M. and O. Etzioni, Extracting product features and opinions from reviews, in Natural language processing and text mining. 2007, Springer. p. 9-28.

[13]. Kobayashi, N., et al. Collecting evaluative expressions for opinion extraction. in International Conference on Natural Language Processing. 2004. Springer.

[14]. Lerman, K., S. Blair-Goldensohn, and R. McDonald, Sentiment summarization: evaluating and learning user preferences. 2009.

[15]. Serrano-Guerrero, J., et al., Sentiment analysis: A review and comparative analysis of web services. Information Sciences, 2015. 311: p. 18-38.

[16]. Go, A., R. Bhayani, and L. Huang, Twitter sentiment classification using distant supervision. CS224N project report, Stanford, 2009. 1(12): p. 2009.

[17]. Canneyt, S.V., N. Claeys, and B. Dhoedt. Topic-dependent sentiment classification on twitter. in European Conference on Information Retrieval. 2015. Springer.

[18]. Dong, L., et al. Adaptive recursive neural network for target-dependent twitter sentiment classification. in Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 2: Short papers). 2014.

[19]. Kaya, M., G. Fidan, and I.H. Toroslu, Transfer learning using Twitter data for improving sentiment classification of Turkish political news, in Information sciences and systems 2013. 2013, Springer. p. 139-148.

[20]. Fernández, J., et al., Sentiment analysis of spanish tweets using a ranking algorithm and skipgrams. 2013.

[21]. Baccianella, S., A. Esuli, and F. Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. in Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). 2010.

[22]. Bradley, M.M. and P.J. Lang, Affective norms for English words (ANEW): Instruction manual and affective ratings. 1999, Technical report C-1, the center for research in psychophysiology

[23]. Choi, Y., et al. Identifying sources of opinions with conditional random fields and extraction patterns. in Proceedings of human language technology conference and conference on empirical methods in natural language processing. 2005.

[24]. Dave, K., S. Lawrence, and D.M. Pennock. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. in Proceedings of the 12th international conference on World Wide Web. 2003.

[25]. Eguchi, K. and V. Lavrenko. Sentiment retrieval using generative models. in Proceedings of the 2006 conference on empirical methods in natural language processing. 2006.

[26]. Fernandez-Lozano, C., et al., Improving enzyme regulatory protein classification by means of SVM-RFE feature selection. Molecular BioSystems, 2014. 10(5): p. 1063-1071.

Cite this article as :

Muhammad Luqman, Amir Yaqoob, Majid Bashir Ahmad, Kanza Majid, "Sentiment Analysis to Predict Movies Success Rate Based on NLTK Movie Review Corpora Using Machine Learning", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 9 Issue 1, pp. 46-54, January-February 2023. Available at doi : https://doi.org/10.32628/CSEIT239013

Journal URL : https://ijsrcseit.com/CSEIT239013