ISSN 2395-1621

SENTIMENT ANALYSIS OF MOVIE REVIEW AND RECOMMENDATION SYSTEM

Shweta Bhosale , Mayuri Borkar, Pushpa Dendage,
Prof. Dr Sushen R. Gulhane

b.shweta1901@gmail.com
mayuriborkar8194@gmail.com
pudendage1999@gmail.com
sushen.gulhane@dyptc.edu.in

Department of Information Technology
D Y Patil College of Engineering, Ambi, Talegaon.

ABSTRACT

Sentiment Analysis or Opinion mining on movie reviews. Sentiment is a thought, view, or attitude, especially one based mainly on emotion instead of reason. For this we need to classify each and every review, this can be done through classifiers such as NAÏVE BAYES, SVM, K-NN, we preferred NAÏVE BAYES classifier, since it is a probabilistic approach, we depend upon each word's probability to be positive and negative, through this we will classify the review to be one of those. Also displaying the details of movie like overview, genera, status, runtime of movie. And all these details we are fetching from IMDB site using IMDB API Key. After performing sentiment analysis on movie reviews, we also implemented Recommendation system for movies.

Keywords: Sentiment Analysis, SVM, K-NN, NAÏVE BAYES.

ARTICLE INFO

Article History

Received: 12th May 2022

Received in revised form : 12th May 2022

Accepted : 14th May 2022

Published online : 15th May 2022

I. INTRODUCTION

Movie reviews are an important way to evaluating the performance of a movies. While providing a numerical/stars rating to a movie tells us about the success or failure of a movie quantitatively, a collection of movie reviews is what gives us a deeper qualitative insight on different aspects of the movie. A textual movie review tells us about a movie's strong and weak points and a thorough analysis of a movie review can tell us whether the movie is in general terms. Sentiment analysis is a major topic in machine learning aimed at extracting subjective information from the textual reviews. The field of sentiment of analysis is closely tied to natural language processing and text mining. It can be used to determine the attitude of the reviewer with respect to various to picsor the overall polarity of review. Using sentiment analysis, we can find the reviewer's state of mind when providing a review and understand whether the person was "happy", "sad", "angry", etc. In this project we intend to use sentiment analysis on the set of movie reviews given by the reviewer's and to understand their overall reaction to the movie , whether they like the movie or not. We aim to use the relation of words in the review to predict the overall polarity of the review.

PROBLEM STATEMENT

Text can be categorized in two types based on its properties in terms of text mining: 'subjectivity 'and 'polarity'. The focus of our project is to find the polarity of the text which means that we are interested in finding if the sentence is positive or negative. We use machine learning techniques classify such sentences and try to find answers to the following questions:

1. What machine learning techniques are useful for this purpose? Which one out of them performs the best and which techniques are better than the others?
2. What are some of the advantages and disadvantages of traditional machine learning techniques for sentiment analysis?
3. How difficult the task of extracting sentiment from short comments or sentences can be as compared to the traditional topic-based text classification?

MOTIVATION

The Emergency in the last Decade of Social Media platforms such as Twitter, Facebook and Instagram enabled people to engage in social activities to express their opinions, thoughts, and emotions on the variety of topics. Open-source platform, large amount of data are produced (e.g.6000tweetspersecond) , this representing an opportunity for companies to assess their social influence and people opinions towards their products. Consequently, a computational framework is desirable to perform opinion mining and sentiment analysis which can adapt to the activity domain of the user. And the recommendation system to help the user to see the related contents from the lots of available contents .

II. LITERATURE SURVEY

Joscha et. al, in their paper [1] devised and compared various techniques like Bag of words models, n-grams for using semantic information to improve the performance of sentiment analysis. The earlier approaches did not consider the semantic associations between sentences or documents parts. Research by A. Hogeboom et al. [2] neither compared the methodological variants nor provided a method to merge disclosure units in the most favourable manner. They aimed to improve the sentiment analysis by using Rhetoric Structure Theory (RST) as it gives a hierarchical representation at the document level. They proposed an integration of the grid search and weighting to find out the average scores of sentiment from Rhetoric Structure Theory (RST) tree. They encoded the binary data into the random forest by using feature engineering as it greatly reduced the complexity of original RST tree. They concluded that machine learning raised the balanced accuracy and gives a high F1 score of 71.9%.

Amir Hossein Yazdavar et al. in this paper [3] provided novel understanding of sentiment analysis problem containing numerated data in drug reviews. They analysed sentences which contained quantitative terms to classify them into opinionated or non-opinionated and also to identify the polarity expressed by using fuzzy set theory. The development of fuzzy knowledge base was done by interviewing several doctors from various medical centres. Although the number of researches has been done in this field (Bhatia, et al., [4]) these do not consider the numerical (quantitative) data contained in the reviews while recognizing the sentiment polarity. Also, the training data used has a high domain dependency and hence cannot be used in different domains. They concluded that their proposed method knowledge engineering based on fuzzy sets was much simpler, efficient and has high accuracy of over 72% F1 value.

Dhiraj Murthy in his paper [5] he identified what roles do tweets play in political elections. He pointed out that even though there were various researches and studies done to find out the political engagement of Twitter, no work was done to find out if these tweets were Predictive or Reactive. In his paper, he concluded that the tweets are more reactive than predictive. He found out that electoral success in not at all related to the success on Twitter and that various social media platforms were used to increase the popularity of a candidate by generating a buzz around them.

Ahmad Kamal in his paper [6] designed an opinion mining framework that facilitates objectivity or subjectivity analysis, feature extraction and review summarization etc. He used supervised machine learning approach for subjectivity and objectivity classification of reviews. The various techniques used by him were Naive Bayes, Decision Tree, Multilayer Perceptron and Bagging. He also improved mining performance by preventing irrelevant extraction and noise as in Kamal's paper[7].in Kamal's paper. [7].

Humera Shaziya et al. in this paper [8] classified movie reviews for sentiment analysis using WEKA Tool. They enhanced the earlier work done in sentiment categorization which analyses opinions which express either positive or negative sentiment. In this paper, they also considered the fact that reviews that have opinions from more than one person and a single review may express both the positive and negative sentiment. They conducted their experiment on WEKA and concluded that Naïve Bayes performs much better than SVM for movie reviews as well as text. Naive Bayes has an accuracy of 85.1%.

Akshay Amolik et. al. in his paper [9] created the dataset using twitter posts of movie reviews and related tweets about those movies. Sentence level sentiment analysis is performed on these tweets. It is done in three phases. Firstly, pre-processing is done. Then Feature vector is created using relevant features. Finally, by using different classifiers like Naïve Bayes, Support vector machine, Ensemble classifier, k-means and Artificial Neural Networks, tweets were classified into positive, negative and neutral classes. The results show that we get 75 % accuracy form SVM. He negated Wu et. al. paper [10] which made an observation that if @username is found in a tweet, it influences an action and also helps to influence the probability. But in this paper Akshay Amolik replaced @username with AT_USER and hashtags were also removed due to which we used Support Vector Machine rather than Naive Bayes which increased the accuracy by 10%.

III. METHOLODGY

1. Dataset

The dataset used for this task was collected from Large Movie Review Dataset which was used by the AI department of Stanford University for the associated publication. The dataset contains10,000 training examples collected from IMDb where each review is labelled with the rating of the movie on scale of 1-10. As sentiments are usually bipolar like good/bad or happy/sad or like/dislike, we categorized these ratings as either 1 (like) or 0 (dislike)based on the ratings. If the rating was above 5, we deduced that the person liked the movie otherwise he did not. Initially the dataset was divided into two subsets containing 25,000 examples each for training and testing. We found this division to be sub-optimal as the number of training examples was very small and leading to under-fitting. We then tried to redistribute the examples as 40,000 for training and 10,000 for testing. While this produced better models, it also led to over-fitting on training examples

and worse performance on the test set. Finally, we decided to use Cross-Validation in which the complete dataset is divided into multiple folds with different samples for training and validation each time and the final performance statistic of the classifiers averaged over all results. This improved the accuracy of our models across the boards. A typical review text looks like this:

I'm a fan of TV movies in general and this was one of the good ones. The cast performances throughout were pretty solid and there were twists I didn't see coming before each commercial. To me it was kind of like Medium meets CSI.

Did anyone else think that in certain lights, the daughter looked like a young Nicole Kidman? Are they related in any way? I'd definitely watch it gain or rent it if it ever comes to video.

Dedee was great. Haven't seen in her in a lot of things and she did her job very convincingly.

If you're into to TV mystery movies, check this one out if you have a chance.

As seen above, one necessary pre-processing step prior to feature extraction was removal of HTML tags like "
". We used simple regular expressions matching to remove these HTML tags from the text.

```
data=data.loc[:,['director_name','actor_1_name','
actor_2_name','actor_3_name','genres','movie_title']]
```

Another important step was to make the text case-insensitive as that would help us count the word occurrences across all reviews and pronoun important words. We also removed all the punctuation marks like '!', '?', etc. as they do not provide any substantial information and are used by different people with varying connotations. This was achieved using standard python libraries for text and string manipulation. We also removed stop words from the text for some of our feature extraction tasks, which is described in greater detail in later sections. One important point to note is that we did not use stemming of words as some information is lost while stemming a word to its root form.

2. Text Pre-processing

After collecting the required datasets, the next step is to clean the data. It is very important step.

Following is the code that we done for cleaning our data.

In any machine learning task, cleaning or pre-processing the data is as important as model building if not more. And when it comes to unstructured data like text, this process is even more important.

Objective of this kernel is to understand the various text pre-processing steps with code examples.

Some of the common text pre-processing / cleaning steps are:

- Lower casing
- Removal of Punctuations
- Removal of Stop words
- Removal of Frequent words
- Removal of emoji

So, these are the different types of text pre-processing steps which we can do on text data. But we need not do all of these all the times. We need to carefully choose the pre-processing steps based on our use case since that also play an important role.

For example, in sentiment analysis use case, we need not remove the emoji or emoticons as it will convey some important information about the sentiment. Similarly, we need to decide based on our use cases.

```
Import pandas as pd
Data=pd.read_csv('datasets/movies_metadata_00.csv')
Data.head()
```

	color	director_name	num_critic_for_reviews	duration	director_facebook_likes	actor_3_facebook_likes	actor_2_name	actor_1_facebook_likes
0	Color	James Cameron	723.0	178.0	0.0	855.0	Joel David Moore	1000.0
1	Color	Gore Verbinski	302.0	169.0	563.0	1000.0	Orlando Bloom	40000.0
2	Color	Sam Mendes	602.0	148.0	0.0	161.0	Rory Kinnear	11000.0
3	Color	Christopher Nolan	813.0	164.0	22000.0	23000.0	Christian Bale	27000.0
4	NaN	Doug Walker	NaN	NaN	131.0	NaN	Rob Walker	131.0

5 rows × 28 columns

Recommendation will be based on these features only

- Lower Casing

Lower casing is a common text pre-processing technique. The idea is to convert the input text into same casing format so that 'text', 'Text' and 'TEXT' are treated the same way.

```
data['movie_title'] = data['movie_title'].str.lower()
```

- Removal of Punctuations

One another common text pre-processing technique is to remove the punctuations from the text data. This is again a text standardization process that will help to treat 'hurray' and 'hurray!' in the same way.

```
data['genres'] = data['genres'].str.replace('|', ' ') data.head()
```

director_name	actor_1_name	actor_2_name	actor_3_name	genres	movie_title
James Cameron	CCH Pounder	Joel David Moore	Wes Studi	Action Adventure Fantasy Sci-Fi	Avatar
Gore Verbinski	Johnny Depp	Orlando Bloom	Jack Davenport	Action Adventure Fantasy	Pirates of the Caribbean: At World's End
Sam Mendes	Christoph Waltz	Rory Kinnear	Stephanie Sigman	Action Adventure Thriller	Spectre
Christopher Nolan	Tom Hardy	Christian Bale	Joseph Gordon-Levitt	Action Thriller	The Dark Knight Rises
Doug Walker	Doug Walker	Rob Walker	unknown	Documentary	Star Wars: Episode VII - The Force Awakens ...

1. Applying Naive Bayes

Naive Bayes works on the principle of probabilities and the Bayes rule given by:

$$P(c|d) = P(c)P(d|c) / P(d)$$

By using Naive Baye's machine learning classifier, we are performing sentiment analysis on the reviews and divide it into two types:

- Positive
- Negative


```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.20,
random_state=42) clf = naive_bayes.MultinomialNB()
clf.fit(X_train,y_train)
```

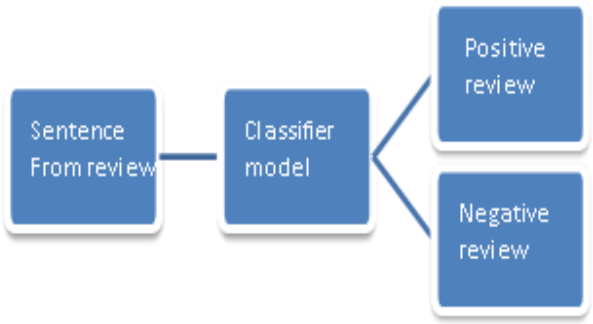

Fig: Sentiment analysis task

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.20, random_state=42)

clf = naive_bayes.MultinomialNB()
clf.fit(X_train,y_train)

MultinomialNB()

accuracy_score(y_test,clf.predict(X_test))*100

97.47109826589595

clf = naive_bayes.MultinomialNB()
clf.fit(X,y)

MultinomialNB()
```

2. Designing Graphical User Interfaces

As one of the main objective of the project is performing highly accurate sentiment analysis, the graphical user interface was not within the initial scope. However, a graphical user interface turned out to be imperative, such that users can interact with the system in a more intuitive manner.
We have used various front end designing technologies to give a better experience to user.

These technologies mentioned below:

1. HTML

HTML provides the basic structure of sites, which is enhanced and modified by other technologies like CSS and JavaScript.HTML stands for Hyper Text Markup Language. "Markup language" means that, rather than using a programming language to perform functions, HTML uses tags to identify different types of content and the purposes they each serve to the webpage

2. CSS

CSS is used to control presentation, formatting, and layout.. CSS is the language we use to style an HTML document.CSS describes how HTML elements should be displayed.

3. JavaScript

JavaScript is used to control the behaviour of different elements. JavaScript is a cross-platform, object-oriented scripting language used to make webpages interactive (e.g., having complex animations, clickable buttons, popup menus, etc.). There are also more advanced server-side versions of JavaScript such as Node.js, which allow you to add more functionality to a website than downloading files (such as real-time collaboration between multiple computers). Inside a host environment (for example, a web browser), JavaScript can be connected to the objects of its environment to provide programmatic control over them.

The flow of GUI is given below:

1. Firstly we have inserted a background image to make the GUI attractive.
2. Then the heading is given as “sentiment analysis of movie reviews and recommendation system” and make it bold and align to centre.
3. We have taken a text box to enter the movie name and a button.
4. After entering the button, we will get the movie details
5. A Card is used to show movie poster and align it to left.
6. We make use of table to show the reviews and its analysed sentiment.
7. Then again using cards to show the recommended movies to the user

IV. CONCLUSION

In this project we have investigated the task of sentiment analysis as a classification problem. We have used datasets: IMDB movie reviews and first performed data cleaning and the perform the machine learning classification algorithm Naive Baye’s theorem. Also created the bag of words for the purpose of feature extraction from the given text.
And then after search the movie by its name, all the details regarding the movie were shown and the online reviews of the respective movie has fetched and we perform sentiment analysis on it and labbled whether the review is good or bad And also shows the top Cast from the film and created recommendation system based on the actor name, director name, title of the movie etc.

V. REFERENCES

[1]Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. (2011). "Learning Word Vectors for Sentiment Analysis. “The 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011).

[2]Kennedy and D. Inkpen. 2006. Sentiment classification of movie reviews using contextual valence shifters. Computational Intelligence, 22:110–125, May

[3]O. Alm, D. Roth, and R. Sproat. 2005. Emotions from a text: machine learning for text-based emotion prediction. In Proceedings of HLT/EMNLP, pages 579–586

[4]M. Steyvers and T. L. Griffiths. 2006. Probabilistic topic models. In T. Landauer, D McNamara, S. Dennis, and W. Kintsch, editors, Latent Semantic Analysis: A Road to Meaning.

[5]Pang, Bo; Lee, Lillian; Vaithyanathan, Shivakumar (2002). "Thumbs up? Sentiment Classification using Machine Learning Techniques". Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP).

[6]Pang, Bo; Lee, Lillian; Vaithyanathan, Shivakumar (2002). "Thumbs up? Sentiment Classification using Machine Learning Techniques". Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP).

[7]Tumasjan, Andranik; O.Sprenger, Timm; G.Sandner, Philipp; M.Welpe, Isabell (2010). "Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment". "Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media

[8]Natural Language Processing from Scratch http://static.googleusercontent.com/media/research.google.com/en//pubs/archive/35671.pd

[9]Scikit-learn API Reference: http://scikit learn.org/stable/modules/classes.htmAvenues in Opinion Mining and Sentiment Analysis". IEEE Intelligent System

