MOVIE REVIEW SENTIMENT ANALYSIS

COURSE PROJECT REPORT
18CSE398J -Machine Learning - Core Concepts with Applications

(2018 Regulation)
III Year/ VI Semester
Academic Year: 2022 -2023 (EVEN)

By

Aastha Aggarwal RA2011026010297

Under the guidance of

Dr. Vadivu G
Professor
Department of Data Science and Business Systems

SRM
INSTITUTE OF SCIENCE & TECHNOLOGY
(Deemed to be University u/s 3 of UGC Act, 1956)

DEPARTMENT OF DATA SCIENCE AND BUSINESS SYSTEMS
FACULTY OF ENGINEERING AND TECHNOLOGY
SRM INSTITUTE OF SCIENCE AND TECHNOLOGY

Kattankulathur, Kancheepuram

MAY 2023

ABSTRACT

Nowadays, social media has become a tremendous source of acquiring user's opinions. With the advancement of technology and sophistication of the internet, a huge amount of data is generated from various sources like social blogs, websites, etc. In recent times, blogs and websites are the real-time means of gathering product reviews. However, an excessive number of blogs on the cloud has enabled the generation of a huge volume of information in different forms like attitudes, opinions, and reviews. All of the collected data is used to improve products and services provided by both private organizations and governments around the world. This project includes sentiment analysis of movie reviews using feature-based opinion mining and supervised machine learning. In this project, the main focus is to determine the polarity of reviews using nouns, verbs, and adjectives as opinion words. Reviews will be Classified into two different categories positive and negative. Reviews of Open Movie Database is used as a source data set and Natural Language Processing Toolkit for Part of Speech Tagging.

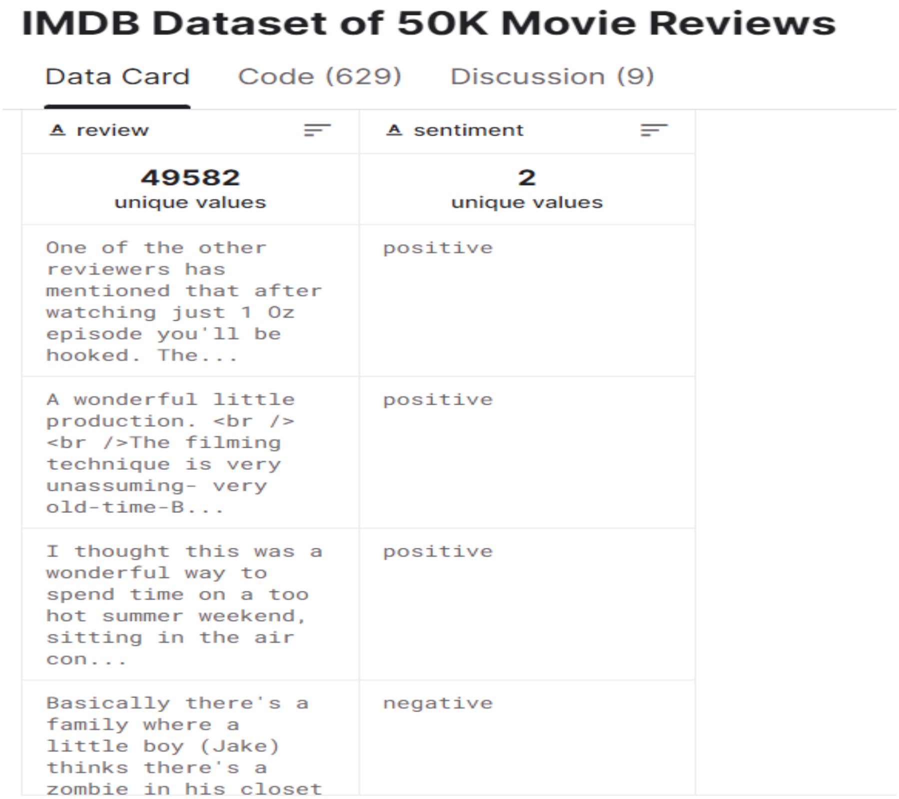
Github Link: Movie Review Sentiment Analysis

INTRODUCTION

Every Human Being makes its decisions based on past experience, sentiments or opinions passed by other human beings. Whenever an individual wants to buy a new item or product, they seek opinions from others about the item or product. Similarly, every organization wants to deliver their best product to the market so they gather opinions from their customers about their product using surveys. Sentiment Analysis is a study of someone's opinions, sentiments or emotions expressed about a product or a movie. Movie reviews help users decide if the movie is worth their time. A summary of all reviews for a movie can help users make this decision by not wasting their time reading all reviews. Movie-rating websites are often used by critics to post comments and rate movies which help viewers decide if the movie is worth watching. Sentiment analysis can determine the attitude of critics depending on their reviews. Sentiment analysis of a movie review can rate how positive or negative a movie review is and hence the overall rating for a movie. Therefore, the process of understanding if a review is positive or negative can be automated as the machine learns through training and testing the data. This project aims to rate reviews using two classifiers and compare which gives better and more accurate results. Classification is a data mining methodology that assigns classes to a collection of data in order to help in more accurate predictions and analysis. Logistic Regression,Naïve Bayes and decision tree classifications will be used and the results of sentiment analysis compared.

DATASET

IMDB dataset is a dataset having about 50K movie reviews for natural language processing or Text analytics. This is a dataset for binary sentiment classification containing substantially more data than previous benchmark datasets. This dataset provides a set of 25,000 highly polar movie reviews for training and 25,000 for testing. So, one can predict the number of positive and negative reviews using either classification or deep learning algorithms.



IMDB Dataset of 50K Movie Reviews

review	sentiment
49582 unique values	2 unique values
One of the other reviewers has mentioned that after watching just 1 Oz episode you'll be hooked. The...	positive
A wonderful little production. <br /><br />The filming technique is very unassuming- very old-time-B...	positive
I thought this was a wonderful way to spend time on a too hot summer weekend, sitting in the air con...	positive
Basically there's a family where a little boy (Jake) thinks there's a zombie in his closet	negative

Dataset Link:
https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews

METHODS

- Collect customer feedback data: You need to gather data from various sources such as customer reviews, social media platforms, and customer service records. This data can be in the form of text, audio or video.
- Data preprocessing: After collecting the data, you need to clean and preprocess it. This includes removing stop words, stemming or lemmatizing words, and converting the text to lowercase.
- Sentiment analysis: Use sentiment analysis techniques to identify the overall sentiment of the customer feedback. This helps to identify the negative and positive feedback.
- Topic modeling: Use topic modeling techniques such as Latent Dirichlet Allocation (LDA) to identify the main topics discussed in the customer feedback. This helps to identify the areas of improvement.
- Word cloud: Use word cloud techniques to visualize the most commonly used words in the customer feedback. This helps to identify the areas of improvement.
- Text classification: Use text classification techniques to categorize the customer feedback into different categories such as product features, customer service, pricing, etc. This helps to identify the specific areas of improvement.
- Data visualization: Use data visualization techniques such as bar charts, pie charts, and heat maps to visualize the results of the analysis. This helps to communicate the findings effectively.

EXPERIMENT AND RESULT

The algorithms used to compare the test accuracies for this sentiment analysis model are:

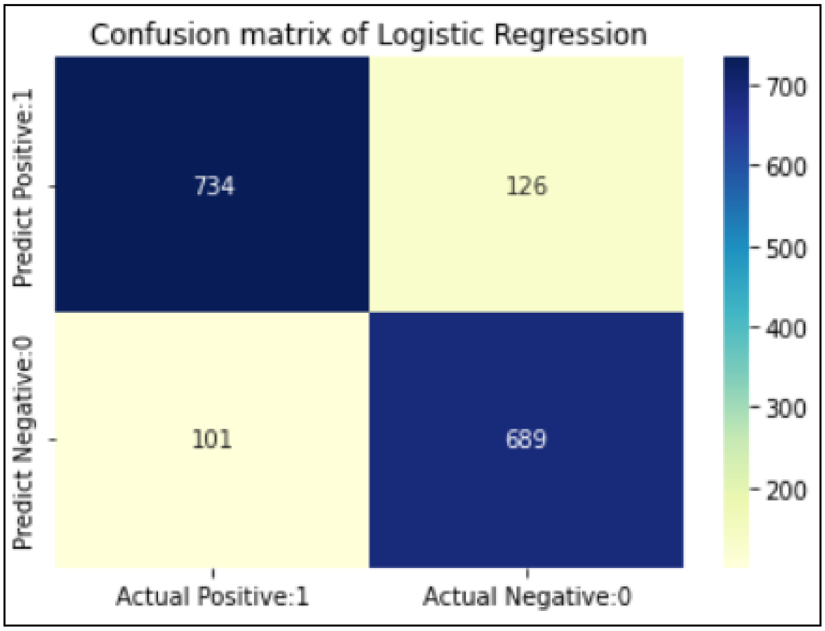
- Logistic Regression: is a statistical analysis method to predict a binary outcome, such as yes or no, based on prior observations of a data set. A logistic regression model predicts a dependent data variable by analyzing the relationship between one or more existing independent variables.
- Multinomial Naive Bayes: is a popular machine learning algorithm for text classification problems in Natural Language Processing (NLP). It is particularly useful for problems that involve text data with discrete features such as word frequency counts. MNB works on the principle of Bayes theorem and assumes that the features are conditionally independent given the class variable.
- Linear Support Vector Classifier: is an algorithm that attempts to find a hyperplane to maximize the distance between classified samples.
- Decision Tree Classifier: is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.

These four algorithms are trained and tested on the training and testing dataset respectively. After which, their test accuracies are calculated for us to conclude that which model performs better.

Out of all four, Logistic Regression has the highest test accuracy and is therefore, the best model for movie review sentiment analysis.

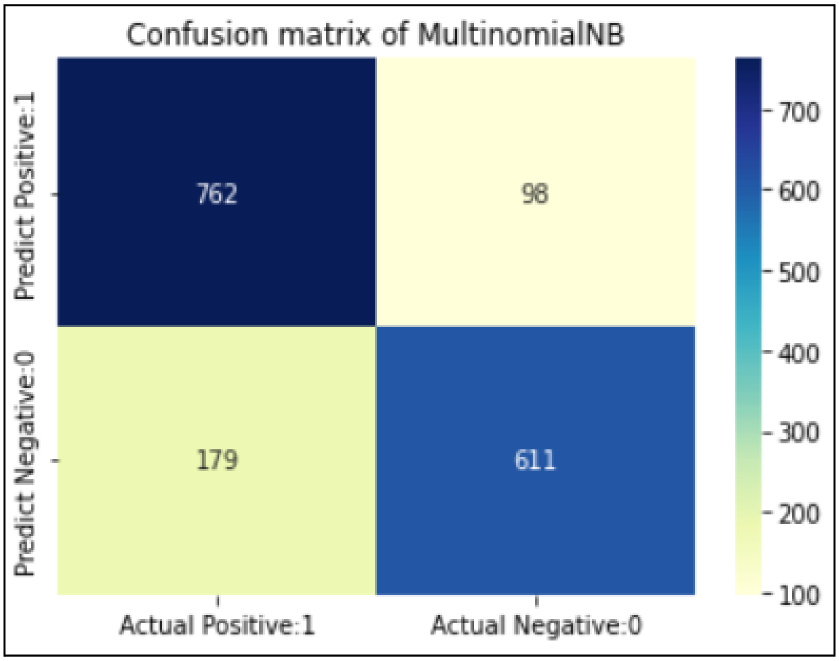
Accuracy and Confusion Matrix for Logistic Regression

```
train score :  0.954002079002079
test score :  0.8624242424242424
Test accuracy: 86.24%
```

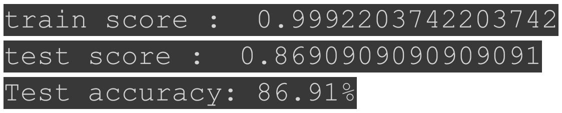


Accuracy and Confusion Matrix for Multinomial NB

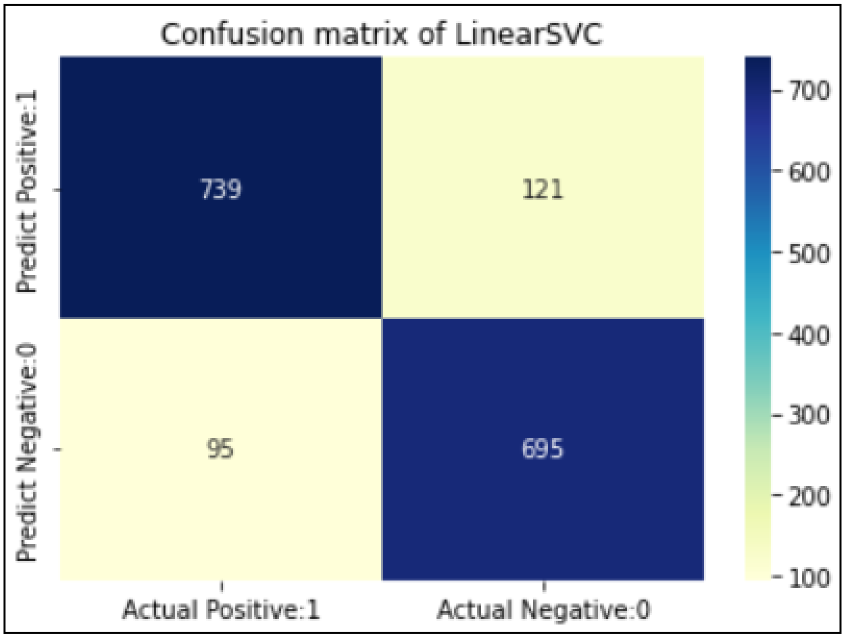
```
train score :  0.9586798336798337
test score :  0.8321212121212122
Test accuracy: 83.21%
```



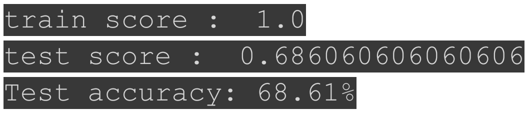
Accuracy and Confusion Matrix for Linear SVC



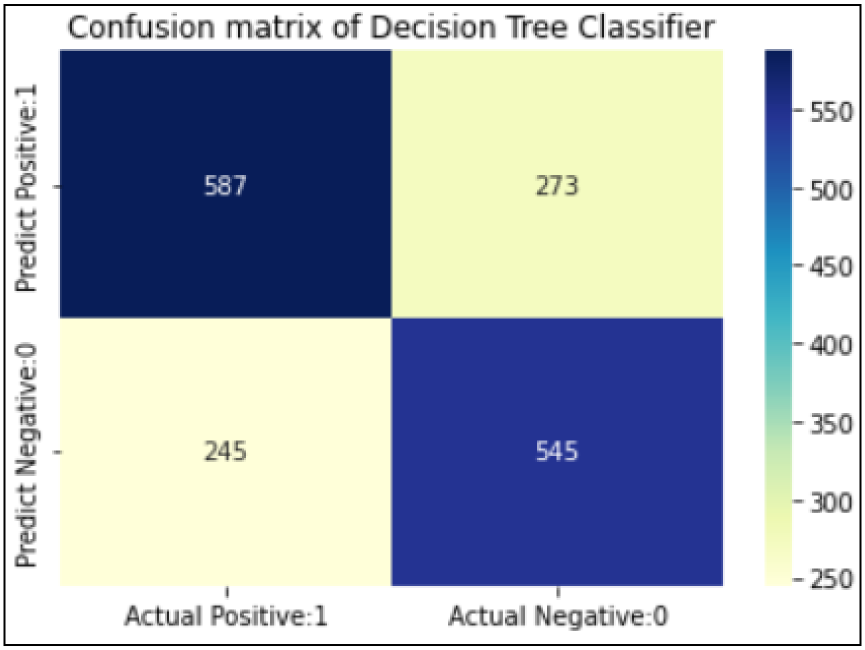
```
train score :  0.9992203742203742
test score :  0.8690909090909091
Test accuracy: 86.91%
```



Accuracy and Confusion Matrix for Decision Tree Classifier



```
train score :  1.0
test score :  0.6860606060606061
Test accuracy: 68.61%
```



CONCLUSION AND FUTURE WORK

From the results above, we can infer that for our problem statement, the Logistic Regression Model is best. One can also use a Naïve Bayes Classifier or an SVC classifier, which provides a good accuracy percentage. One peculiar thing to note is the low accuracy of the Decision Tree classifier. This might be because of the overfitting of decision trees to the training data.

Sentiment analysis of a movie review can rate how positive or negative a movie review is and hence the overall rating for a movie. Therefore, the process of understanding if a review is positive or negative can be automated as the machine learns through training and testing the data. This model can be useful to production houses and movie directors to know the public reviews and this will also give them a better insight into the success or failure of their movies.

This is only the beginning of sentiment analysis. Further methodologies could utilize bigrams (groupings of two words) to endeavor to hold more relevant importance, utilizing neural networks like LSTMs (Long Momentary Memory) to expand the distance of relationships among words in the audits and that's only the tip of the iceberg.

REFERENCES

1. Twitter Sentiment Analysis Gunjan Goyal — Published On June 11, 2021 and Last Modified On March 3rd, 2023https://www.analyticsvidhya.com/blog/2021/06/twitter-sentiment-analysis-a-nlp-use-case-for-beginners/
2. Stemming and Lemmatization Shipra Saxena — Published On March 23, 2021 and Last Modified On March 30th, 2021https://www.analyticsvidhya.com/blog/2021/03/tokenization-and-text-normalization/
3. Sentiment Analysis. (n.d.). Retrieved from https://monkeylearn.com/sentiment-analysis/
4. Natural language toolkit. http://www.nltk.org
5. Scikit-learn toolkit: https://scikit-learn.org/stable/
6. S. Bird, E. Klein, and E. Loper. Natural Language Processing with Python – Analyzing Text with the Natural Language Toolkit. O'Reilly Media, 2009.
7. B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. 2002.
8. Kuat Yessenov and Sasa Misailovic. Sentiment Analysis of Movie Review Comments. May 17, 2009
9. Kamal A., 2015, Review Mining for Feature Based Opinion Summarization and Visualization
10. Humera Shaziya, G.Kavitha, Raniah Zaheer, 2015, Text Categorization of Movie Reviews for Sentiment Analysis , International Journal of Innovative Research in Science, Engineering and Technology, Vol. 4, Issue11
11. Akshay Amolik, Niketan Jivane, Mahavir Bhandari, Dr.M. Venkatesan, Twitter Sentiment Analysis of Movie Reviews using Machine Learning Techniques, School of Computer Science and Engineering, VIT University, Vellore
12. Y. Wu and F. Ren, 2011, Learning Sentimental influence in twitter, Future Computer Science and Application (ICFCSA), 2011, International Conference IEEE vol. 119122.
13. A. Hogenboom, F. Frasincar, F. de Jong, and U. Kaymak,. 2015, Using Rhetorical Structure in Sentiment Analysis, Communications of the ACM, vol. 58, no. 7, pp. 69–77.

14. Palak Baid, Apoorva Gupta and Neelam Chaplot, Sentiment Analysis of Movie Reviews using Machine Learning Techniques, December 2017
15. A Hybrid CNN-LSTM Model for Improving Accuracy of Movie Reviews Sentiment Analysis, Anwar Ur Rehman, Ahmad Kamran Malik, Basit Raza & Waqar Ali, June 2019
16. Al-Smadi M, Talafha B, Al-Ayyoub M, Jararweh Y (2018) Using long short-term memory deep neural networks for aspect-based sentiment analysis of Arabic reviews. International Journal of Machine Learning and Cybernetics, 1-13
17. Cheng Z, Xiaojun C, Lei Z, Rose C, Mohan K (2019) MMALFM: Explainable recommendation by leveraging reviews and images. ACM Trans Inf Syst 37(2):16
18. Elghazaly T, Mahmoud A, Hefny HA (2016) Political sentiment analysis using twitter data. In Proceedings of the ACM International Conference on Internet of things and Cloud Computing, pp.11
19. Fang X, Zhan J (2015) Sentiment analysis using product review data. J Big Data 2(1):5
20. Govindarajan M (2013) Sentiment analysis of movie reviews using hybrid method of naive bayes and genetic algorithm. Inte J Adv Comput Res 3(4):139
21. Hassan A, Mahmood A (2018) Convolutional Recurrent Deep Learning Model for Sentence Classification. IEEE Access 6:13949–13957
22. Himelboim I, Smith MA, Rainie L, Shneiderman B, Espina C (2017) Classifying twitter topic-networks using social network analysis. Social Media+ Society, 1-13