Online Movie Recommendation System Using Sentimental Analysis

Software Construction & Maintenance

Project Report

Submitted by

19MIS1186 – KHUSHI KANTULA

19MIS1084 – AKASH RAVI 19MIS1156 – NIRANJANA O

Semester: Fall 2021-2022

### *Abstract—* The Online Movie Recommendation System provides reviews and ratings to released movie and suggest movie to the user. This system generates a common review related to a movie by using Latent-Semantic Analysis (LSA) algorithm. LSA algorithm analyses the relationships between a set of documents and the terms they contain by producing a set of concepts related to the documents and terms. Here analysis of comments given by various users will be done and a common review will be generated. This generated review will be a simple English statement and will help user to take a correct decision while selecting any movie. Also, the system is providing ratings out of five. With the help of this, user will be able to search his/her favourite movie with in no time. There is a section in our website which also provides a brief information about the movie and genres. If user wants to recommend his/her favourite movie to any other user/friend then he/she can.

***Index Terms—*Movie review, Latent Semantic Analysis, Sentiment analysis, Recommendation system, Opinion mining.**

1. INTRODUCTION

People‟s opinion has become one of the extremely important sources for various services in ever-growing popular social networks. In particular, online opinions have turned into a virtual currency for businesses looking to market their products, identify new opportunities, and manage their reputations [6]. In general, recommender systems are defined as the supporting systems which help users to find information, products, or services (such as books, movies, music, digital products, websites, and TV programs) by aggregating and analysing suggestions from other users, which means reviews from various authorities, and user attributes. After viewing such reviews, they take their decisions. So, such reviews must be correct and proper [3].

Generally, the reviews are generated in graphical format that is in star ratings. Users just have to see the ratings which are generated by analyzing the ratings given by other users to that product and have to take his/her decisions. Such ratings are easily understandable by any user. But they don’t give clear idea of how the product is. They are helpful only in the scenario where if any product is excellent or very poor. The scenario where product is average, star ratings prove bit confuse for any user. They don’t get clear views of what the other users think of that product. If the reviews are in simple English statement it would be easy for any user to understand the feelings of the other users too, about the product. Also, star ratings will be there for his/her help. So, the review about any product will give clear idea to any user so that he can easily take his/her decisions in such confusing scenario too.

Our system is a movie recommendation system which will generate reviews related to the movies which are released. Unlike other systems, we are going to generate a common review by analyzing only the comments of the people (no heavy feedback). This will reduce the overhead of any user who is commenting on any movie and will make the system more user-friendly. So, the system will generate better review which will be a simple English statement. Also, our system will provide star ratings.

1. LITERATURE SURVEY

Sentiment Analysis is an application of Machine Learning (ML) which is used to find the sentiments of users‟ reviews, comments etc. on the internet. Nowadays, social websites like Facebook, Twitter are widely used for posting the users reviews about different things such as movies, news, food, fashion, politics and much more. Reviews and opinions play a major role in identifying the level of satisfaction of users regarding a particular entity. These are then used to find the polarity i.e. positive, negative and neutral. In this paper an approach to Sentiment Analysis on movie reviews in Hindi language is discussed [2].

There are various papers which have used machine learning based approach for sentiment analysis on product reviews [1][4] and it showed better results than lexical based approach [6].

The main problem of this sentiment analysis of customer reviews is that classification of opinions of customers as negative, positive or neutral is a very tedious and hard task to be accomplished. The foremost challenge is the identification of fake reviews [7]. This isn't useful for buyers those need to look through the audits of items preceding buy yet additionally for organizations those need to watch the open's response to their items. In web time, advance in use of online life destinations, for example, twitter, Facebook, and survey webpage delivers an enormous measure of printed data. The printed data fills an imperative source to recognize open/client's assumption towards ideological group, items or an occasion. The slants communicated by open/clients are in the structure of positive, negative or unbiased extremity. The printed data in online life destinations assumes a significant job in choice emotionally supportive networks and individual choice makers. The procedure of

Mechanizing recognizable proof of assessment in a content is known as Sentiment Analysis. The framework can screen and assess continuously online perspectives, to exhibit how the entire world is responding to an idea/belief system/occasion. Growing such a framework which appoints extremity to a tweet is a hard assignment. In this paper [14], a scoring system to discover the slant extremity of the Twitter messages is proposed.

Recommender system is used to recommend items and services to the users and provide recommendations based on prediction. The prediction performance plays vital role in the quality of recommendation.

FUNCTIONAL REQUIREMENTS

The functional requirements of the Sentiment Analysis System for Movie Review are as follows:

1. The system can manage (add, read, update, and delete) movie review datasets

2. The system can manage (add, read, update, and delete) the dictionary of root words

3. The system can manage (add, read, update, and delete) the dictionary of stop words

4. The system can perform the classifier training process and display the model in the form of feature sets of the term data from the training data

5. The system can display the test data result and display confusion matrix generated from the classifier testing

6. The system can display a set of movie review dataset terms derived from tokenizing, filtering, and stemming processes

7. The system can display sentiment analysis result derived from reviews submitted by users

NON – FUNCTIONAL REQUIREMENTS

Meanwhile, the non-functional requirements of the Sentiment Analysis System for Movie Review are as follows:

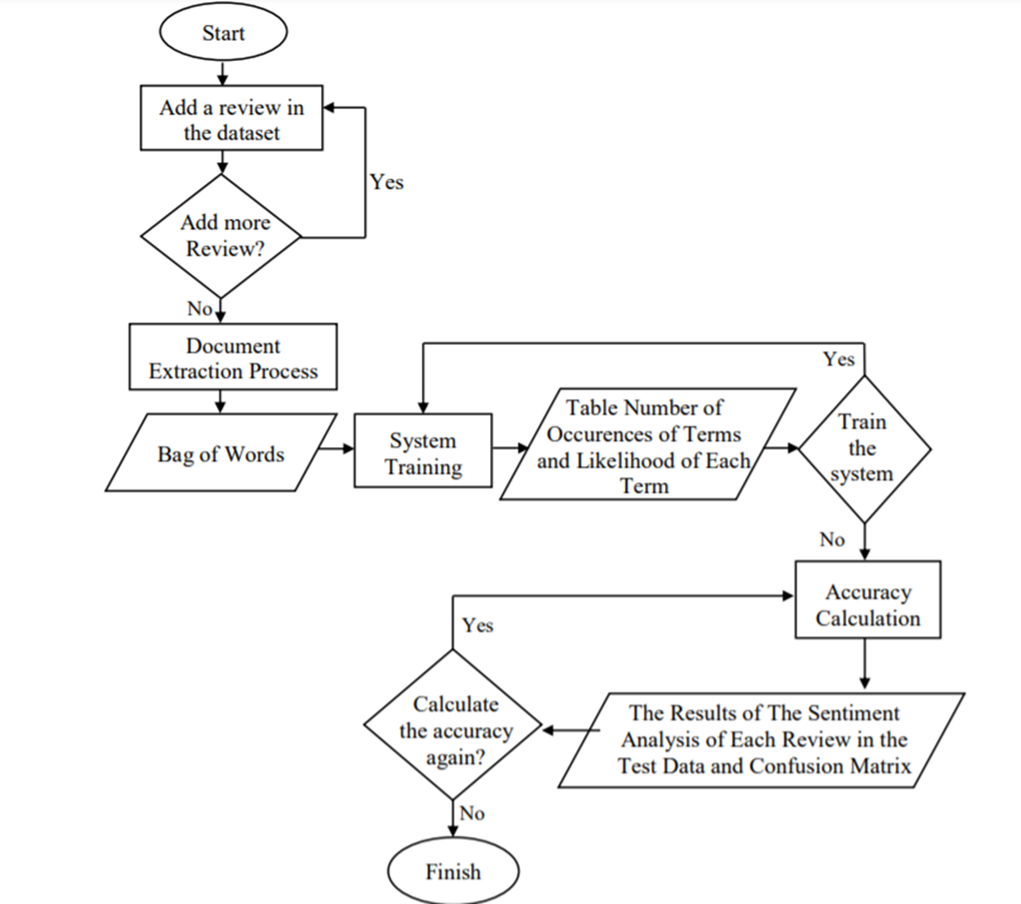
1. The system uses an authentication through login process in order to differentiate user level.

2. The system can run in various web browsers which support the system environment

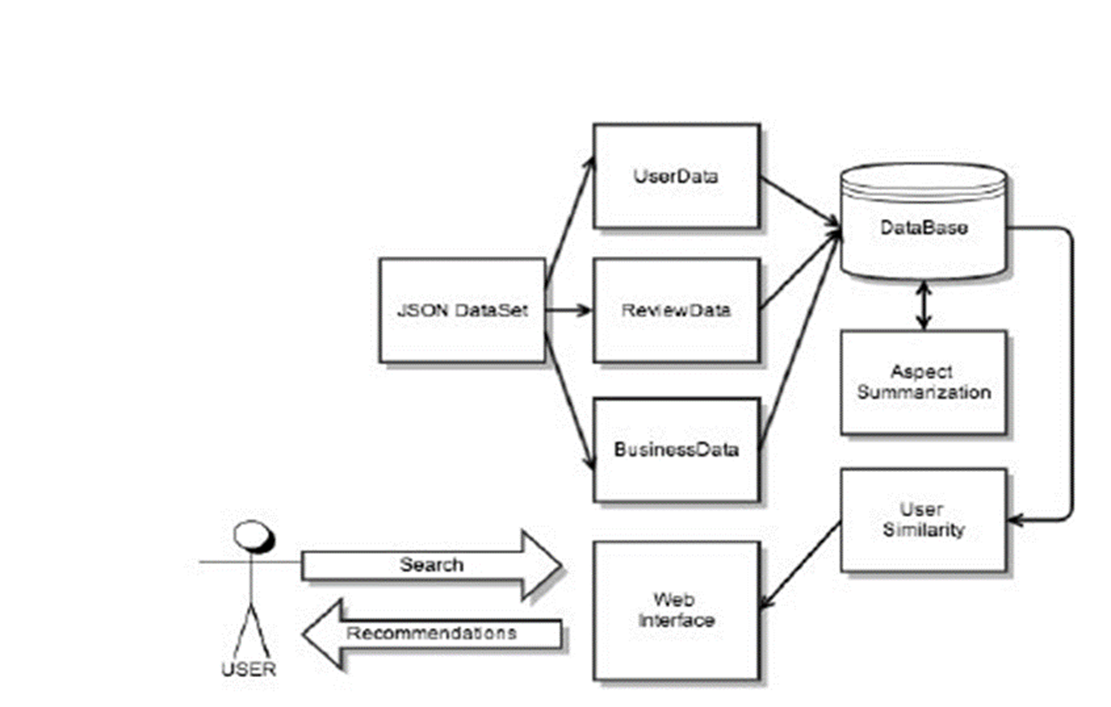
3. The system gives a fast response

4. The system has a user-friendly interface design.

OVER VIEW OF THE SYSTEM



SYSTEM ARCHITECTURE



1. METHODOLOGY

Online Movie Review system is designed to overcome the drawbacks of existing system. It also provides fast searching for any movie and also viewing reviews of that movie and recommending movie to any of the friend by sending him/her a mail. Main Aspects:

* Review Generation
* Recommendation

This system will provide fast searching as we implement Alignment algorithm. This area introduces the proposed system to expand the feature extraction which is helpful in sentiment analysis. The proposed system uses a combination of NLTK (Natural Language Tool Kit) systems for preparing datasets for positive and negative reviews (using python) and supervised learning.

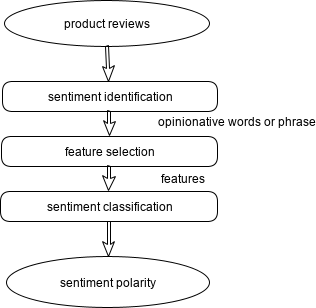


Figure 1. Flow chart for sentiment analysis on product reviews

The above figure shows the architecture of proposed system. In this movie review dataset is used to perform the operations.

Data Preprocessing: The preprocessing is done to expel the noise information from the dataset. In other word, dataset is advanced according to prerequisite. The strategy utilized as a part of preprocessing is feature extraction. The following pre-processing steps are followed to scoring the sentence.

* + *Tokenization* is the process of breaking a sequence of string into pieces. It may be words, keyword, symbols, keyword, sentences and other elements called tokens. Tokens can be a words, phrases or they may be whole sentences. In this process some characters like punctuation mark are removed and these tokens become the input to the other task such as parsing or text mining.
  + *POS (Part of speech)* tagging is the preprocessing technique. In this techniques words are tagged to specific part of speech such as a nouns, adjective, adverbs, verbs etc. based on their association with resulting words. E.g. “This movie is so wasteful of talent, it is truly disgusting” this sentence will be tagged as “This (Determiner) movie (Noun) is (Verb) so (Adverb) wasteful (Adjective) of (Preposition) talent (Noun), it (Pronoun) is (Verb) truly (Adverb) disgusting (Adjective)”. As per relationship of words we will give score for sentence.

After the Data Pre-processing, the bag of words for each positive and negative reviews is extracted from all the reviews received. Now, comes the part of movie review classification into either positive or negative review. For this, an algorithm called LSA (Latent Semantic Analysis) is used. It is also compared with the results of direct correlation of movie ratings results achieved.

For performing LSA, the following methodology is used:

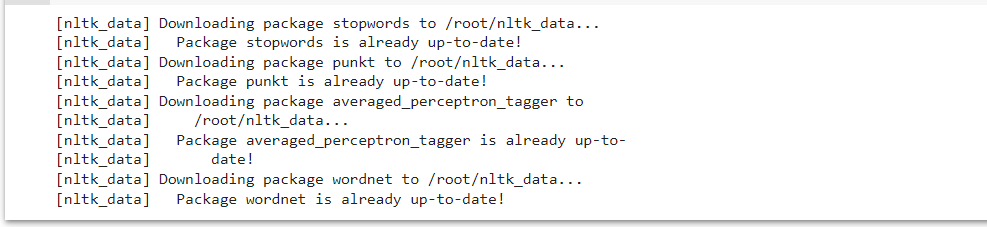
* + A matrix „A‟ is made with rows as bag of words and 2 columns, one of the positive review and the other of negative review.
  + A query column vector is formed for the review to be classified based on the presence of query review words in bag of words. (If present, 1 is added in the corresponding row, else 0).
  + Decomposition of matrix „A‟ into U, S and V matrices where A=USVT. This is done by finding the eigen values of the matrix formed by multiplication of A and transpose of A.
  + Rank approximation is done.
  + A new query vector is formed using q=qTUS-1.
  + Using cosine similarity, the maximum score corresponds to the positive or negative review achieved accordingly.

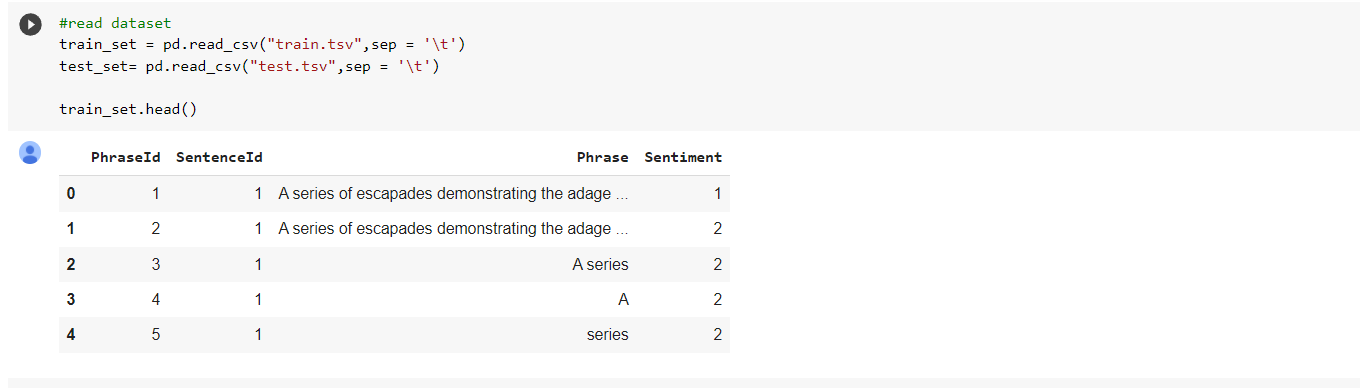
1. EXPERIMENTATIONS AND RESULTS
   * Dataset: In this paper, to perform sentiment analysis on movie review by using some approach for feature extraction using opinion lexicon English is done. For the analysis of sentiment large dataset is used. The dataset contain total 250000 number of movie review. In which 12500 are the positive review and 125000 are the negative review.
   * Technology Used: For performing all this models python is used. Python is widely used high level programming language. Pre- processing is done by using NLTK. It is most popular library for natural language processing.

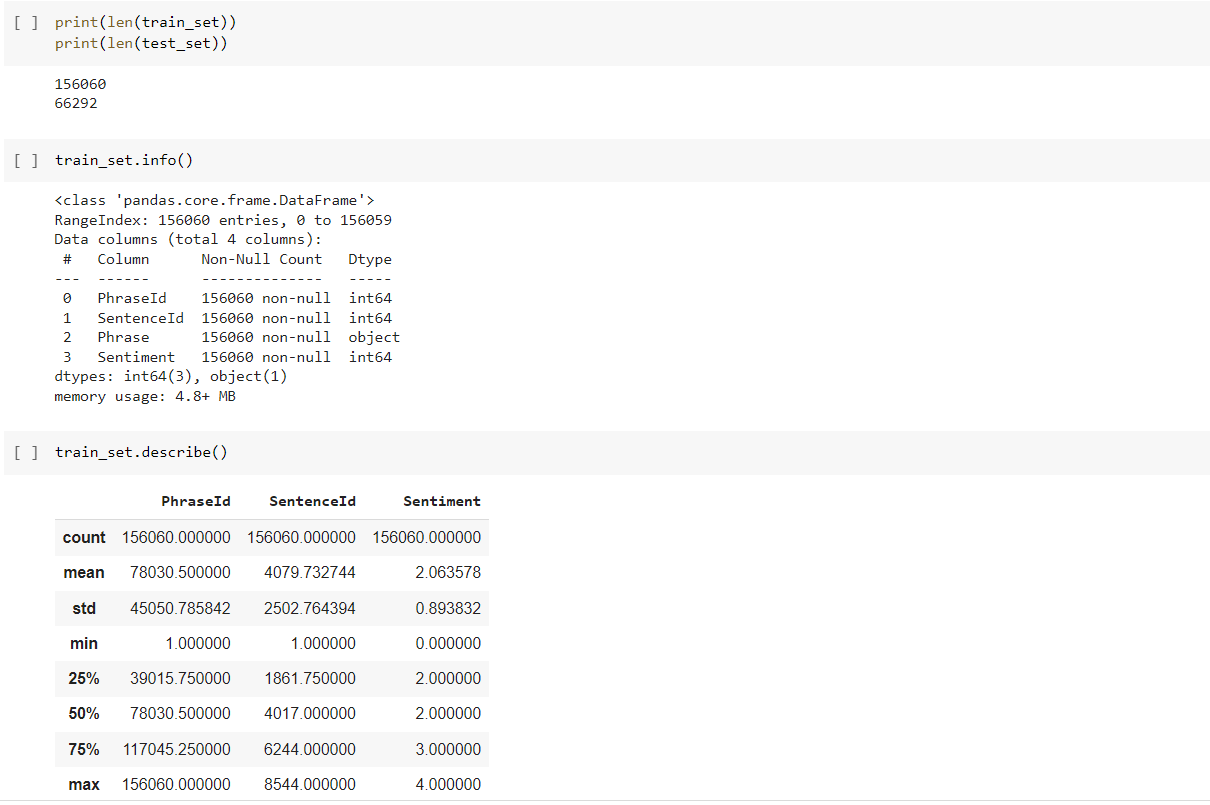
Here, the users can rate and review the movies. Also, they get the ratings in the form of stars and in the form of numerical score out of 5 (representing the number of stars). The users can view various movie reviews and accordingly get the movie recommendation. In the website, they can also choose from different genres to make our search field narrower. Some of the famous genres are: horror, comedy, drama, romcom, scifi, tech, fictional, musical, etc.

CODE:

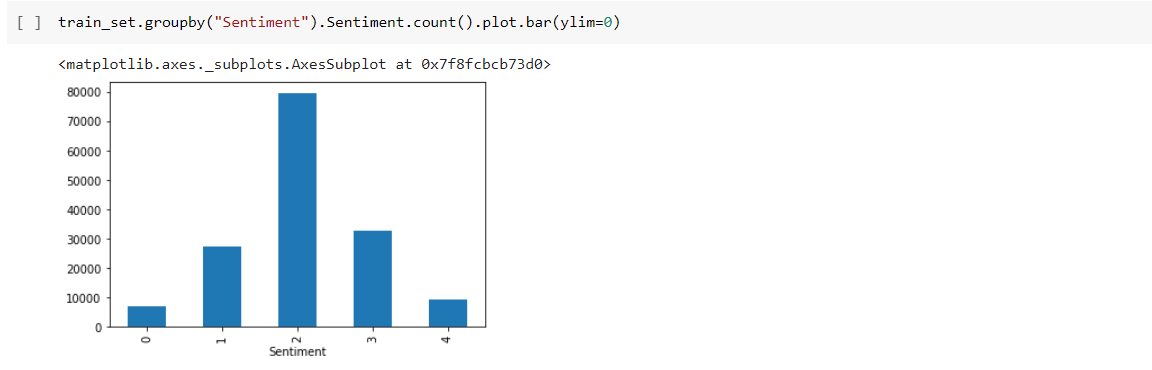


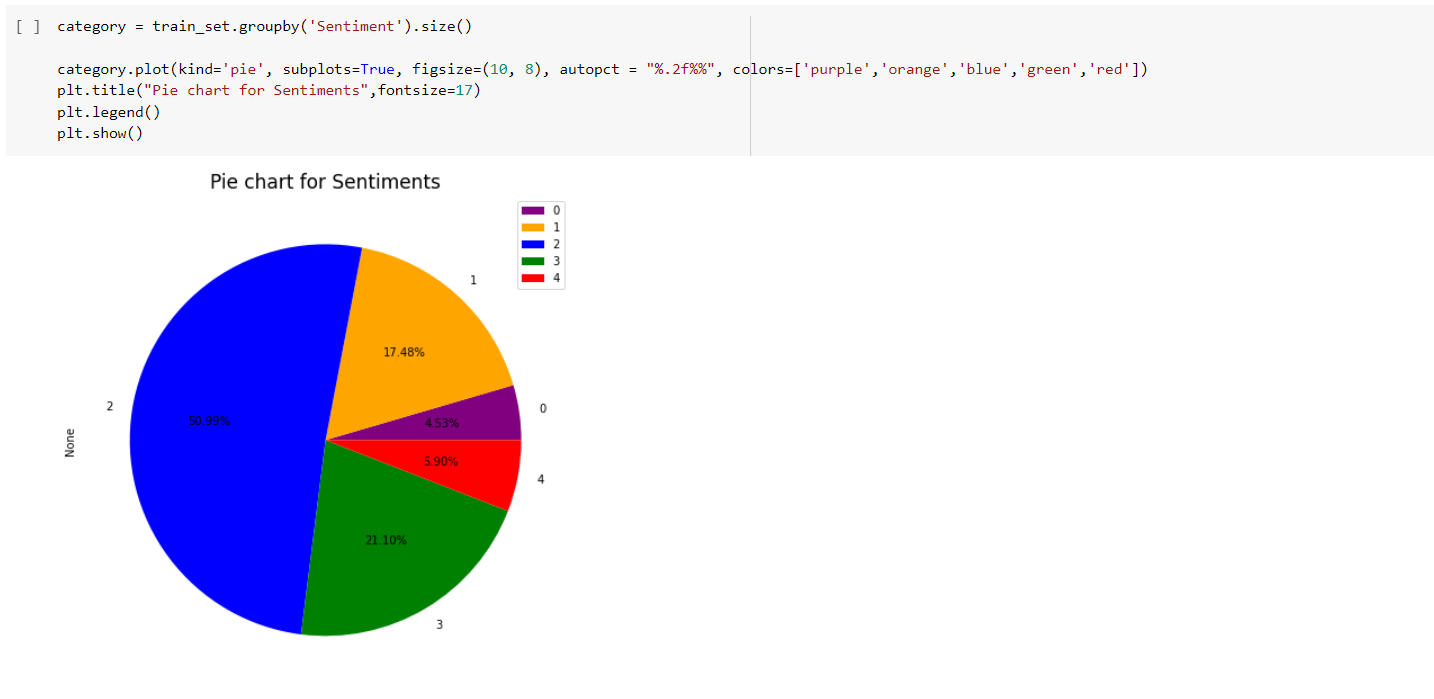


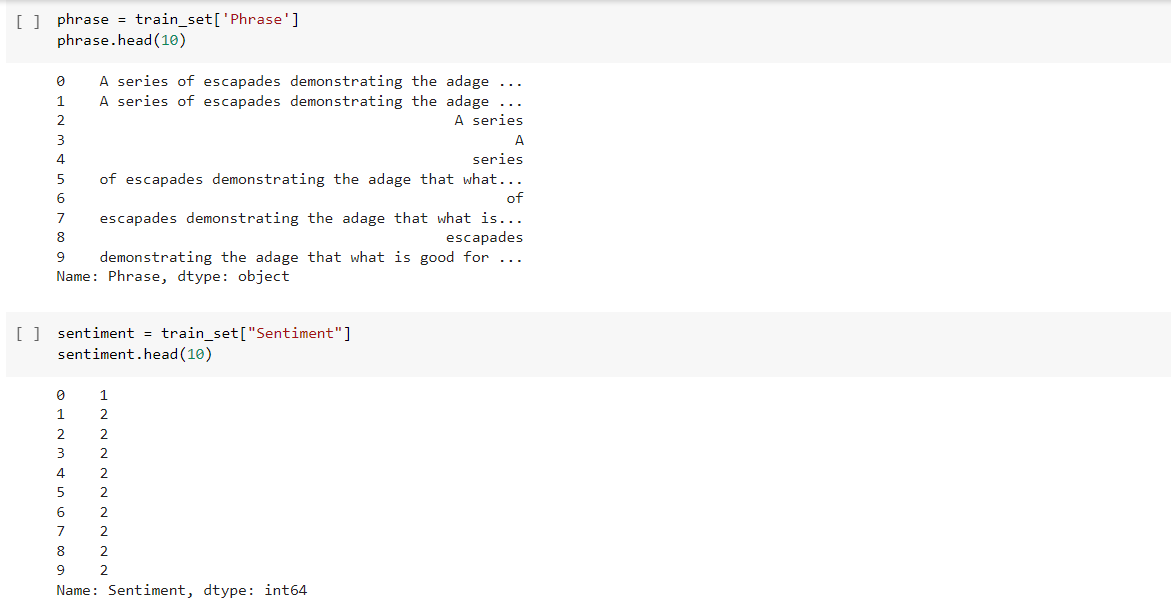




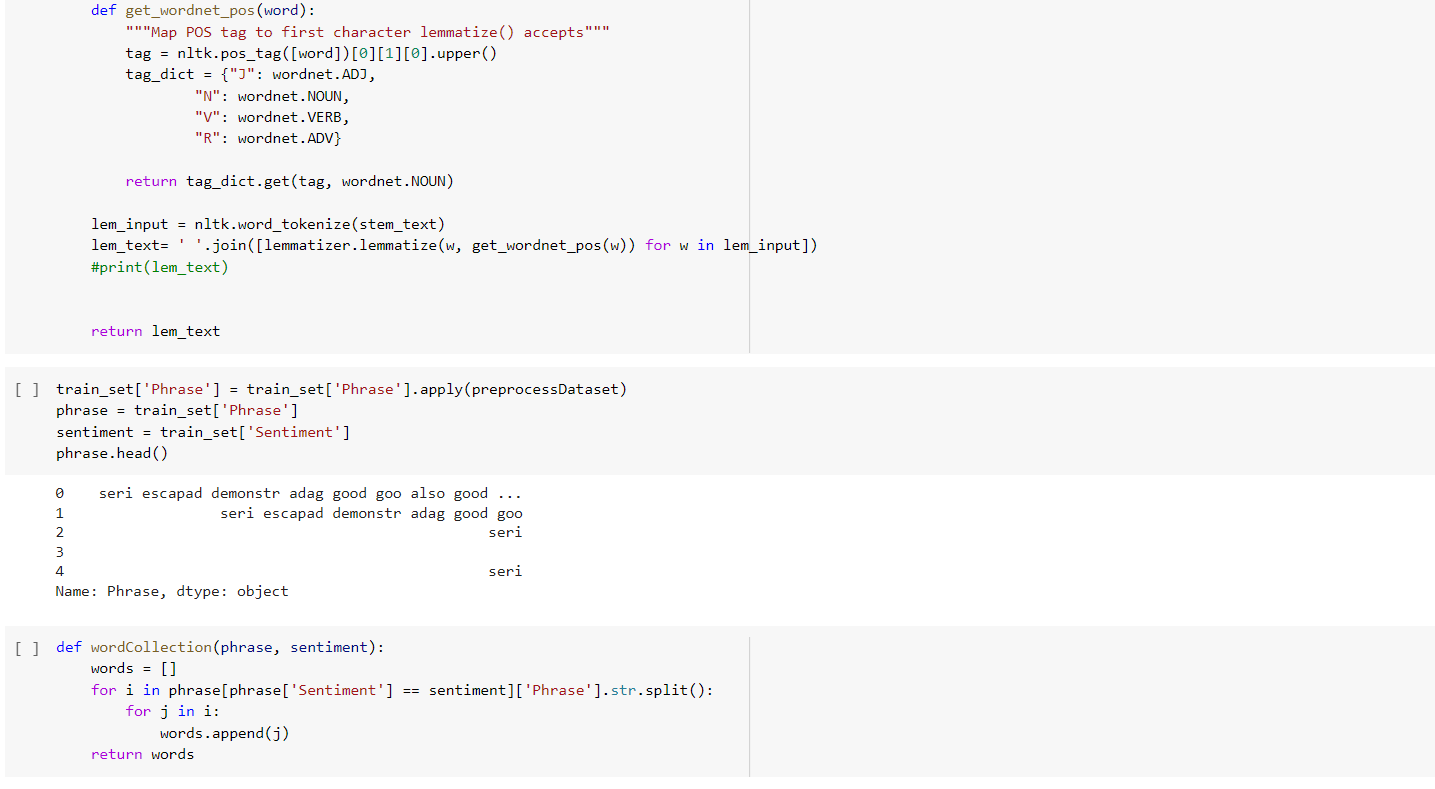


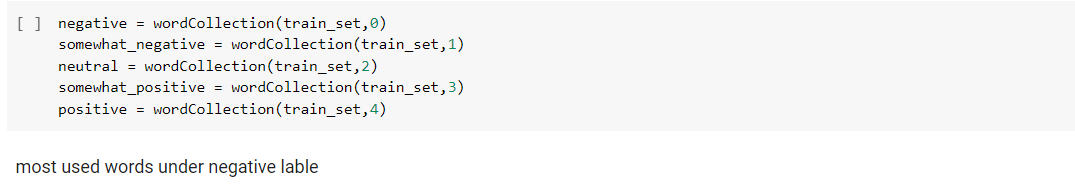


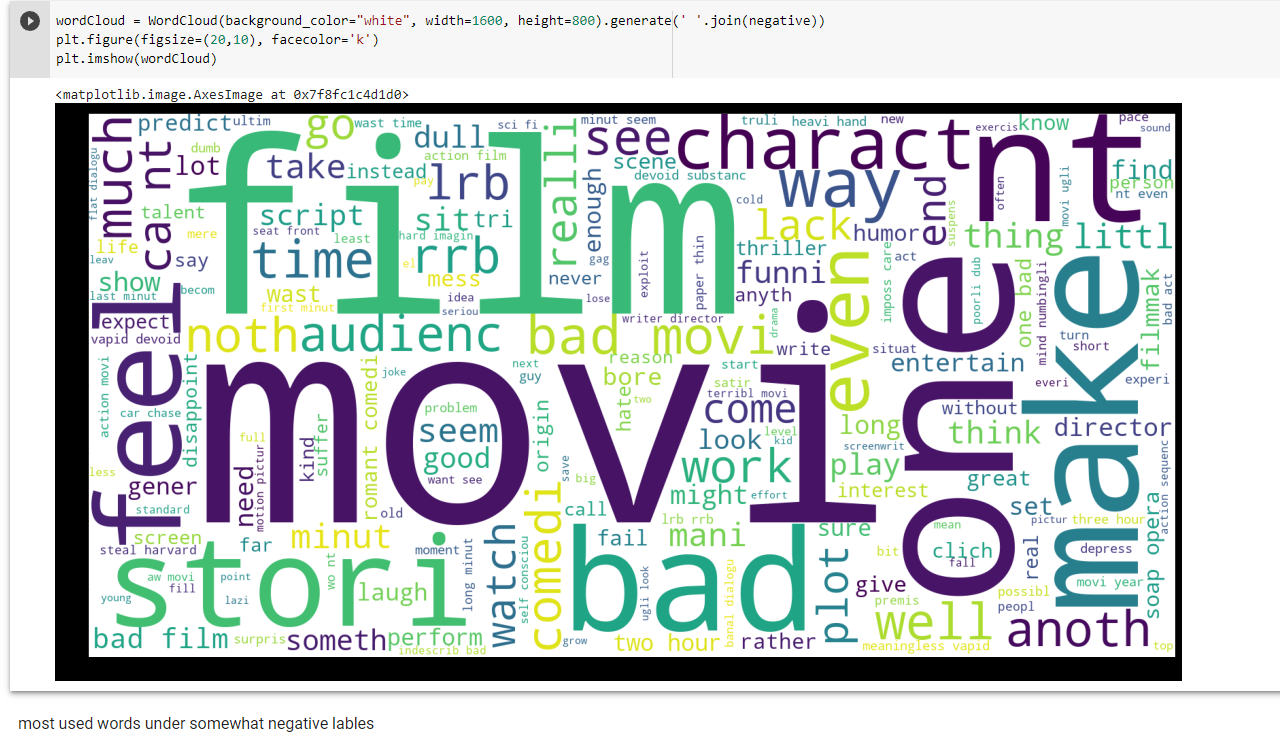




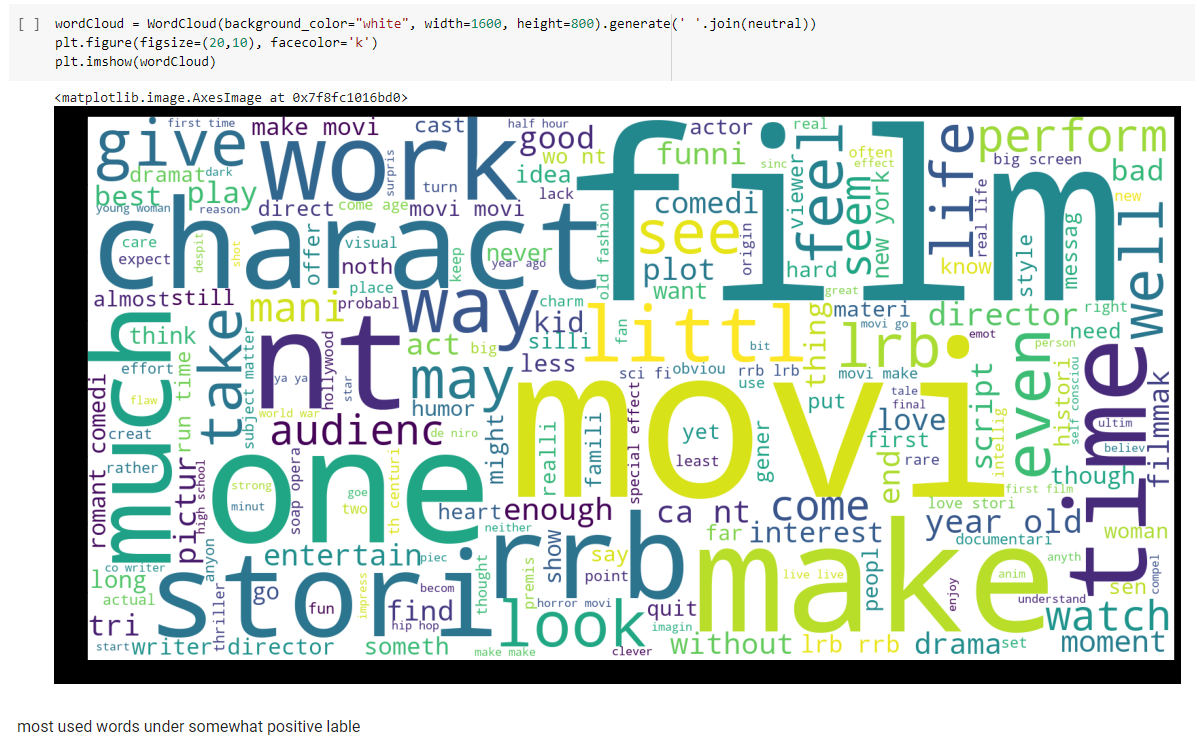


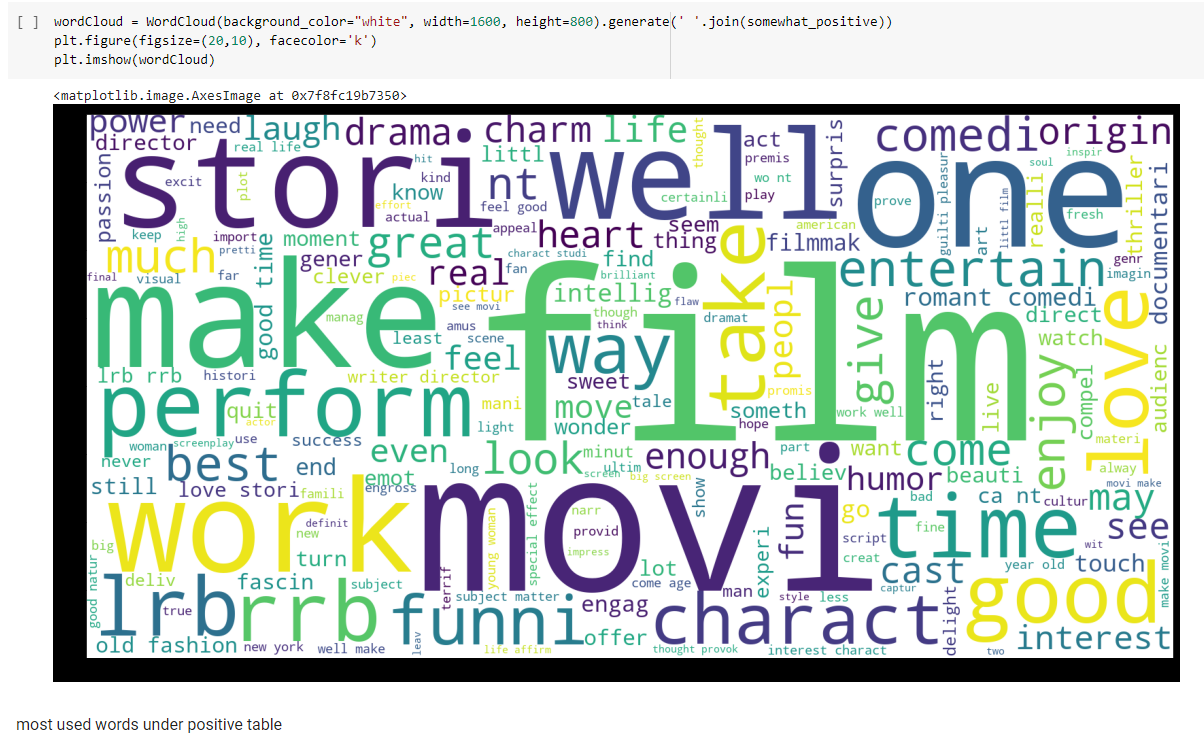


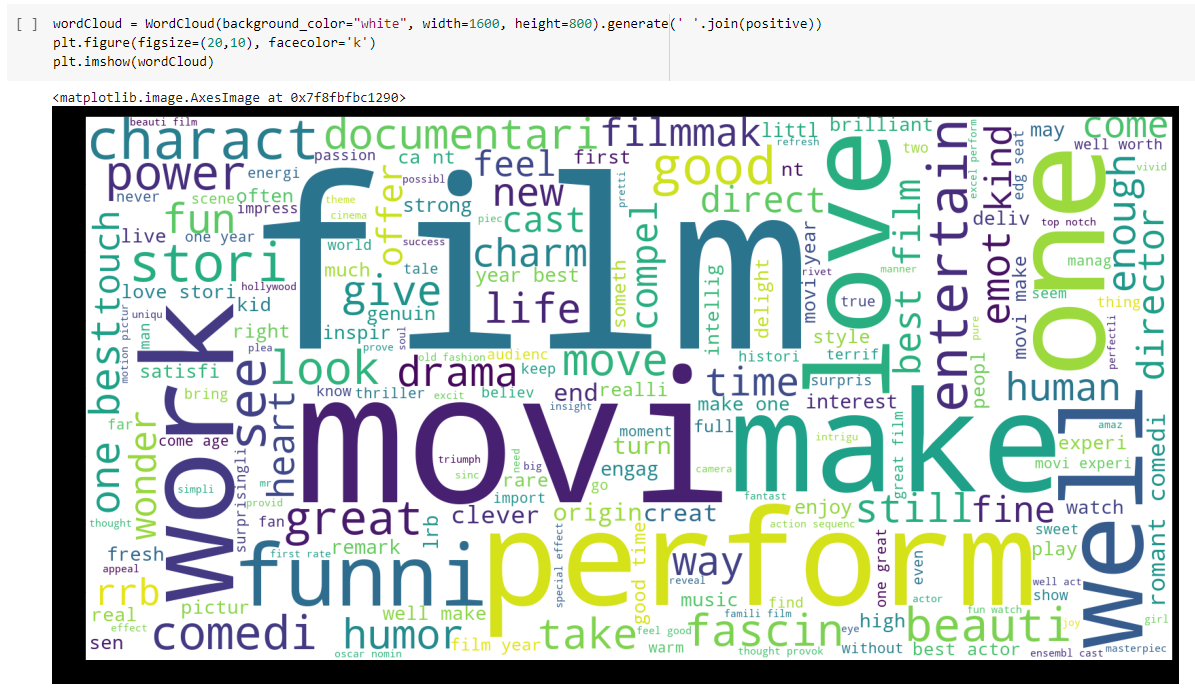


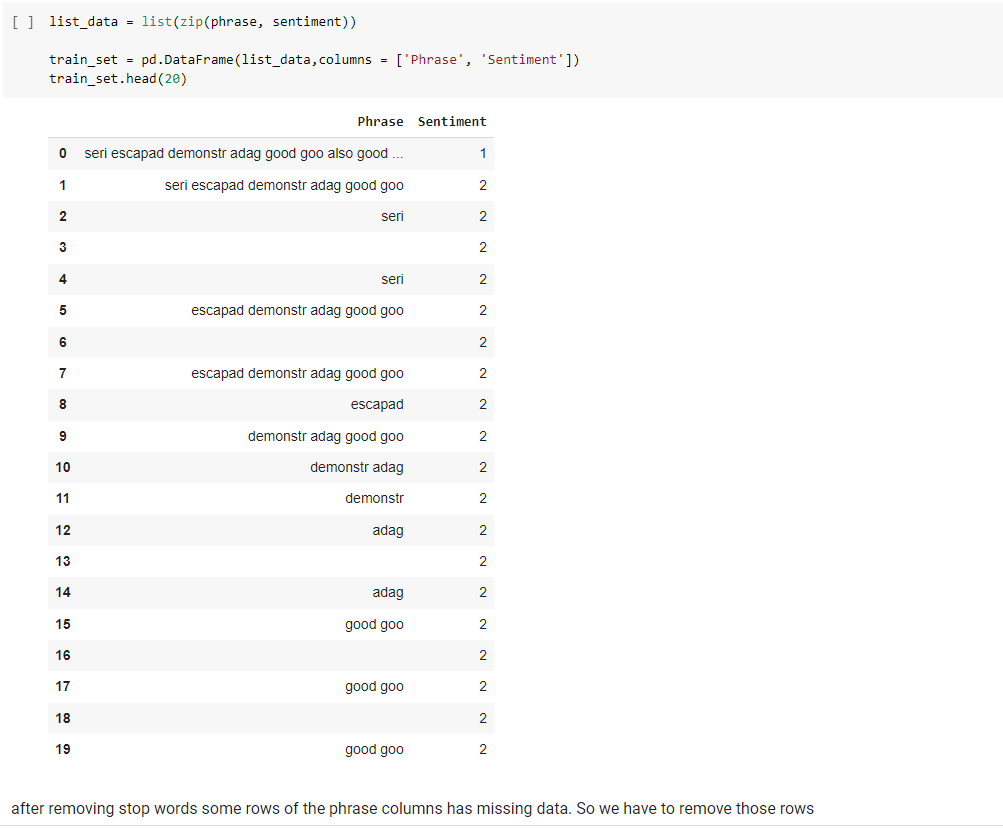


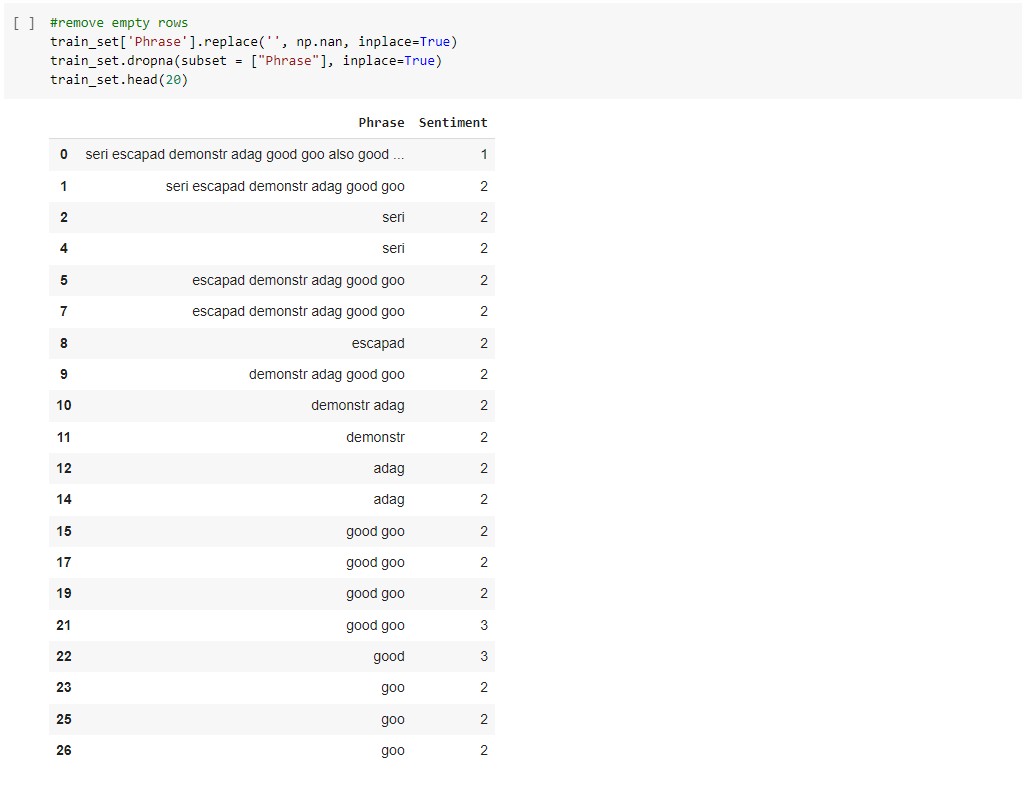




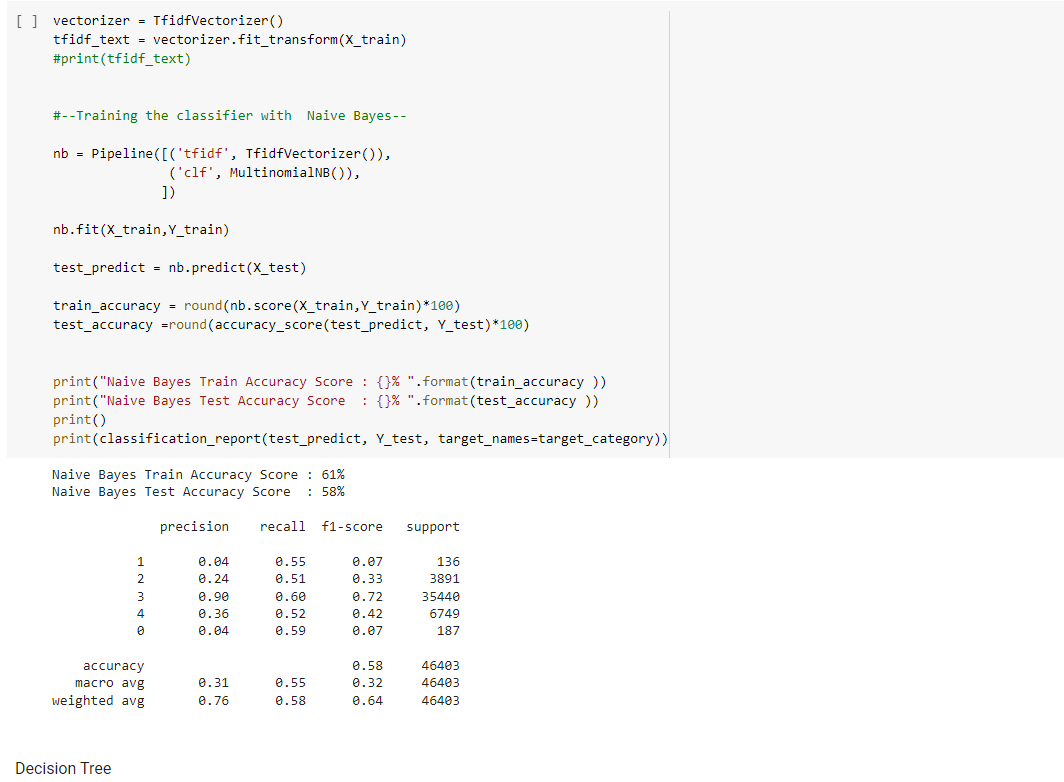


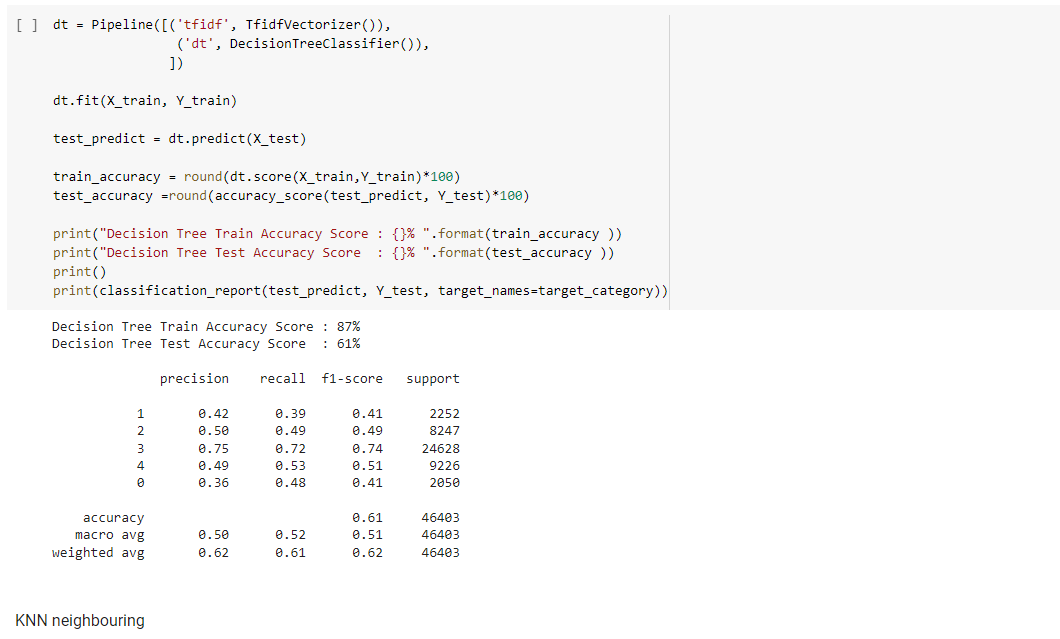


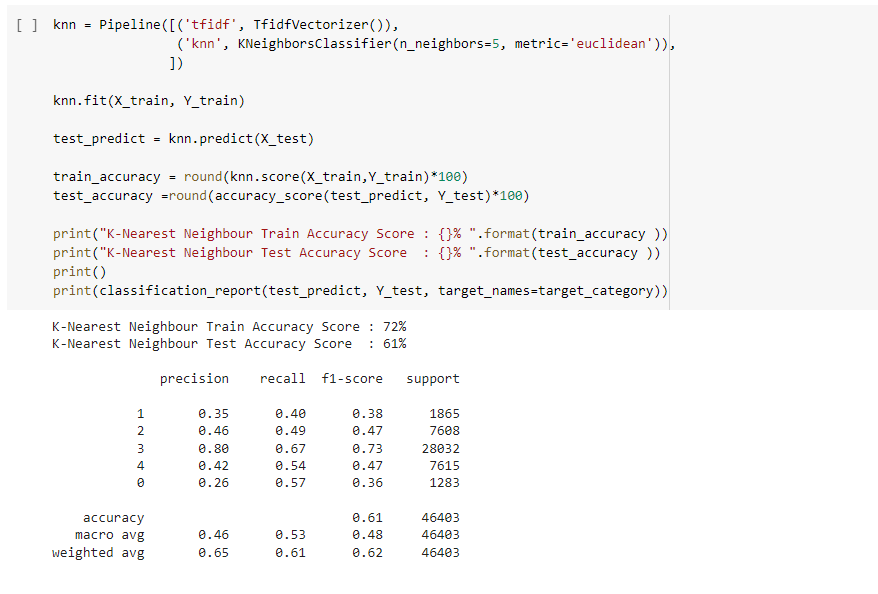


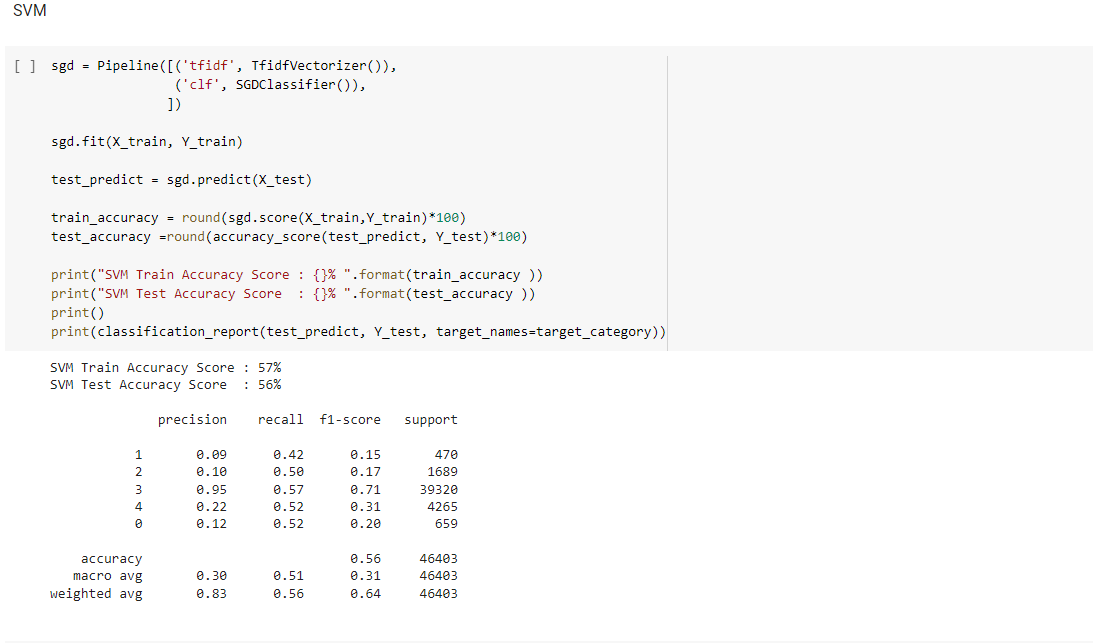


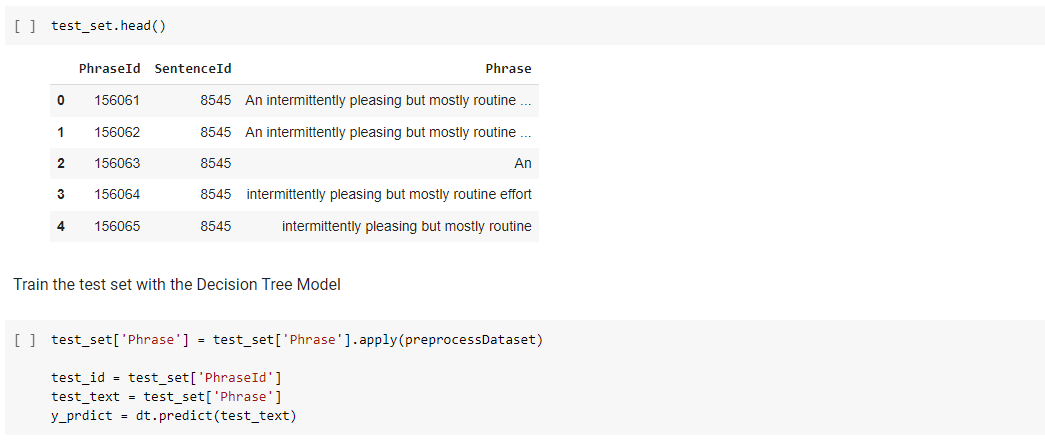


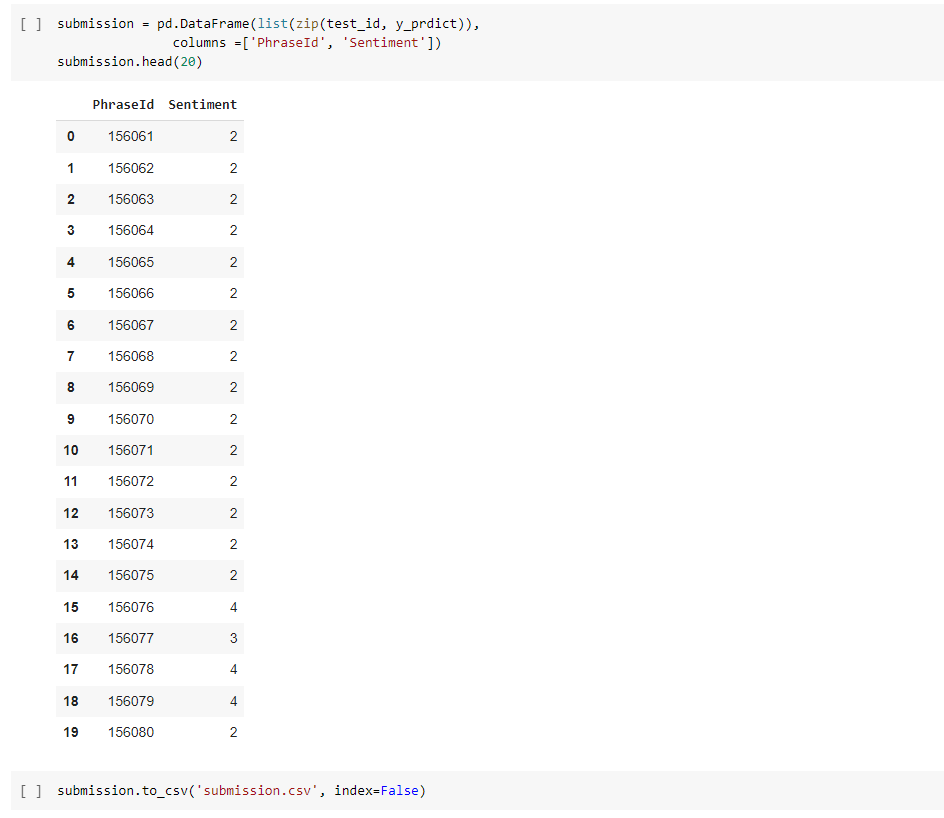












V. Conclusion

From the results above, we can infer that for our problem statement or model, Decision Tree model with feature set using mixture of Unigrams and Bigrams, preprocessing the model is best. Apart from this one can also use a KNN (k neighbors classifiers) as this provides a good accuracy percentage. Among the four classifiers used, Decision Tree with the highest accuracy rate of 87% train set and 61 % of test set followed by KNN with accuracy rate of 72% train set and 61% of test set and Naïve Bayes with accuracy rate of 61% train set and 58% test set with the worst accuracy rate of SVM of 57% train set and 56% test set.

FAQS

**What is sentiment analysis in NLP?**

Sentiment Analysis (also known as opinion mining or emotion AI) is a sub-field of NLP that tries to identify and extract opinions within a given text across blogs, reviews, social media, forums, news etc.

**How many types of sentiments are there?**

There are two main types of sentiment analysis: subjectivity/objectivity identification and feature/aspect-based sentiment analysis.

**How accurate is sentiment analysis?**

When evaluating the sentiment (positive, negative, neutral) of a given text document, research shows that human analysts tend to agree around 80-85% of the time. ... But when you're running automated sentiment analysis through natural language processing, you want to be certain that the results are reliable.

**How is sentiment analysis useful?**

Sentiment analysis is extremely useful in social media monitoring as it allows us to gain an overview of the wider public opinion behind certain topics. ... The overall customer experience of your users can be revealed quickly with sentiment analysis, but it can get far more granular too.

**What is polarity and subjectivity in sentiment analysis?**

Subjective sentences generally refer to personal opinion, emotion or judgment whereas objective refers to factual information. ... We can see that polarity is 0.8, which means that the statement is positive and 0.75 subjectivity refers that mostly it is a public opinion and not a factual information.

**What are sentiment indicators?**

A sentiment indicator refers to a graphical or numerical indicator designed to show how a group feels about the market or economy. ... Sentiment indicators show how bullish or bearish a group of people are, which may help forecast this group's future behavior, often in a contrarian way.

**What is polarity score in sentiment analysis?**

The key aspect of sentiment analysis is to analyze a body of text for understanding the opinion expressed by it. Typically, we quantify this sentiment with a positive or negative value, called polarity. The overall sentiment is often inferred as positive, neutral or negative from the sign of the polarity score.

**What is emotion analysis?**

Sentiment analysis and emotional analysis are two key methods experts use to quantify audiences' emotional engagement. You can use them in your content strategy to unveil readers' emotional responses to your content. If you manage to uncover how people feel about your content, you can easily make it perfect.

# REFERENCES

1. M, Mamtesh, and Seema Mehla. “Sentiment Analysis of Movie Reviews Using Machine Learning Classifiers.” International Journal of Computer Applications, vol. 182, no. 50, Nov. 2019, pp. 25–28., doi:10.5120/ijca2019918756.
2. Nanda, Charu, et al. “Sentiment Analysis of Movie Reviews in Hindi Language Using Machine Learning.” 2018 International Conference on Communication and Signal Processing (ICCSP), 2018, doi:10.1109/iccsp.2018.8524223.
3. Wang, G., Sun, J., Ma, J., Xu, K., & Gu, J. (2014). Sentiment classification: The contribution of ensemble learning. Decision Support Systems,57, 77-93. doi:10.1016/j.dss.2013.08.002
4. Tripathy, A., Agrawal, A., & Rath, S. K. (2016). Classification of sentiment reviews using n-gram machine learning approach. Expert Systems with Applications,57, 117-126. doi:10.1016/j.eswa.2016.03.028
5. Kaur, B., & Kumari, N. (2016). A Hybrid Approach to Sentiment Analysis of Technical Article Reviews. International Journal of Education and Management Engineering,6(6), 1-11.doi:10.5815/ijeme.2016.06.01
6. Melville, Prem, et al. “Sentiment Analysis of Blogs by Combining Lexical Knowledge with Text Classification.” Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD 09, 2009, doi:10.1145/1557019.1557156.
7. Mukherjee, Arjun, et al. “Spotting Fake Reviewer Groups in Consumer Reviews.” Proceedings of the 21st International Conference on World Wide Web - WWW 12, 2012, doi:10.1145/2187836.2187863.
8. Liu, Bing, and Lei Zhang. “A Survey of Opinion Mining and Sentiment Analysis.” Mining Text Data, 2012, pp. 415–463., doi:10.1007/978-1-4614-3223-4\_13.
9. Khan, Khairullah, et al. “Mining Opinion Components from Unstructured Reviews: A Review.” Journal of King Saud University - Computer and Information Sciences, vol. 26, no. 3, 2014, pp. 258–275., doi:10.1016/j.jksuci.2014.03.009.
10. Araque, Oscar, et al. “A Semantic Similarity-Based Perspective of Affect Lexicons for Sentiment Analysis.” Knowledge-Based Systems, vol. 165, 2019, pp. 346–359., doi:10.1016/j.knosys.2018.12.005.
11. Araque, Oscar, et al. “Enhancing Deep Learning Sentiment Analysis with Ensemble Techniques in Social Applications.” Expert Systems with Applications, vol. 77, 2017, pp. 236–246., doi:10.1016/j.eswa.2017.02.002.
12. Rathor, A. S., Agarwal, A., & Dimri, P. (2018). Comparative Study of Machine Learning Approaches for Amazon Reviews. Procedia Computer Science,132, 1552-1561. doi:10.1016/j.procs.2018.05.119
13. Cernian, A., Sgarciu, V., & Martin, B. (2015). Sentiment analysis from product reviews using SentiWordNet as lexical resource. 2015 7th International Conference on Electronics, Computers and Artificial Intelligence (ECAI). doi:10.1109/ecai.2015.7301224
14. J, A., & G, J. (2018). Sentiment Classification of Tweets with Non-Language Features. Procedia Computer Science,143, 426-433. doi:10.1016/j.procs.2018.10.414
15. Devi, S. Sathiya, and G. Parthasarathy. “A Hybrid Approach for Movie Recommendation System Using Feature Engineering.” 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), 2018, doi:10.1109/icicct.2018.8473335.