SENTIMENT ANALYSIS ON MOVIE REVIEWS

Prepared for

Professor: CHRISTOPHE SERVAN

Subject: NATURAL LANGUAGE PROCESSING

Prepared by

NAVEEN KUMAR CHOLLANGI

&

CHETNA JAYAKUMAR

Specialization: Data Science and Analytics

GitHub Link

EPITA - SCHOOL OF ENGINEERING AND COMPUTER SCIENCE

09 July 2022

ABSTRACT

The demand for the automatic classification of electronic documents has skyrocketed along with the rapid development of text sentiment analysis. In recent years, a lot of research has been done on the text classification or text mining paradigm. In this project, we offer a term frequency-inverse document frequency technique for text sentiment classification (TF-IDF)

After applying our suggested model to three different text mining algorithms, we discovered that the Linear Support Vector Machine (LSVM) is best suited to use with our suggested model. When compared to earlier methods, the results obtained show a significant increase in accuracy.

INTRODUCTION

Opinion mining also referred to as sentiment analysis or emotion AI, is the systematic identification, extraction, measurement, and analysis of affective states and subjective data using natural language processing, text analysis, computational linguistics, and biometrics. Sentiment analysis is frequently used in marketing, customer service, and clinical medicine applications. It is applied to the voice of the customer materials like reviews and survey responses, online and social media, and healthcare materials.

Generally speaking, sentiment analysis aims to ascertain the general contextual polarity or emotional response to a document, interaction, or event as well as the attitude of a speaker, writer, or other subjects concerning a given topic. The attitude could be an assessment or judgment (see the appraisal theory), an affective state (i.e., the author's or speaker's emotional state), or the intended emotional communication (that is to say, the emotional effect intended by the author or interlocutor).

ABOUT THE DATASET

This is a dataset for binary sentiment classification containing substantially more data than previous benchmark datasets. The goal would be to produce a high-performing sentiment analyzer by training it on the available rows. The structure of the CSV file is quite simple, it has two columns, one containing the reviews, and the other one the sentiment. Once the model will re-classify part of the reviews on the testing portion, we'll be able to calculate how many were correctly classified which indicates the overall accuracy of the SVM model. https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews

GOAL

This project aims at deploying a machine learning model called "Support Vector Machines" with a particular focus on text cleaning and hyperparameters optimization, these two techniques will most likely increase the model accuracy.

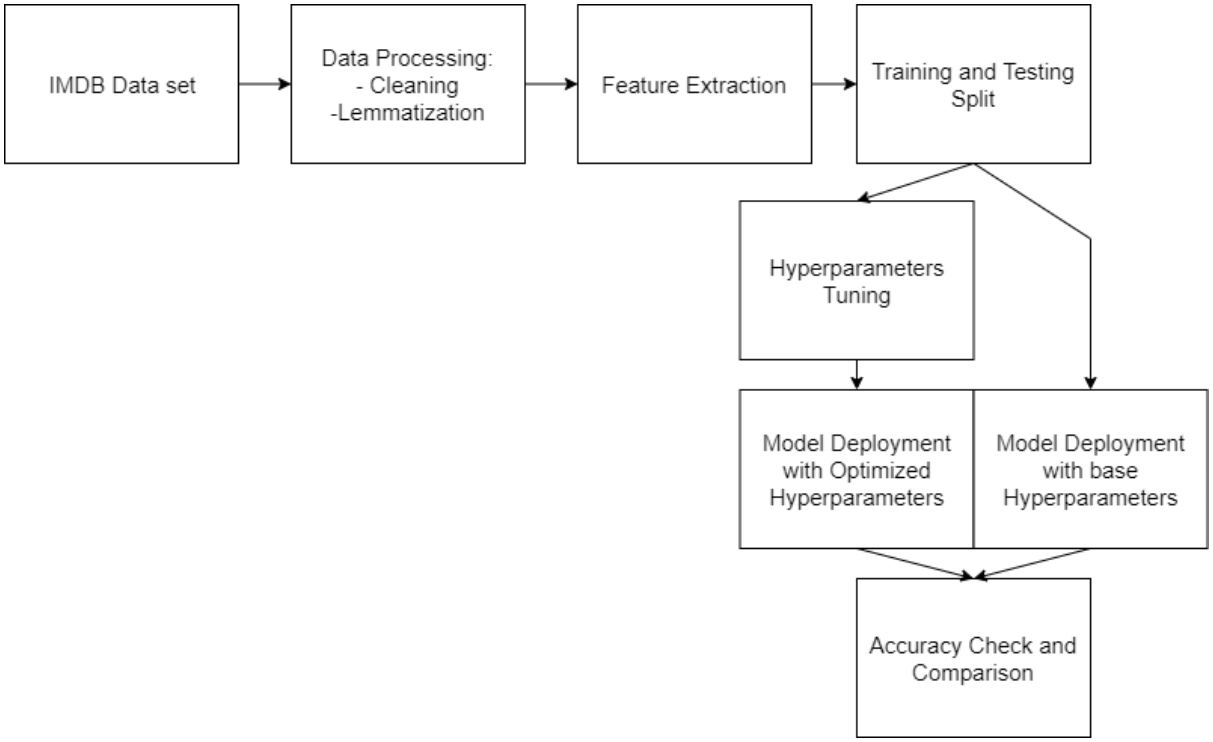
TEXT CLEANING / TEXT PROCESSING

Cleaning, pre-processing, and normalizing text to bring text components like phrases and words to some standard format is one of the key steps before going into the process of feature engineering and modeling.

TEXT NORMALIZATION

Words are tokenized. To separate a statement into words, we utilize the Spacy library and import the STOP_WORDS method.

We have used the **Spacy library** as it gives the best ways to access many algorithms. The dataset has two major fields namely Reviews and Sentiments for training our model.



REMOVING HTML STRIPS AND NOISE TEXT

Here in the data head, we can see some HTML code so, in the beginning, we need to clean those HTML strips. Also, removing some noisy texts along with square brackets.

REMOVING SPECIAL CHARACTERS

Because we're working with English-language evaluations in our dataset, we need to make sure that any special characters are deleted.

REMOVING STOP WORDS AND NORMALISING

Stop words are words that have little or no meaning, especially when synthesizing meaningful aspects from the text. Stop words are words that are filtered out of natural language data (text) before or after it is processed in computers. While “stop words” usually refer to a language’s most common terms, all-natural language processing algorithms don’t employ a single universal list. Stop words include words such as a, an, the, and others.

Text normalization is the process of converting previously uncanonical text into a single canonical form. Because the input is guaranteed to be consistent before operations are done on it, normalizing text before storing or processing it allows for separation of concerns.

As a part of Data exploration & preparation, the data_prep() function checks for null values, converting all the texts to lower to ease the pre-processing, removes emails, removes URLs, removes special characters, removal HTML tags using **beautiful soup,** removal of stop words using the spacy library and then applying this function to the reviews field to implement these modifications.

TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY MODEL

(TF-IDF)

It is used to convert text documents to a matrix of tf-idf features. The term frequency-inverse document frequency statistic is a numerical measure of how essential a word is to a document in a collection.

Word Embeddings, also known as Word Vectorization, is an NLP technique for mapping words or phrases from a lexicon to a corresponding vector of real numbers, which can then be used to derive word predictions and semantics. Vectorization is the process of translating words into numbers.

To convert the texts in the Reviews column into valuable and meaningful vectors, we are using the **TF-IDF vectorization** technique. Data Splitting into x and y variables to use for training and testing the model. Also, performing fitting even on the TF-IDF variable to increase the accuracy of our model.

For modeling, we have chosen the **LinearSVC model** as we have compared 3 other models namely XGboost, Random Forest, multinomial and naïve Bayes

In the end, we tested with two random reviews picked from another dataset,to test our model. The predicted results are accurate and it displays in terms of the reviews on movies.

CONCLUSION

The base feature extractor and base model configurations already had a high accuracy score, but the hyperparameter optimization process was able to raise it even higher. Machine learning continues to astound us with its ability to "learn" and classify records more quickly and effectively than humans. It is truly amazing that a "simple" statistical algorithm may be able to perform tasks faster and more accurately than people. Although there are still some restrictions, technology has finally made it possible for almost everyone to create their own models and research this extraordinary field.

RELATED WORKS

https://www.researchgate.net/publication/346511493_Sentiment_Analysis_of_IMDb_Movie_Reviews_Using_Long_Short-Term_Memory

https://www.researchgate.net/publication/330014159_Sentiment_Analysis_on_IMDb_Movie_Reviews_Using_Hybrid_Feature_Extraction_Method

http://cs229.stanford.edu/proj2020spr/report/Wu_Shin.pdf

https://ieeexplore.ieee.org/document/9137994

https://arxiv.org/ftp/arxiv/papers/1806/1806.06407.pdf