**Sentiment-Analysis-on-Movie reviews**

Prepared by

NAVEEN KUMAR CHOLLANGI, CHETNA JAYAKYMAR

To

Professor: Christophe Servan

EPITA: SCHOOL OF ENGINEERING AND COMPUTER SCIENCE

**Abstract**

The demand for 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.

## **Sentiment Analysis**

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 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 subject with respect to 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).

# Dataset

This is a dataset for binary sentiment classification containing substantially more data than previous benchmark datasets. We provide a set of 25,000 highly polar movie reviews for training and 25,000 for testing. The goal would be to produce a high-performing sentiment analyser by training it on the available rows. The structure of the csv file is quite simple, it has two columns, one contains 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.

**Introduction:**

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

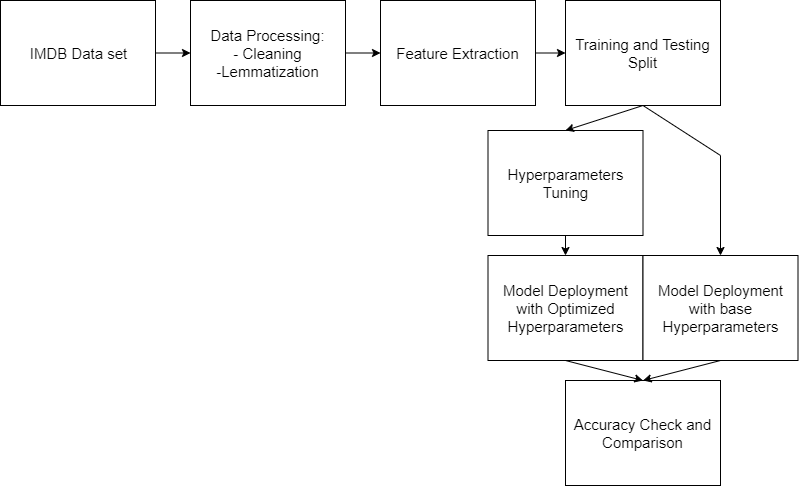
**Text cleaning/ text processing :**

Cleaning, pre-processing, and normalising 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 modelling.

**Text normalization:**

Words are tokenized. To separate a statement into words, we utilise the Spacy library and importing STOP\_WORDS method.

We have used **Spacy library** as it gives 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 data head we can see some html code so first we need to clean that 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 normalization**

Stop words are words that have little or no meaning, especially when synthesising 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 refers 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 normalisation is the process of converting previously uncanonical text into a single canonical form. Because input is guaranteed to be consistent before operations are done on it, normalising text before storing or processing it allows for separation of concerns.

As a part of Data exploration & preparation, 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 of html tags using **beautiful soup,** removal of stop words using spacy library and then applying this function to the reviews field to implement these modifications.

**Term Frequency-Inverse Document Frequency model (TFIDF)**

It is used to convert text documents to matrix of tfidf 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 Reviews column into valuable and meaningful vectors, we are using **TF-IDF** **vectorization** technique. Data Splitting into x and y variables to use them for training and testing the model. Also, performing fitting even on TF-IDF variable to increase the accuracy of our model.

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

As the end, we have tested with two random reviews picked from other 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 hyperparameters 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.