MIT | Academy of Engineering

Machine Learning Practical Examination

Group Members :
Nehaal Pandey   Sunit Trivedi   Divyansh Sawant

Course Instructor : Prof.Diptee Ghusse

Introduction

1. Sentiment analysis, also known as opinion mining, is a popular area of natural language processing that involves identifying and categorizing subjective information in text.
2. With the rise of social media and other online platforms, sentiment analysis has become increasingly important for businesses and organizations to gauge public opinion and sentiment about their products, services, and brands.
3. In this project, we aim to design and develop a binary classification model for sentimental analysis based on user input answers to three questions.
4. The model will be trained on a dataset of labeled user responses, and will classify user responses as either positive or negative sentiment.
5. Additionally, we will provide generalized video suggestions based on the user's mood, to help them find content that aligns with their emotional state.
6. Overall, our goal is to create a user-friendly tool that can provide valuable insights into user sentiment and help users discover relevant video content based on their mood.

Problem Statement

To design and develop a Binary Classification model for Sentimental Analysis based on User-Input answers to three questions and to provide generalised video suggestions based on the mood of the user.

Objectives

- To collect information from the user.

- Preprocess Text Data

- Train a sentiment analysis model.

- Evaluate model accuracy

- Give generalized Video suggestions

About Dataset

The dataset being used is the **sentiment140 dataset**. It contains 1,600,000 tweets extracted using the **Twitter API**. The tweets have been annotated **(0 = Negative, 4 = Positive)** and they can be used to detect sentiment.
**It contains the following 6 fields:**

1. **sentiment**: the polarity of the tweet *(0 = negative, 4 = positive)*
2. **ids**: The id of the tweet *(2087)*
3. **date**: the date of the tweet *(Sat May 16 23:58:44 UTC 2009)*
4. **flag**: The query (lyx). If there is no query, then this value is NO_QUERY.
5. **user**: the user that tweeted *(robotickilldozr)*
6. **text**: the text of the tweet *(Lyx is cool)*

We require only the **sentiment** and **text** fields, so we discard the rest.
Furthermore, we're changing the **sentiment** field so that it has new values to reflect the sentiment. **(0 = Negative, 1 = Positive)**

Methodology
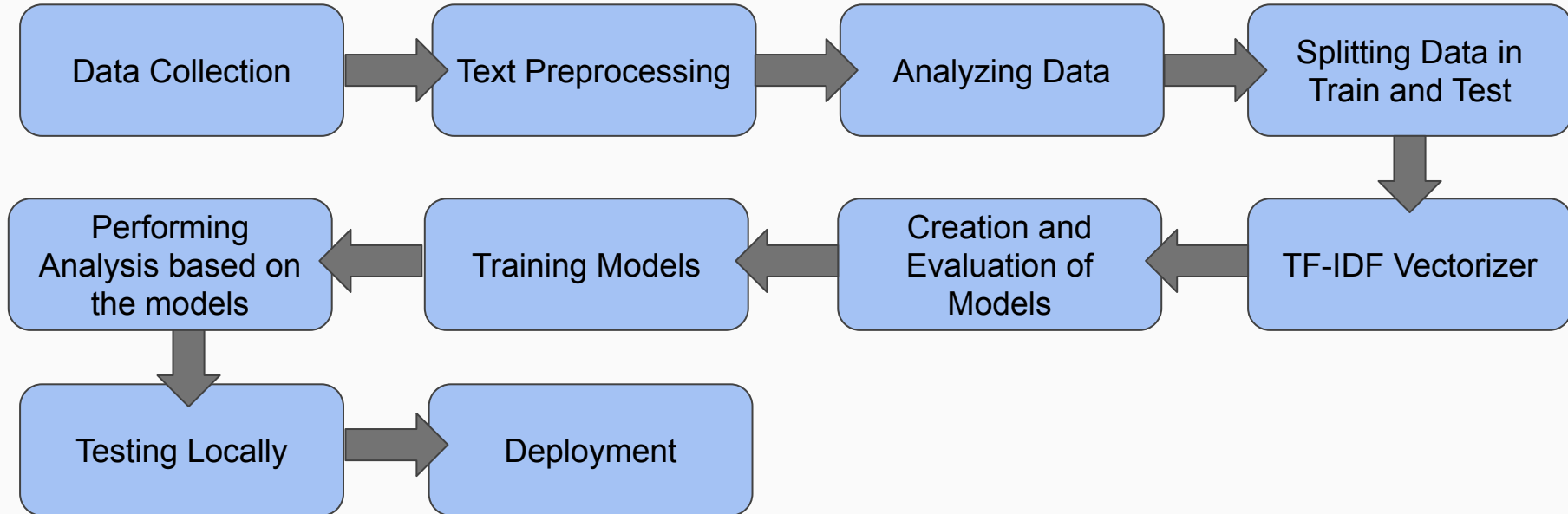
Preprocessing of Dataset

The preprocessing steps followed on this dataset are :

1. **Lower Casing** : Each text is converted to lowercase.
2. **Replacing URLS** : Links starting with 'www' or 'http' are replaced with 'URL'.
3. **Replacing Emojis** : The emojis mentioned in the tweets are replaced with a meaning which describes it with the help of predefined dictionary containing emojis along with their meanings.
4. **Replacing Usernames** : Replace @usernames with the word 'USER'.
5. **Removing Non Alphabets** : Replacing characters except digits and alphabets with space.

Preprocessing of Dataset

6.   **Removing Consecutive Letters** : 3 or more consecutive letters are replaced by 2 letters.

7.   **Removing Short Words** : Words with length less than 2 are removed.

8.   **Removing Stopwords** : Stopwords are the words which do not much add meaning to the sentence.

9.   **Lemmatizing** : Lemmatization is the process of converting a word to its base form.

Training and Testing Model

We're creating 3 different types of model for our sentiment analysis problem:

- **Logistic Regression (LR)**
- **Support Vector Classification (SVC)**
- **Naive Bayesian**
- **Decision Tree**
- **Random Forest**

Results Analysis and Comparison with different models
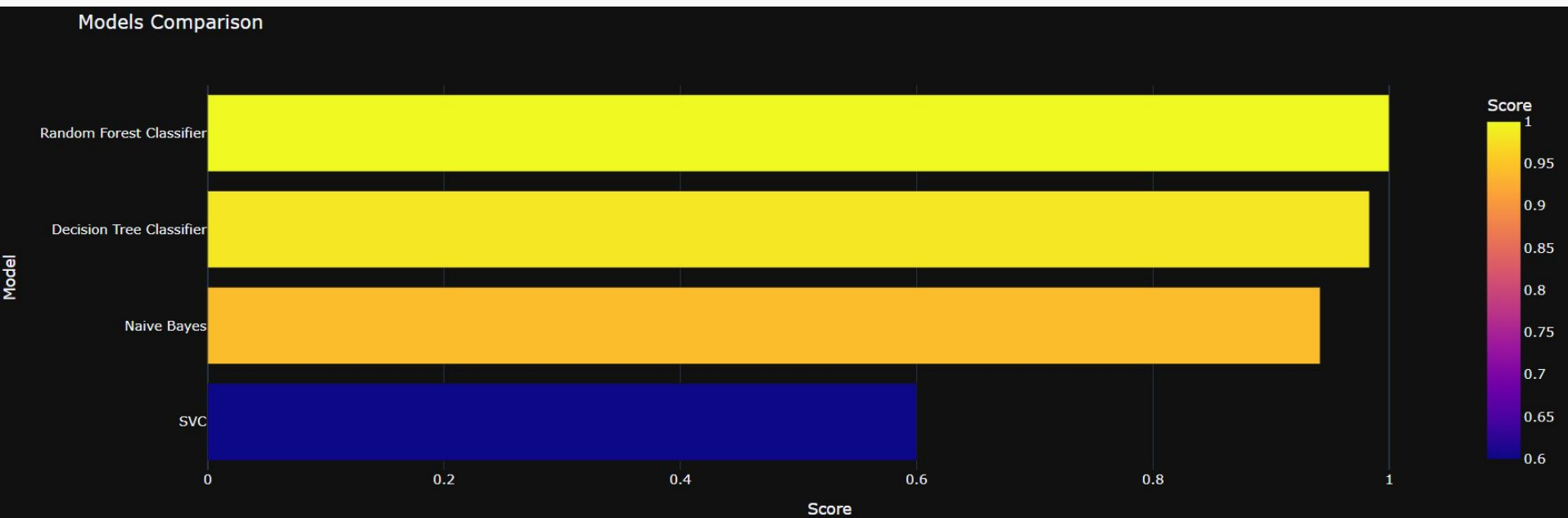
Conclusion

- In conclusion, developing a sentiment analysis project that recommends generalised videos based on a user's mood can be a powerful tool for improving well-being and promoting positive emotions.

- This project has the potential to be useful in a variety of settings, such as mental health counseling, employee well-being programs, and general entertainment applications. As natural language processing and machine learning continue to advance, it is likely that such projects will become even more powerful and sophisticated in the future.