Project Report: Social Media Sentiment Analysis

1. Introduction

The rapid expansion of social media platforms has created an enormous volume of user-generated content, reflecting diverse sentiments. Understanding these sentiments is crucial for businesses, researchers, and policymakers. This project focuses on analyzing sentiments from social media posts, classifying them into four categories: positive, negative, neutral, and irrelevant.

2. Dataset Collection

For this project, the dataset was obtained from a publicly accessible repository containing tweets labeled with corresponding sentiment classes. The specific dataset used is the "Twitter Sentiment Analysis" dataset, which includes numerous tweets along with their sentiment labels. This dataset was chosen for its relevance and richness in textual data, allowing for effective training and evaluation of sentiment analysis models.

The dataset was loaded directly from a CSV file hosted on GitHub (link: https://raw.githubusercontent.com/laxmimerit/All-CSV-ML-Data-Files-Download/master/twitter_sentiment.csv)

After loading the dataset, only the relevant columns were selected, which consist of sentiment labels and the tweet texts. The columns were renamed for better clarity ['sentiment', 'text'].

Sentiment dataset contains:

Negative - 22808

Positive - 21109

Neutral - 18603

Irrelevant - 13162

3. Data Preprocessing

Data preprocessing is essential in preparing raw data for analysis. In this project, several steps were taken to clean and preprocess the dataset:

- **Lowercasing**: All text data was converted to lowercase to ensure uniformity across the dataset.
- **Removing Unwanted Content**: URLs, HTML tags, special characters, and retweets were removed to focus on the sentiments expressed in the text.
- **Handling Missing Values**: Any missing text entries were filled with empty strings to avoid disruptions during processing.
- **Text Cleaning**: Various preprocessing functions were applied to clean the data thoroughly.

These preprocessing steps significantly enhanced the quality of the input data, contributing to the improved performance of the sentiment analysis models. For URLs, HTML tags, special characters, and retweets were removed using the preprocess-kgptalkie library.

4. Model Training

4.1. Applied Algorithms

Four different machine learning algorithms were selected for training and evaluation:

1. **Naive Bayes Classifier**: A simple probabilistic model that is quick to train and effective for text classification tasks.
2. **Decision Tree Classifier**: A model that creates a tree-like structure for making decisions based on feature values, providing a clear interpretation of the decision-making process.
3. **Support Vector Machine (SVM)**: A powerful classifier that constructs hyperplanes in high-dimensional space to separate different sentiment classes, known for its robustness.
4. **Random Forest Classifier**: An ensemble method that builds multiple decision trees and combines their predictions, offering enhanced accuracy and reduced risk of overfitting.

4.2. Train-Test Split

The dataset was divided into training and testing sets, with 80% of the data used for training and 20% for testing.

4.3. Text Vectorization

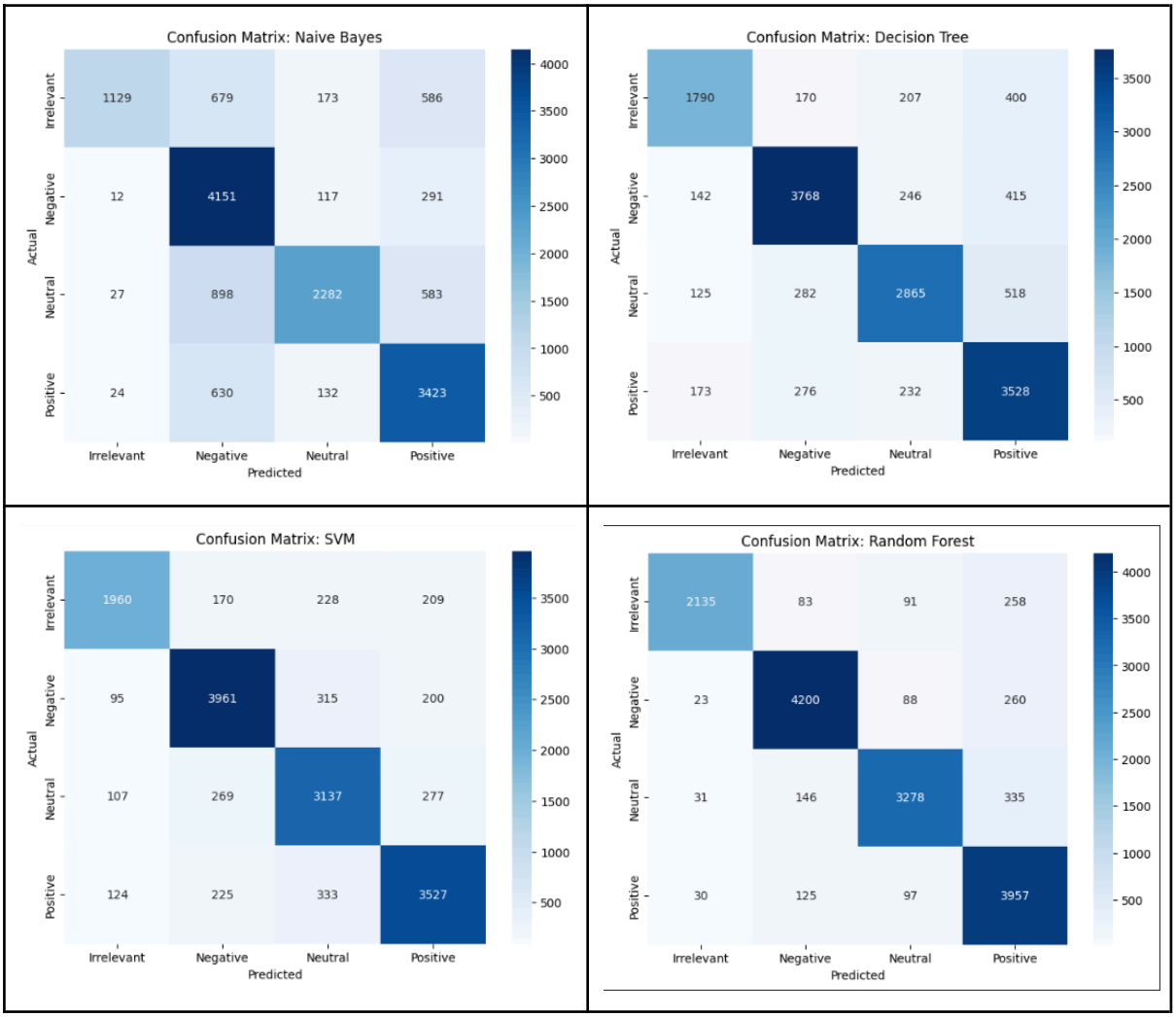
TfidfVectorizer is a feature extraction technique from the sklearn library. It converts text data into a matrix of numerical features, which machine learning models can understand. TF-IDF stands for Term Frequency-Inverse Document Frequency, which is a measure that evaluates how important a word is to a document in a collection (or corpus). This vectorization was utilized to convert the text data into a format suitable for model training.

5. Model Evaluation

The performance of each model was evaluated based on its accuracy, which measures the percentage of correctly classified instances. The results of the models are summarized in the following table:

Model	**Accuracy(%)**
Naive Bayes	72.57
Decision Tree	78.95
Support Vector Machine (SVM)	83.14
Random Forest	89.65

Confusion Matrix:



Analysis of Results

- **Naive Bayes Classifier**: Achieved an accuracy of 72.57%. While it is a straightforward approach, it struggled with the complexities of the text data.
- **Decision Tree Classifier**: Improved accuracy to 78.95%, effectively capturing non-linear relationships but susceptible to overfitting.
- **Support Vector Machine (SVM)**: Achieved an accuracy of 83.14%, demonstrating robustness in high-dimensional spaces.
- **Random Forest Classifier**: Outperformed all other models with an accuracy of 89.65%, effectively handling diverse text patterns and minimizing overfitting risks.

6. Conclusion

This project successfully demonstrated the effectiveness of multiple machine learning algorithms in classifying sentiments from social media posts. The Random Forest classifier emerged as the most effective model, providing the highest accuracy and robust generalization capabilities.

The findings of this project highlight the importance of selecting appropriate models based on specific requirements in sentiment analysis. Future work may include exploring advanced deep learning techniques for even better performance and expanding the dataset for improved generalization.

7. Future Work

Future enhancements could focus on the following areas:

- Implementing deep learning models such as LSTM (Long Short-Term Memory) or Transformers to improve accuracy further.
- Expanding the dataset to include more diverse social media platforms for broader applicability.
- Incorporating sentiment analysis of multilingual data to increase the project's relevance in global contexts.

8. References:

[1] https://raw.githubusercontent.com/laxmimerit/All-CSV-ML-Data-Files-Download/master/twitter_sentiment.csv

[2] https://github.com/laxmimerit/preprocess_kgptalkie

[3] https://scikit-learn.org/stable/supervised_learning.html

[4] https://jakevdp.github.io/PythonDataScienceHandbook/