Tweet Sentiment Analysis

AI Project

Group Members

Muhammad Ahsan Rahim -14817

Ammar Ahmed Alvi - 14867

Video Link:

https://drive.google.com/drive/folders/1J2P84Io-5JDSUQ9UUSgIUQo-YhUdhaQK?usp=sharing

Project Title

Tweet Sentiment Analysis

Problem Overview:

Social listening is an extremely important part of any brand's social media audit, as brands need to monitor the sentiment towards their brand, products and niche and twitter has become the hub of customer feedback these past days.

Therefore we have decided to create a tweet sentiment analyzer in which we've tried and trained multiple machine learning models to fit a set of random and general tweets to identify emotions like happy, sad, angry etc and we've then selected the best performing model and used it to perform sentiment analysis on tweets related to a particular keyword extracted directly off twitter using the Twitter API.

Data Collection

For training and testing, we have used the SMILE Twitter Emotion dataset which contains around 3000 tweets that can be found at:

https://figshare.com/articles/smile_annotations_final_csv/3187909/2

For the application users, we have added a tweet extractor code in which a person can pass in a keyword and number of tweets to extract directly off twitter using the Twitter API, however in order to use this a person must authenticate the Twitter API using their own keys.

Data Preprocessing

- Removed tweets with no sentiment and not-relevant tags
- Removed tweets with multiple sentiments
- Removed Stopwords from tweets
- Used a Term frequency - inverse document frequency (TF-IDF) vectorizer
- Tested multiple n-grams. (Bigrams performed best)

Models:

Naive Bayes:

A Naive Bayes classifier is an algorithm that uses Bayes' theorem to classify objects. Naive Bayes classifiers assume strong, or naive, independence between attributes of data points.

Random Forest with Gini:

Random forest, like its name implies, consists of a large number of individual decision trees that operate as an ensemble.

K - Nearest Neighbours:

K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (which in our case is the euclidean distance).

Gradient Boosting:

Gradient Boosting trains many models in a gradual, additive and sequential manner. It identifies the shortcomings of models in an ensemble by using high weight data points and uses gradients in the loss function to optimize the ensemble.

Multi-Layer Perceptron (MLP) Neural Network:

Neural networks are a set of algorithms, modeled loosely after the human brain, that are designed to recognize patterns.

Support Vector Machine:

A support vector machine (SVM) is a supervised machine learning model that uses classification algorithms for two-group classification problems. After giving an SVM model sets of labeled training data for each category, they're able to categorize new text.

Model	***Accuracy***
Naive Bayes	0.9108
Random Forest	0.9134
K-Nearest Neighbour	0.8450
Gradient Boosting	0.9060
Multi-Layer Perceptron	0.9160
Support Vector Machine	0.8022