TWITTER(X) SENTIMENT ANALYSIS USING NATURAL LANGUAGE PROCESSING

CS595 – Termination Project

Venkata Sri Sai Surya Vinay Upadhyayula
B00981964

Table of Contents

Abstract

This project aims to study the sentiment dataset from Twitter(X) and analyze it using Natural Language processing and Machine Learning. The aim is to preprocess the labelled dataset, train a machine learning model, predict and evaluate the model's performance.

Sentiment analysis, also known as opinion mining, is a natural language processing (NLP) and machine learning (ML) technique used to determine the sentiment expressed in a piece of text. The goal of sentiment analysis is to identify and extract subjective information from the text and classify it as positive, negative, or neutral.

The following are the general steps one usually follows for sentimental analysis:

Text Preprocessing: The input text is preprocessed to remove irrelevant information, such as stop words, punctuation, and special characters. It may also involve stemming or lemmatization to reduce words to their base form.

Feature Extraction: Relevant features are extracted from the preprocessed text. These features could include words, n-grams (sequences of adjacent words), or other linguistic elements.

Model Training: Machine learning models, such as classification algorithms, are trained on labeled datasets. These datasets consist of text samples with corresponding sentiment labels (positive, negative, or neutral).

Sentiment Classification: Once the model is trained, it can be used to predict the sentiment of new, unseen text. The model assigns a sentiment label (positive, negative, or neutral) based on the learned patterns from the training data.

Evaluation: The performance of the sentiment analysis model is evaluated using metrics like accuracy, precision, recall, and F1 score. This helps assess how well the model generalizes to new data.

There are two libraries namely 'spacy' and 'nltk' used for dealing with natural language related processing.

Spacy: NLP library for implementing algorithms. It is faster and efficient, but user doesn't have a choice to choose different algorithms that we need.

NLTK

Natural Language Tool Kit is an open-source NLP library that has many functionalities and choice for the user to choose his best fit of algorithms but is less efficient comparative to Spacy.

In the context of sentiment analysis, spacy doesn't include models which is typically easier to perform using NLTK.
Owing to the above reasons I have used nltk for this project.

Stemming and lemmatization are both techniques used in natural language processing (NLP) to reduce words to their base or root form. However, they operate differently and serve different purposes.

Stemming

- Stemming is the process of reducing words to their word stem or root form by removing suffixes.
- The goal of stemming is to reduce words to a common base form, even if the stem itself may not be a valid word.
- Stemming is typically faster and more straightforward than lemmatization, but it may not always produce valid words.

Example:

- Stemming the words "running", "runs", and "runner" would result in the common stem "run".

NLTK used 2 algorithms for stemming:

1)Porter's algorithm – based on longest suffix rule (simple phrase)
2) Snowball – English/ Porter 2 stemmer

Lemmatization

- Lemmatization, on the other hand, is the process of reducing words to their base or dictionary form (lemma) while considering the context and meaning of the word.
- Lemmatization uses linguistic rules and morphological analysis to accurately identify the lemma of a word.
- Lemmatization usually produces valid words that exist in the dictionary.

Example:

- Lemmatizing the words "running", "runs", and "runner" would result in the lemma "run".

Stop Words

Words like "a" and "the" appear so frequently that they don't require tagging as thoroughly as nouns, verbs and modifiers. We call these stop words, and they can be filtered from the text to be processed. NLTK holds inbuilt library that holds common stop words in the English language. It can be used as follows,

```
import nltk
from nltk.corpus import stopwords
nltk.download('stopwords')
STOPWORDS = set(stopwords.words("english"))
```

Scikit Learn

Scikit learn library is used for all the functions like numpy, pandas, matplotlib and many others for data preprocessing, visualizing data and machine learning model training, prediction and performance metrics.

Term Frequency - Inverse Document Frequency (TF-IDF) Vectorization

Most machine learning models understand numeric data where most of the real world data cannot be and is not and so is out twitter sentiment analysis data. After cleaning the data as part of further preprocessing we assign and vectorize the text data into numeric vector using TF - IDF vectorization.
Once all the preprocessing steps including stemming, lemmatization, stop word removal, removing irrelevant words(cleaning) is done, we feed the data into the machine learning model by splitting the model into train and test dataset.

Train-Test Split

Train-test split is a technique used to evaluate the performance of machine learning models.
It involves splitting the dataset into two subsets: the training set and the testing set.
The training set is used to train the model, while the testing set is used to evaluate its performance.
Typically, the training set contains a larger portion of the data, around 70-80%, while the testing set contains the remaining 20-30%.

Model Fitting

Model fitting is the process of training a machine learning model using the training data.
The model learns the patterns in the training data and adjusts its parameters to minimize the prediction error.
I have used Logistic Regression and Support vector machine models to classify the data.

Logistic Regression

This is one of the powerful classifiers in machine learning. It comes to the class of discriminative classifier and uses a logistic/sigmoid function as its hypothesis space that maps our features to labels. In our code let us break done each parameter passed to the Logistic Regression function in the sklearn library. **The accuracy achieved with this model is 89%.**
LogisticRegression(C=2, max_iter=1000, n_jobs=-1)

Larger values of `C` reduce the regularization strength, allowing the model to fit the training data more closely.
`max_iter` is the maximum number of iterations taken for the optimization algorithm to converge.
Optimization algorithms like gradient descent are used to find the optimal coefficients for logistic regression.

Support Vector Machine

It is one of the powerful machine learning algorithms used for regression and classification problems. SVM comprises of a hyperplane that divides the classes (say in our case). So, the key idea behind is to find a hyperplane that best separates different classes in the feature space. But the true goal is to find the plane that not only separates the classes but also maximizes the margin, providing the best generalization of unseen data. It can be used for both linear and nonlinear classification. It focuses more on correctly classifying the most critical instances (support vectors). **The accuracy achieved with this model is 92%.**

Model Evaluation

Model evaluation is the process of assessing the performance of the trained model using the testing data.
In real world scenario it is not enough to have a single metric as it may not encompass all the scenarios needed to evaluate the model. So, we have many metrics such as:

1) Accuracy = no of correct predictions/ total no of predictions
-- It is useful when we have balanced featured data (i.e.., 50% HAM, 50% SPAM data in the feature set)
--Not very useful when we have un - balanced data (i.e., 99%SPAM, 1% HAM data in the feature set)

2)Precision = no of True Positive / no of True Positive + no of False positive
- It is the ability to identify only relevant data points i.e., proportion of what is relevant

3)Recall = no of True Positive / no of True Positive + no of False negative
- It is the ability to find all the relevant cases within a dataset

4)F1 Score = 2* (precision*recall/precision +recall)
- It is the harmonic mean of recall and precision.

Confusion matrix: It gives the ratio of TP, TN, FP and FN's

Technologies/Tools

Python, Sklearn, pandas, numpy libraries.

Code

The code is executed in a google collab runtime environment.

References

To succeed in this project, I chose NLP - Natural Language Processing with Python by Jose Portilla as a guide