Twitter Sentiment Analysis

by Chionh Wan Sim, Hiew Shani, Kushioka Nodoka, Yeo Fu Kai Marcus (Group 13)

Introduction

This project aims to analyse different machine learning techniques in their ability to classify a Twitter tweet to be of positive or negative sentiment. Techniques to be analysed are: Random Forest, K-nearest Neighbours (KNN), and Logistic Regression.

Setting Up

Using Google Colab:

1. Download the .ipynb file onto your local desk drive
2. Go to https://colab.research.google.com/
3. Go to 'Upload' and press 'Browse' or drag the downloaded file into the page.

Requirements

You may install the packages inside the requirements.txt file:

```
pip install -r /path/to/requirements.txt
```

Instructions

1. Visualising the data

1. Run: 'Importing the relevant libraries'
2. Click on the 'Exploratory Data Analysis' section in the Google Colab table of contents
3. To visualise the label distribution of the dataset, run the code under the subsection 'Visualising the label distribution of the dataset'.
4. To visualise the distribution of the length of tweets using a histogram, run the code under the subsection 'Visualising the distribution of length of tweets'.
5. To visualise the distribution of tweet characteristics, run the code under the subsection 'Visualising number of different characteristics of tweets with respect to the total number of tweets'.

2. Running the models

1. Run the following sections:
    a. 'Importing the relevant libraries'
    b. 'Importing twitter set from the Google Drive link' under the section 'Exploratory Data Analysis'
    c. 'Data Preprocessing'
    d. 'One hot encoding (Bag of words)'
2. To run random forest and logistic regression, run 'Model Training: Random Forest & Logistic Regression'
3. To run k-Nearest Neighbors, run 'Model Training: K-Nearest Neighbours (KNN)'.

Contents

1. Importing of Relevant Libraries
2. Exploratory Data Analysis
    a. Importing Twitter Dataset
    b. Visualising label distribution of dataset
    c. Visualising distribution of length of tweets
    d. Visualising distribution of tweet characteristics
3. Data Preprocessing
    a. Removing Usernames
    b. Removing Links
    c. Removing Punctuations, Numbers, and Special Characters
    d. Tokenisation
    e. Stemming and Lemmatisation
    f. Removing Stopwords
4. Data Visualisation (After Preprocessing)
    a. Visualising most common words across entire preprocessed dataset
    b. Visualising most common words that are being labelled as ‘Negative’
    c. Visualising most common words that are being labelled as ‘Positive’
5. One hot encoding
6. Model Training: Random Forest & Logistic Regression
    a. Hyperparameter Tuning for Random Forest & Logistic Regression
7. Model Training: K-Nearest Neighbours (KNN)
    a. Hyperparameter Tuning for KNN