PROJECT DOCUMENTATION

-BY SHYAM PRAKASH J
CODALAB USERNAME: shyam_j08

PROJECT TITLE: EMOTION INTENSITY FOR TWEETS

TABLE OF CONTENTS

INTRODUCTION:

We utilize words to express not only the emotion we are experiencing, but also its strength. For example, our words can convey that we are really angry, slightly sad, completely pleased, and so on. The degree or amount of an emotion, such as anger or grief, is referred to as intensity. Detecting the speaker's emotional intensity automatically has applications in business, public health, intelligence gathering, and social welfare.
The ultimate goal of this project is to recognize the emotional intensity of a tweet's speaker.
This particular goal was realized through the use of three separate types of approaches: statistical, deep learning, and machine learning. In this article, we'll go over each of these approaches in detail.

BACKGROUND AND SIGNIFICANCE:

Existing emotion datasets are primarily tagged categorically, with little indication of emotional intensity. Furthermore, practically all of the problems are structured as categorization tasks (identify 1 among n emotions for this sentence). In contrast, knowing the extent to which an emotion is expressed in text is frequently useful for applications. This is the first task in which systems must determine the intensity of emotions in tweets automatically.

TASK:

Determine the intensity or degree of emotion X felt by the speaker given a tweet and an emotion X — a real-valued score between 0 and 1. The highest attainable score 1 denotes experiencing the most

quantity of emotion X. (or having a mental state maximally inclined towards feeling emotion X). The lowest possible score of 0 indicates that you are experiencing the least amount of emotion X. (or having a mental state maximally away from feeling emotion X). A tweet with the emotion X will be used as an example. It's important to note that the absolute scores have no inherent meaning; they're only utilised to show that cases with higher scores correspond to a higher level of emotion X than instances with lower scores.

DATA:

For four emotions, training and test datasets are provided: joy, sorrow, fear, and rage. The anger training dataset, for example, contains tweets with a real-valued score between 0 and 1 reflecting the speaker's level of rage. Only the tweet text is included in the test data. After the evaluation period, the gold emotion intensity scores will be released.

PROJECT DESCRIPTION:

Importing the necessary libraries for Data Pre-processing, Plotting, NLP, Model implementation, and evaluation is the first and most important stage in this project.
The exploratory data analysis is the next crucial step to do. EDA is used to uncover the underlying structure of a data source and is beneficial to a corporation since it reveals trends, patterns, and linkages that might otherwise go undetected.
The dataset was previously available in four different text files, which were then concatenated for the purpose of training the model.
Then, while exploring the data, some NLP tasks were completed.
Data was cleaned up by eliminating links, numerals, @, stop words, and converting everything to lower case.
The sentences were then converted into vectors using Count vectorizer and Tfidf Vectorizer.
Count vectorizer: It is used to transform a given text into a vector on the basis of the frequency (count) of each word that occurs in the entire text.
Tfidf vectorizer: It uses an in-memory vocabulary (a python dict) to map the most frequent words to features indices and hence compute a word occurrence frequency (sparse) matrix.

MODEL IMPLEMENTATION:

1. STATISTICAL APPROACH:

Three different approaches were implemented for developing the pure statistical models such as

- Ordinary Least Square
- Generalised Linear Model
- Robust Linear Model

Ordinary Least Square:

Ordinary Least Squares regression (OLS) is used to predict values of a continuous response variable using one or more explanatory variables and can also identify the strength of the relationships between these variables (these two goals of regression are often referred to as prediction and explanation).

Generalised Linear Model:

The general linear model and the generalized linear model (GLM) are two commonly used families of statistical methods to relate some number of continuous and/or categorical predictors to a single outcome variable.

Robust Linear Model:

It is designed to overcome some limitations of traditional parametric and non-parametric methods. Regression analysis seeks to find the relationship between one or more independent variables and a dependent variable

2. DEEP LEARNING APPROACH:

Artificial Neural Networks:

The term "Artificial Neural Network" is derived from Biological neural networks that develop the structure of a human brain. Similar to the human brain that has neurons interconnected to one another, artificial neural networks also have neurons that are interconnected to one another in various layers of the networks. These neurons are known as nodes

Each node, or artificial neuron, connects to another and has an associated weight and threshold. If the output of any individual node is above the specified threshold value, that node is activated, sending data to the next layer of the network.

3. MACHINE LEARNING APPROACH:

This study used a variety of machine learning models, including Linear Regression, Ridge, K Nearest Neighbours, Decision Tree Regressor, and Support Vector Regressor.

All these models were implemented on the data which has been converted into vectors using count vectorizer and tfidf vectorizer.

INTERPRETATION:

Statistical model:

After the implementation of these models, metrics such as MAE, MSE, RMSE were used to determine the efficiency of the model. It has been found that all the models implemented were of similar efficiency. So, we have randomly chose generalised linear model incorporated with count vectorized data as the best model.

Deep learning model:

The ANN model incorporated over the tfidf vectorized data found to be working better with lesser errors. so, it is considered to be the best model.

Machine learning model:

Out of all the models used in the ML approach, support vector regressor found to be a better model incorporated with the tfidf vectorized data.

CONCLUSION:

Finally, we have found the best model on each of the three different approaches. So, we have used these 3 models to find & predict intensity for tweets in the testing data.