AlmaBetter

CORONAVIRUS TWEET SENTIMENT ANALYSIS

CAPSTONE PROJECT - III

MADE BY: SANKET CHOURIYA

Technical Document

ABSTRACT

In this project, I focused on the sentiment analysis of English tweets, contained in the covid-19-nlp-text classification dataset, using several classification methods, from the simplest ones to the most complex.

INTRODUCTION

Nowadays, we live in a world where billions millions of terabytes of text data are produced every day. Thanks to recent studies we are able to analyze them in several ways and to extract from them knowledge about the world around us. In this project, I used different machine learning techniques to classify some Twitter posts about the recent Covid-19 pandemic, considering the sentiment that they imply. In the first part of the project, the classification considers five different sentiment labels (Positive, Negative, Extremely Positive, Extremely Negative, Neutral), while, in the second part, considers just three different labels.

The methods used are: Support Vector Machine (SVM), AdaBoost with Decision Tree, Multinomial Naïve Bayes Classifiers (MNBC), Multinomial Logistic Regression and SGD Classifier. The results obtained with these techniques are shown in this report.

DATA UNDERSTANDING

In Coronavirus tweets sentiment analysis, the data set contain 41157 row and is composed by 6 columns:

- UserName and ScreenName, which refer to the Twitter user who created the tweet;

- Location and TweetAt, which describe respectively the state or city from which the tweet was posted and the date of posting;

- OriginalTweet, which contains the text of the tweet;

Out of 41157 tweets, there are 11422 Positive tweets, 9917 Negative Tweets, Neutral Tweets, 6624 Extremely Positive and 5481 Extremely Negative. This are the original 5 classes in the Target label aka Sentiment Column in Dataframe. As this classification is MultiClass. Thus, to make it simple, the original classes has been merged down to 3 classes now i.e. Positive (18046), Negative (15398) and Neutral (7713) to make the classification much easier.

DATA CLEANING

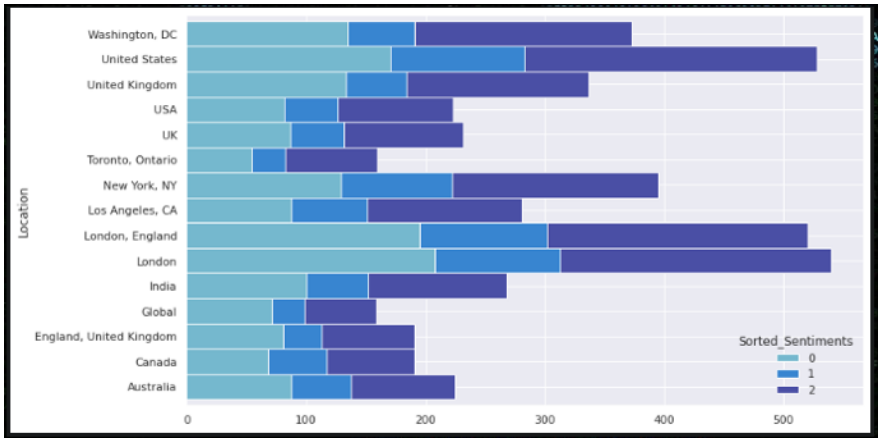
Firstly, I defined a function to tokenize the tweets using tools by using Neattext and RegExp libraries. Following are the necessary steps taken to clean data:
- Substitute all space characters with a single space
- Remove URLs, mentions and hashtags
- Remove Non-ASCII characters
- Lemmatize, Stemming and lower case
- Remove different language stopwords and punctuations
- Auto correcting misspelled words

Subsequently, in order to proceed with the classification, I used the **TfidfVectorizer**, with as tokenizer, the Neattext function previously described, to generate the feature space and the vectors for every single token in that space. Then, it was performed the feature selection using the Chi2 method and the vectors were weighted with the TfidfTransformer.

EXPLORATORY DATA ANALSIS

is an approach of analyzing data sets to summarize their main characteristics, often using statistical graphics and other data visualization methods. The data is reach in information thus it`s logical to dive deep into it.

LOCATIONS WITH MOST OF THE TWEETS

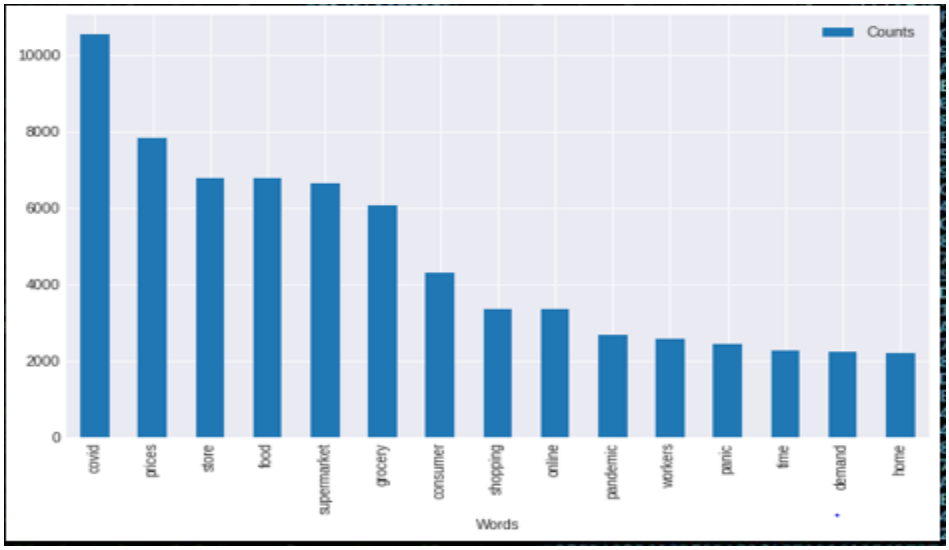


Seems, like London had most no of tweets made during the 2020. Both Most of number of Negative and positive tweets has been made from London itself while Americans were on second place for their presence in Twitter.

MOST USED HASH TAGS AND MENTIONS

Treemap charts visualize hierarchical data using nested rectangles. The input data format is the same as for Sunburst Charts and Icicle Charts: the hierarchy is defined by labels (names for px.treemap) and parents attributes.

MOST USED WORDS IN TWEETS



It's not a surprise that covid is the most common word used in the tweets. It made more than 12000 appearances in tweets globally. After then that words like, prices, store, food, supermarket and grocery were most used words.

Three-Labels Classification:

SVC, AdaBoost with Decision Tree, SGD Classifier, M-Naïve Bayes Classifier and M-Logistic Regression

Multiclass classification:

Whether it's spelled multi-class or multiclass, the science is the same. Multiclass classification is a machine learning classification task that consists of more than two classes, or outputs.

We are given a set of training samples separated into K distinct classes (In this dataset it is POSITIVE, NEGATIVE & NEUTRAL sentiments) and we create an ML model to forecast which of those classes some previously unknown data belongs to. The model learns patterns specific to each class from the training dataset and utilizes those patterns to forecast the classification of future data.

Some of the most popular algorithms for multi-class classifications that I have used in this project:
- Multinomial Naive Bayes Classifier
- Stochastic Gradient Descent - SGD Classifier
- SVM (Support Vector Machine)
- AdaBoosted Decision Trees
- Multinomial Logistic Regression

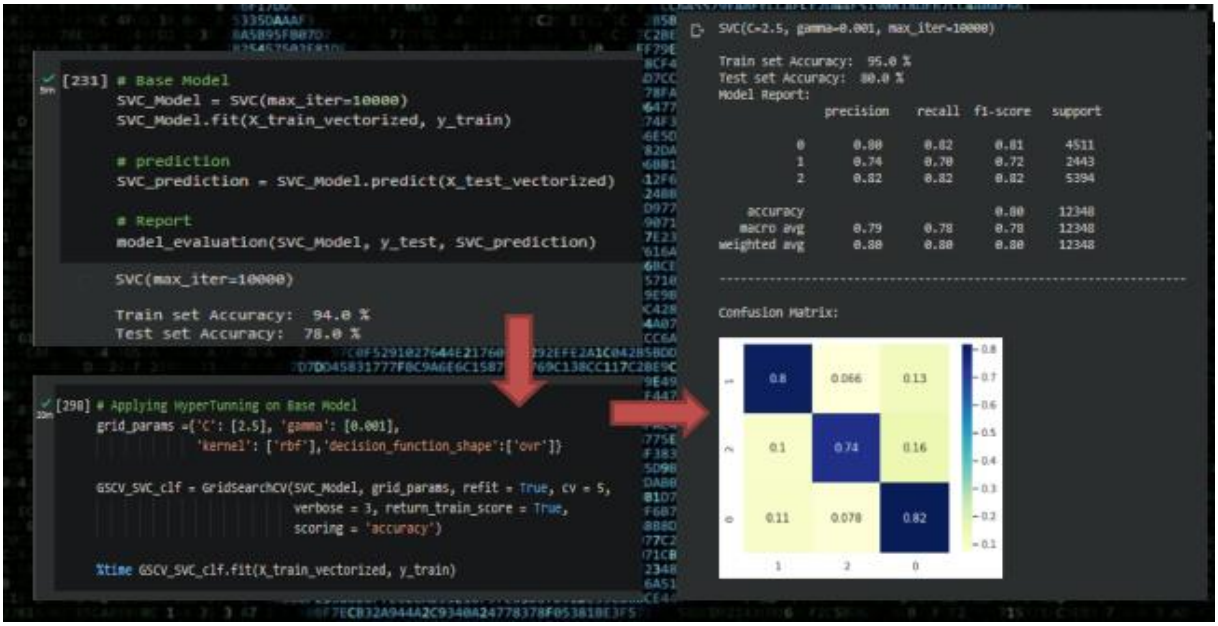
MODEL COMPARISON

	Model	Train_Accuracy	Test_Accuracy	F1_Score
0	Naive Bayes Classifier	83.0	65.0	[0.684, 0.495, 0.69]
1	Stochastic Gradient Descent	96.0	74.0	[0.765, 0.63, 0.777]
2	Support Vector Machine	94.0	78.0	[0.803, 0.687, 0.809]
3	AdaBoosted Decision Trees	84.0	78.0	[0.789, 0.729, 0.793]
4	Logistic Regression	97.0	71.0	[0.739, 0.578, 0.748]

Support Vector Machine and AdaBoost Classifier performed better than the other three classifiers, with Support Vector Machine having a slight advantage with a mean F1 score of around 0.77 combined while overall model accuracy is 78%

HYPER PARAMETER TUNING

Support Vector Classifier model came out to the best base model. Thus, I decided to perform hyper-tuning. I used it in two different ways, changing the values of some parameters. In the first case, the classification was done setting the number of iterations at 10000 and using as C value of the SVC the default one (C= 1.0). After that, I applied the **GridSearchCV** method on the train set to find the gamma at 0.001 and the best value for the parameter C among 0.01, 0.1, 1, 10, 100. A 5 fold cross-validation was performed on the train set to discover the best values. The best parameters chosen by the optimization method were at gamma **0.001**, kernel as '**rbf**; Decision_function_shape as '**OVR'** and **2.5** for C value.



CONCLUTION

- Speaking of locations, most of the tweets has been made anonymously or without a location share, which contribute around 21% of the total tweets made globally. On the second place, It is London and London, England both contributed around 2.5% of total tweets made globally.
- #coronavirus and other versions of it were the most trending hashtags during the timeline.
- @realDonaldTrump and @Tesco were the most tagged and active users on the twitter.
- It's not a surprise that covid is the most common word used in the tweets. It made more than 12000 appearances in tweets globally. After that mostly, Food and Service related words were mostly used.
- Initially, 5 target sentiments were given for the classification later then converted into 3-Class targets i.e. POSITIVE, NEGATIVE and NEUTRAL which made Sentiment Classification a little easier.
- For vectorization, TF-IDF vectorizer has been used, which Convert a collection of raw documents to a matrix of TF-IDF features.
- For a Multiclass Classification, SVC Base model scored 77% Overall accuracy. After which, Increase to 3% when hyper-tuned with GridSearchCV.