Interim Report of Twitter Data Analysis

Abstract

We examine the sentiment analysis of twitter data. In this project we are addressing the problem of sentiment analysis in twitter; where we are classifying the tweets that we have to a positive, negative or a neutral expression. This we are able to achieve based on the use of certain keywords.

Twitter is a growing online platform for social media where people express their thought on various topics. These expressions are usually in texts that are not more than 140 characters. This platform allows one to send and receive short post that are called tweets.

The information that is disseminated in twitter are not always entitled to be read. Twitter services is rapidly expanding and it has a user of close to 200 million who are registered. Out of this big number, almost half of them are active users of twitter, who daily log into twitter and can generate up to 250 million tweets per day. Due to this big number, we are hoping to achieve a reflection of what the public says, based on the sentiments that the public has on different topics.

Introduction

Social media has a way in which the sentiments from the public can be useful in the society. It is the purpose of this data analysis, by the sentiment analysis that we are carrying out to identify, by the use of the models that are going to be applied to make sense in the twitter data analysis. We are going to use the different methods that are used in the development of models in the carrying out of the model creation. Twitter data allows one to track the brand and also to identify the customers that are able to contribute to the development of a brand or

a product. The focus of this analysis is not only to get the sentiment analysis of covid19 in Africa but as well to get the sentiment analysis of the different countries that are in Africa. We are intending that at the end of this data analysis we are able to classify, and to present out data in a manner that one is able to get the most used words. This is achieved by a proper visualization technique. We are also going to clean our data and preprocess it for the creation of a model that is able to classify the data in the appropriate class that it falls.

In the process of this analysis, it is very important to set up the MLOps pipeline so as to identify the different steps up to the deployment stage of the data analysis. We are able to understand the scope of the data and to come up with a technique of solving our problem by creating a procedurised way in which we are to do this. We began by getting the data that we are going to use in the analysis. This is gotten from tweeter, and since the data that we get is not all that clean, we have to first clean and preprocess the data by the removing the null values that are not very important. Its also proper to note that the data from the twitter is also not easy to get, thus it needs an API and a web scrapping knowledge to do so.

Understanding the data.

After acquiring the data, we needed to understand what we can do with the data. This process of understanding what the data is, what it entails, and if it suits our need for the purpose of this study. We are able to do this by first getting to know the attributes of the data and looking at the values that the data contains. This is also important in handling of the data, so that we are able to know what we are going to need in the data and what is appropriate for us to do without. The data presented in the JSON format (JavaScript Object Notation), contained a lot of jumbled data. To get to understand the data, we first cleaned and organized the data. This process of cleaning the data involved the use of packages like pandas. We removed the

unwanted columns from the data. The unwanted columns that were in the data resulted from the collection of the data. After this we also removed the duplicated data that are in the data, and this was targeted on the rows of the data.

We also were able to find the time to which the tweets were collected. This was done by conversion to date format. By this we are able to classify the data to the time stands that the data was collected.

The data that we are going to use in this analysis is a data that is availed via a json file, but collected in a tweeter. The data was collected by the use of the APIs and thus, the process of the scrapping of the data and the preprocessing stage of the data was very useful as to ensure that the data we collected was the required data that was needed for the creation of the model. The data was as well provided in two forms, the first data was one that included the Africa alone, while the second data is the data that involves the individual countries.

The data that we are dealing with had a total row of 6533 and there are a total of 13 columns. This data that was generated after the cleaning of the JSON file from the preprocessing process.

Method

In our analysis we were able to implement the machine learning algorithms that involve the use of scikit learn to create a model that was used in the classification of the data that we have to tell the if the tweet was a positive, negative or a neutral. This was consequently applied in the polarity score. There after we were able to measure the score of accuracy of the model that we had. This score gave an accuracy of 98%. This enabled us to see that the model that we had created was able to function in the manner that we needed, and we could see that the accuracy was dependable on.

In the deployment of the data, by consistently pushing the data to GitHub, we created a Travis CI which worked out efficiently in the process to be able to deploy the model. The continuous integration using Travis CI enabled us to have a look at the integration of the model as we proceeded on with the development and the deployment. After this we were able to install streamlit to ensure that our code is able to display successfully in a web browser. Implementation details e.g. Travis CI, unit testing coverage, Streamlit for dashboard, etc.

Result and Discussion

From the dashboard that we have we truly realize that the data can be explored in the manner that we need it.

The challenge you faced

During the analysis of this data, there are challenges that were faces in the course of the analysis. The data was not connecting to MySQL and thus there was a lot of errors in the process of the analysis.