Big Data Project Report
Machine Learning with Spark

Team Members: **Date** : 06-12-2021

Abhay D A (PES1UG19CS011)
Akash S (PES1UG19CS042)
Akshay S P (PES1UG19CS046)
Akshay S D (PES1UG19CS045)

Project Title : Sentiment Analysis of twitter data.

Design details:

We have split up our code into 4 main functions which does the following:

1) Collects the data from the TCP socket stream and converts the data into a pyspark dataframe.
2) Preprocesses the tweets.
3) Trains the algorithm.
4) Tests the algorithm

Surface level implementation details about each unit:

1)

- In the very first step we take the incoming data from the TCP socket and convert it to json format using the json.loads() method.
- Then we convert the json data into a pyspark dataframe using the createDataFrameMethod.
- We do this for all the batches streamed during the batch interval.
- We then take this dataframe and pass it as a parameter to another function which preprocesses the data.

2)

- In this step we take the data from the previous step and perform preprocessing on the data.
- Here we take each row and perform the following preprocessing methods on it:

Removing all URLs.
Removing all hashtags.
Removing all special characters
Removing all user mentions.
Converting the data to lowercase.
Removing consecutive repeated characters in a given word.
Removing single characters.
Removing more than one white space between words.

- We then return this transformed dataframe.

3)

- In this function we train the models.
- First we split the dataset into X_train and Y_train data and then convert the data into a numpy array.
- We then perform tokenization of the words since the ML algorithm cannot understand text data. We are using incremental tokenization of data using the partial_fit() method of the HashingVectorizer. We are performing this using pickle files.
- We next use this tokenized data to train the ML models. Just like before we use the partial_fit method of the various classification and clustering algorithms. We are using pickle files to store the trained models.
- For every batch of data we train the algorithm on the current batch and dump the trained model into a pickle file. When another batch of data arrives we load this model and retrain the model on the new data without forgetting the past learnings.

4)

- Here we load the trained model from the pickle files and make predictions based on the test data.
- We then get the predictions and analyse the various metrics of the model such as accuracy , confusion matrix, etc.

Reason behind design decisions:

We thought it was best to structure our program this way as this structure efficiently splits the different functionality of the program. This helps us in debugging and finding out where the error was caused.

Take Away from the project :

This project really gave us a thorough understanding of how to combine both machine learning and spark streaming to make the best use of both worlds. This project also taught us how these technologies are combined in the real world and really helped us understand all the concepts thoroughly.
