PES UNIVERSITY

BIG DATA PROJECT

UE19CS322

TEAM ID: BD2_141_175_403_425

1. Project Title Chosen

SSML- Machine Learning with Spark Streaming

We decided to implement Machine learning models using Spark streaming on the "Sentimental analysis" dataset- which has a column with tweets of users and a column with 0 or 4- a classification of that tweet as negative or positive respectively.

2. Design Details

Streaming: through socket object and conversion to json string, in turn to json dictionary, to spark data frame.

We have a main.py function that contains all our code.

3. Surface Level Implementation

Fetching and pre-processing Data:

We start with making a connection to local host through port 6100. A socket object is created for this. The streamed data is received as a json string. Since the socket is for transport layer communication- we decode the binary form that it is in.

Then this json string is converted to a json dictionary using "loads". We then convert it to a spark data frame.

For pre-processing, we used a python function made by us to remove hashtags, URLs, '@', and digits. We also used a tokenizer, stop words remover in our pre-processing.

Building The Model:

We implemented 3 models- logistic regression, Naiive bayes classifier and random forest to classify the data. Logistic regression is very effective on text data, plus it is very easy to implement and interpret. It's very good to solve a binary classification model (in our case, 0s and 4s).

For Naiive Bayes, the conditional probabilities are easy to evaluate. This classifier's main application is text classification due to how it considers words and the number of times they are present.

We used Random Forest especially because they are good at dealing with high dimensional noisy data in text classification. It comprises a set of decision trees- each which is trained with random subsets of features.

4. Takeaway From The Project

A major takeaway from this project was how large twitter's database can be and how data is generated at rapid rates at all times. We got to learn how to handle large streams of data, plus how to apply Machine Learning tasks on these to come up with conclusions and comparisons. It also offers quick recovery from any failures. Also, we learnt how to combine streaming data with interactive queries.