BIG DATA
PROJECT REPORT

Date: 06/12/2021

TEAM-MEMBERS:
Jaaswin D Kotian PES2UG19CS156
Chintamani Bhat PES2UG19CS099
Chandrahas L G PES2UG19CS096
Hemanth Hegde PES2UG19CS148

Spark Streaming for Machine Learning-SSML
Dataset Chosen: Sentiment

Design Details:

Importing libraries: We start by importing libraries like sparkcontext, streaming context, spark session, reg exp, and so on.

Streaming: Data streaming refers to the process of obtaining data from a csv file and transferring it to our working environment for preparation.

Preprocessing: We clean the model here to get nice clean data for forecasting results by removing undesirable words, emojis, and spaces. Emoji remover, stopword removal, tokenizing, steaming, and hashing are some of the approaches we utilise.

Model construction: To improve the accuracy of the results, we use different types of models such as MultinomialNB, Preceptron, and Multinomial, and we use the results to get: **accuracy** (we add the results of each model cumulatively to increase the accuracy without overfitting or underfitting the results), **graph** (we plot the graph using the output of the models to compare the accuracy of each model), **clustering** (we split the data (tweets) into two parts as positive and negative so that we can compare).

Surface Level Implementation:

Importing lib: We import required lib for processing of data.

Streaming: We stream data to get the tweets for processing. We get data from a csv file and transmitting it to our working environment for preparation.

Preprocessing: To process and achieve the most accurate results, we clean the data/tweets.

To begin, we gather the dataframe and remove any undesirable noise such as emoji and stopwords. The data frame is then tokenized, and stopwords that aren't needed for model prediction are removed. The dataset is then stemmed to acquire the cleaned data.

Feature extraction: The pre-processed data is transformed into vector form, and the features are sent to the model to be trained.

Model construction: in order to achieve high accuracy, we select specific models that are best suited to our data.

We used SKLearn's built-in models: Multinomial Nave Bayes, Passive Aggressive Classifier, and Perceptron model, which uses features as parameters to train.

Testing: To test the models and assess their respective accuracies, we use the test dataset.

Plotting Graphs:

For each model, plotted accuracy vs. batch size.

For each model, plotted the average accuracy.

For each model, plotted total accuracy, recall, and precision.

Reason behind design decisions:

Importing libraries: We used the pyspark library since it allows us to develop Spark apps with Python APIs.

Streaming: Data streams have been employed because they allow us to extract and process data in real-time.

Preprocessing: We preprocessed the dataset to convert the raw data into a clean data set because analysing data in raw format is impossible.

Model construction: We developed a model that can predict if a tweet is sentimentally pleasant or negative.

Graph: We utilised graphs to graphically show the model's results.

Clustering: Used to classify data into structures that are simple to comprehend and manipulate.

Takeaway from project:

We will be able to estimate the sentiments/mood of the user who has tweeted as a result of this project; we will be able to stream the data and process it to obtain the feelings.

This project may also be seen from a commercial standpoint in order to enhance consumer experience.