UE19CS322 - BIG DATA
PROJECT REPORT

Team Id : BD_092_118_166_289

Aryan Kumar Jha **-** PES1UG19CS092
Vinayak CS -PES1UG19CS118
Gauthami S - PES1UG19CS166
Nandana C - PES1UG19CS289

Project Title : Spark Streaming for Machine Learning - Sentiment Analysis

Design details

The original dataset has two features: 'tweet' and 'sentiment'.
streamtrain.py and streamtest.py are used for streaming train and test dataset respectively.

Classification :

pipeline.py : Preprocesses each batch of data and trains the models
testcode.py : Preprocesses each batch of data. Loads the trained model and predicts the output. Calculates accuracy score of each model

Sentiment : Contains train.csv and test.csv

Surface level implementation details about each unit

Classification

Three models have been implemented using the sklearn library - Naive Bayes Multinomial, SGD Classifier and Passive Aggressive Classifier

1. **Naive Bayes Multinomia**l : The multinomial Naive Bayes classifier is **suitable for classification with discrete features** . The multinomial distribution normally requires integer feature counts.

2. **SGD Classifier** : The class SGDClassifier implements a plain stochastic gradient descent learning routine which supports different loss functions and penalties for

classification. Below is the decision boundary of a SGDClassifier trained with the hinge loss, equivalent to a linear SVM.

3. **Passive Aggressive Classifier** : Passive Aggressive Classifier belongs to the category of online learning algorithms in machine learning. It works **by responding as passive for correct classifications and responding as aggressive for** any miscalculation.

Each batch of data undergoes preprocessing steps for both training and testing data.

- Remove special characters/punctuations if any
- Tokenization is the process of taking text and breaking it into individual words.
- Removal of stopwords
- TfidfVectorizer: Maps the most frequent words to features indices and hence compute a word occurrence frequency (sparse) matrix

Training : After each batch has been preprocessed, it has been fed into each machine learning classifier model. The pickle module is used to save and load the machine learning models. Each model uses the method partial_fit to perform one epoch of stochastic gradient descent on given samples.

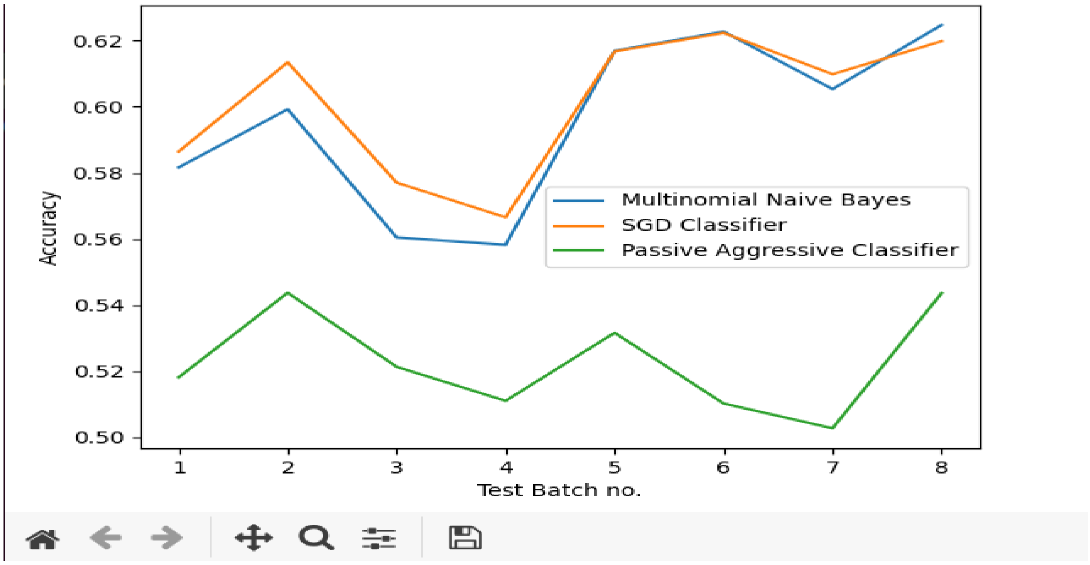
Testing : Preprocessing steps are executed for each batch. The trained models are loaded in the testcode.py file. The models are now used to predict the target. Accuracy score of each model is calculated accordingly.

Clustering

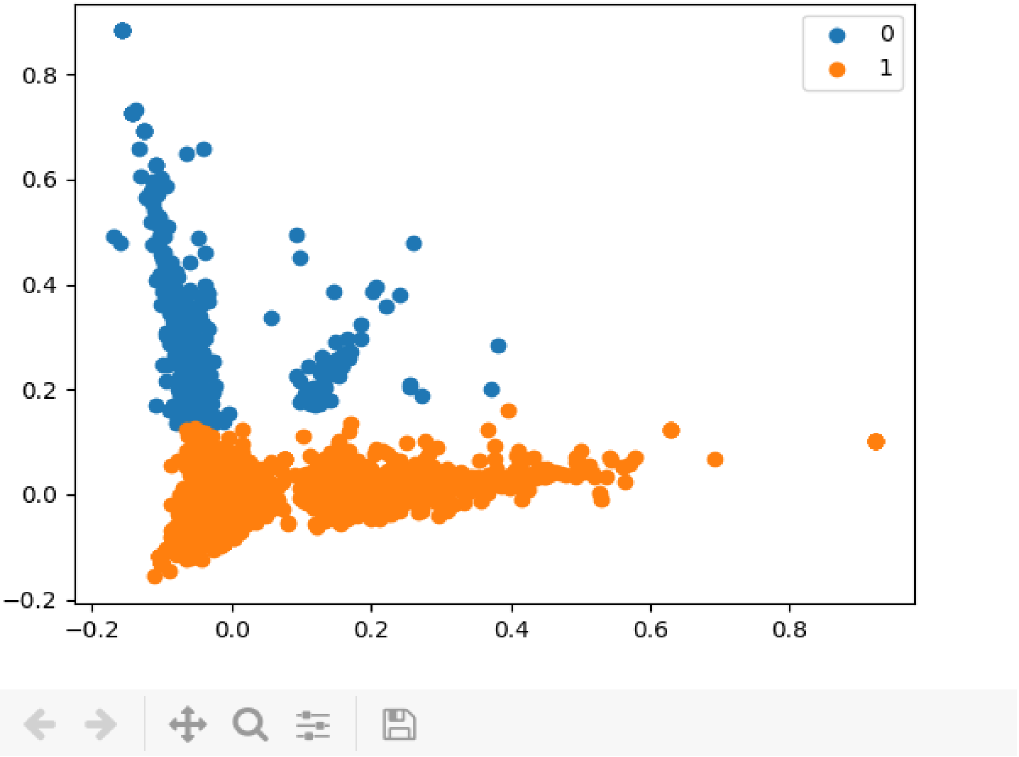
We cluster to find groups in the data, with the number of groups represented by the variable K.

We've implemented the sklearn.cluster.MiniBatchKMeans model for clustering. n_clusters have been initialised to 2. partial_fit is used to update k means estimate on a single mini-batch. The model is saved and loaded using pickle.

Classifiers accuracy score comparison



K means clustering



Reason behind design decisions

Multinomial Naive Bayes model is generally used where there are discrete features. It generally works with the integer counts which are generated as frequency for each word. All features follow multinomial distribution. In such cases TF-IDF(Term Frequency, Inverse Document Frequency) also works.
SGDClassifier model and PassiveAggressiveClassifier have been implemented to compare the performance of each model with the selected dataset.
To evaluate the performance of a classification model such as the one that we just trained, we can use metrics such as the confusion matrix, F1 measure, recall score and the accuracy.

K Means: The goal of this algorithm is to find groups in the data, whereas the no. of groups is represented by the variable K. The data have been clustered on the basis of high similarity points together and low similarity points in the separate clusters.

Take Away from the project

Planning the order in which components are built and managing time is critical. We learned about working with streaming and pyspark and implementing incremental learning. Also, how refining understanding through discussion and explanation is crucial at every step.