BIG DATA PROJECT
SENTIMENT ANALYSIS

Name	SRN
Bhuvan Kumar	**PES1UG19CS113**
Nishanth M	**PES1UG19CS303**
Prajwal Kamath K	**PES1UG19CS337**
Prerana Hadadi	**PES1UG19CS352**

1.Project title chosen

Sentiment Analysis

1.1 Design Details

- We have used the concept of global schema to convert the streaming data to data frame
- We have used certain processing techniques to improve the quality of the dataset .
- We built models to classify the Tweets
- For testing our models, we have used metrics such as F1 score , accuracy , precision , recall and confusion matrix .
- We have used the k means -mini batch algorithm on our dataset to analyze the clusters .

1.2 Surface level implementation

- We created global schema using the StructType , StructField and for each batch of streaming data, we converted to RDD and later to data frame .
- For pre-processing we have used the techniques such as Tokenizer , Stop words , count Vectorizer and TF-IDF to filter out the unwanted data .
- Incremental Learning models such as SGD , PAC and Multinomial are build and the concept of partial fit has been used.

- We followed the same techniques used in streaming our file and predicted the metrics
- We have used k means mini batch algorithm and we plot the graph for each iteration of the batch

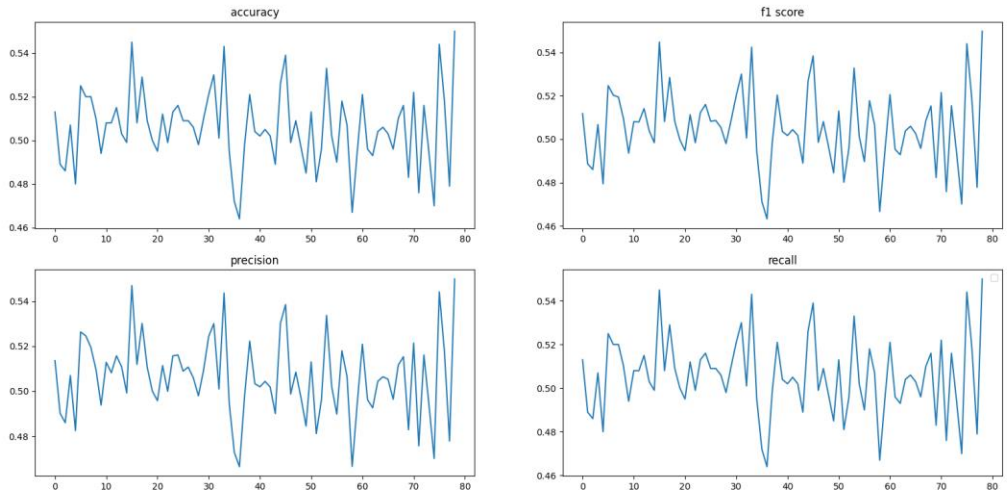
1.3 Reason behind design decisions

- We choose tokenizer as helps in interpreting the meaning of the text by analyzing the sequence of the words.
- Stop words are commonly eliminated from many text processing applications because these words can be distracting, non-informative (or non-discriminative) and are additional memory overhead.
- We choose TF-IDF as it measures the importance in a particular document .
- We choose multinomial as our 1st classifier as it's easy to implement and supports incremental learning and its scalable and easy to handle large dataset**.**
- We choose SGD as our 2nd classifier as its It is easier to fit into memory due to a single training sample being processed by the network. It is computationally fast as only one sample is processed at a time.
- We choose PAC as our 3rd classifier as it works by responding as passive for correct classifications and responding as aggressive for any miscalculation
- The advantage of using K means - mini batch is to reduce the computational cost by not using all the dataset each iteration but a subsample of a fixed size.

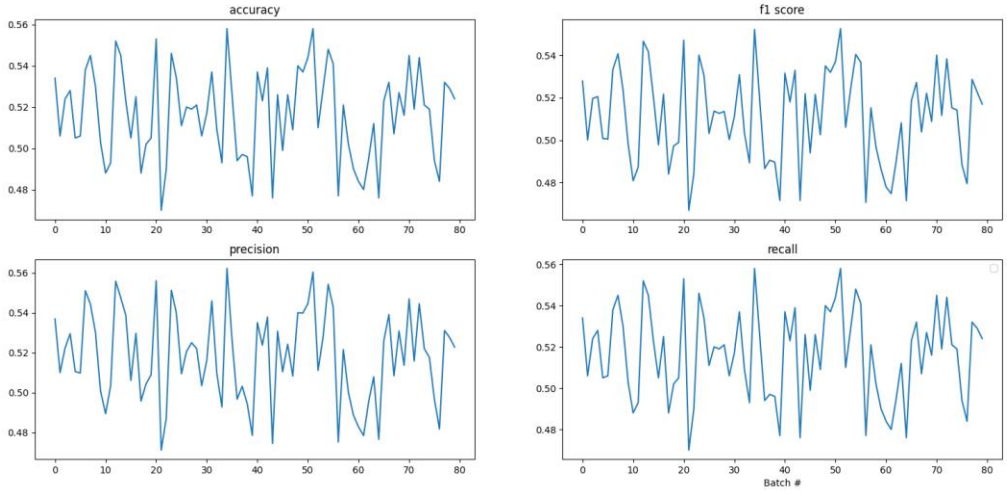
1.4 Takeaway from the project

- We learnt how to train models and predict accuracy , F1 score , computation matrix and recall.
- We learnt how to classify tweets into positive and negative category using SVG, PAC and multinomial as the baseline and to predict better results .
- We learnt certain concepts of scikit learn , and certain pre-processing techniques and how they are helpful in building the accuracy .
- We learnt to construct plots given different models and compare and analyze between them.

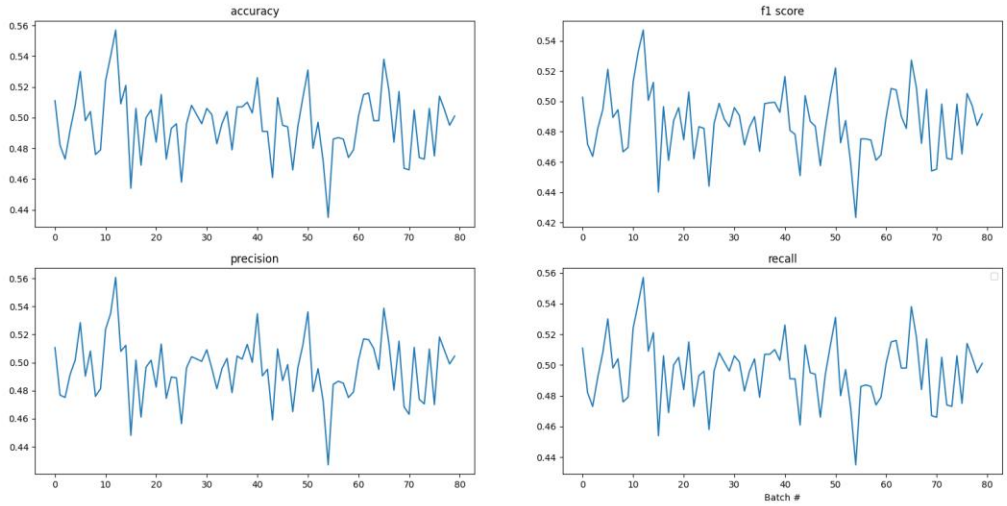
Accuracy plot for 3 different Models



Multinomial



Passive Aggressive Classifier



SGD