SENTIMENT PREDICTIONS FOR TWITTER POSTS

November 9, 2020

Michael Ortiz
VP
Alert! Analytics

Re: Sentiment Predictions for Twitter posts

Dear Mr. Ortiz,

Project background

Alert!Analytics' client, Helio, has requested to build predictive models to detect positive or negative sentiment for Twitter posts using Apache Spark clusters in Microsoft Azure Databricks. The main goal is to confirm that a content of a tweet can be classified as having a certain level of sentiment.

Data characteristics

The data, to be used for model training and testing, represents a large dataset with 1.6 million observations. The data attributes mainly consist of Twitter posts, users' names, post date and time, and manually determined sentiment for each post.

The sentiment includes 2 categories: "0" – negative and "4" – positive.

The dataset has the equal amount of observations for each sentiment, and it is considered balanced (see Figure 1.1).

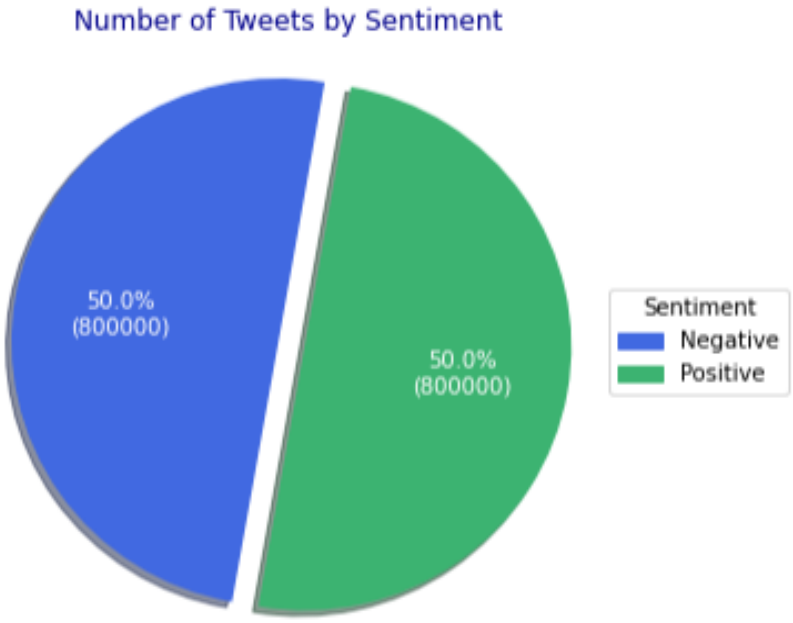


Figure 1.1

SENTIMENT PREDICTIONS FOR TWITTER POSTS

The data was collected between April 6, 2009 and June 25, 2009 (see Figure 1.2).

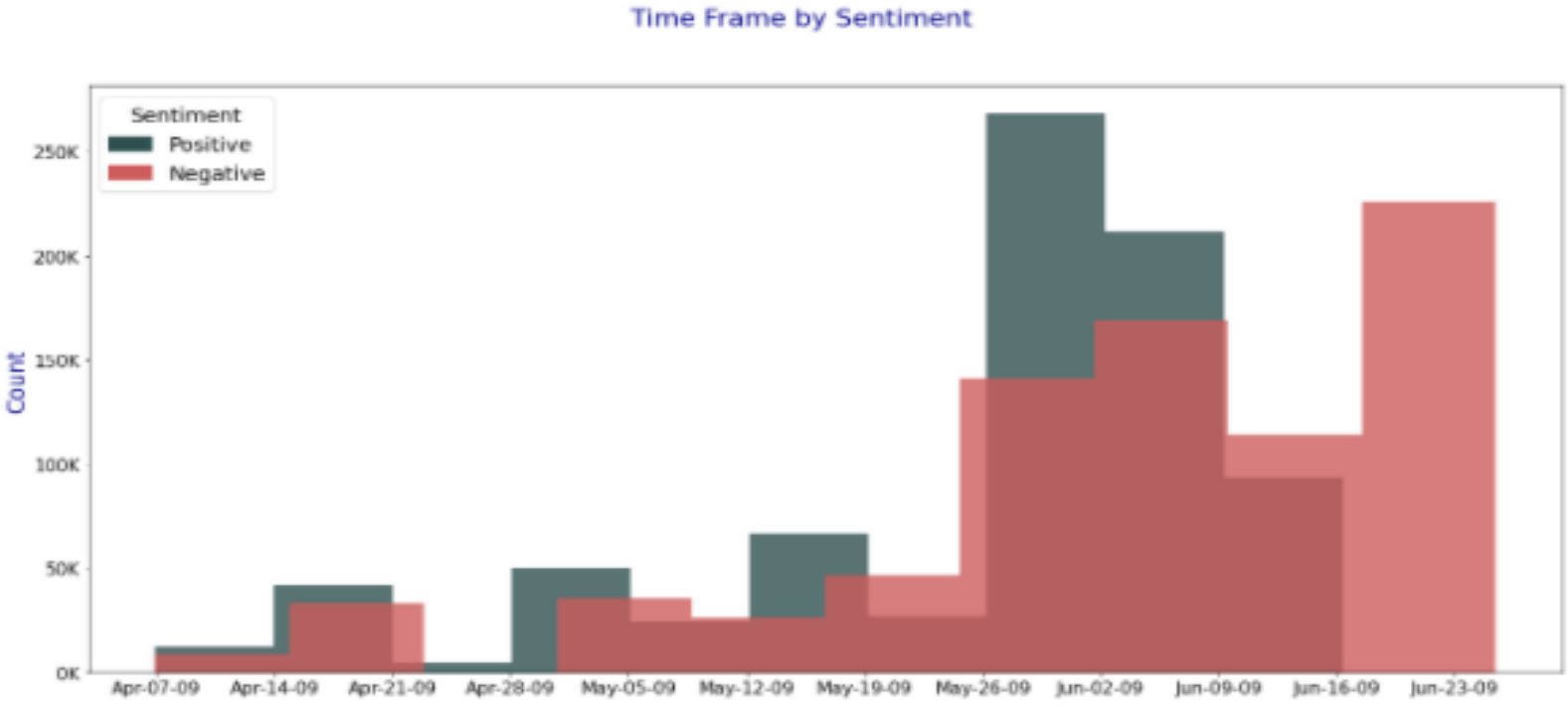


Figure 1.2

Top 5 Twitter users with negative sentiment tweeted more often than 5 Top Twitter users with positive sentiment during that time (see Figure 1.3).

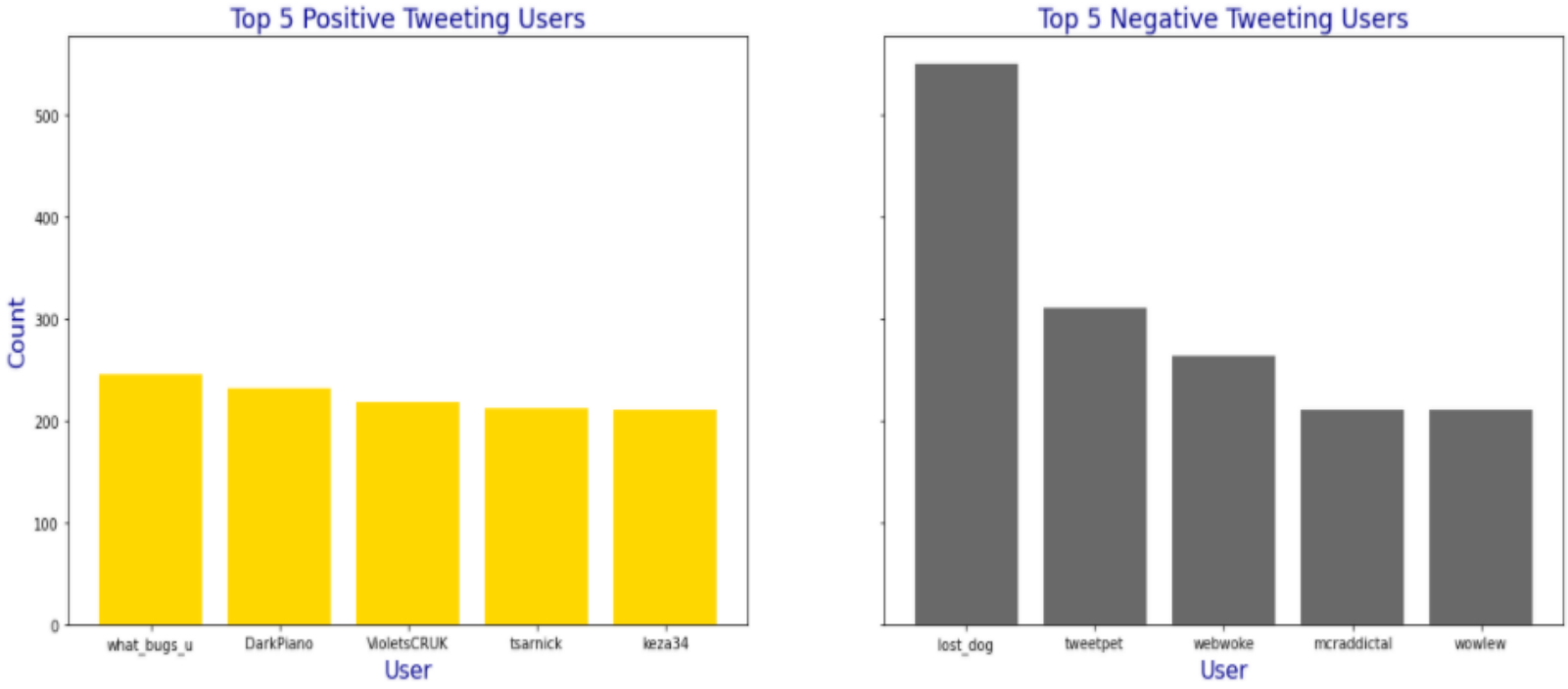


Figure 1.3

SENTIMENT PREDICTIONS FOR TWITTER POSTS

The average length of 5 Top tweets with positive sentiment is almost twice longer than the average length of 5 Top tweets with negative sentiment (see Figure 1.4).

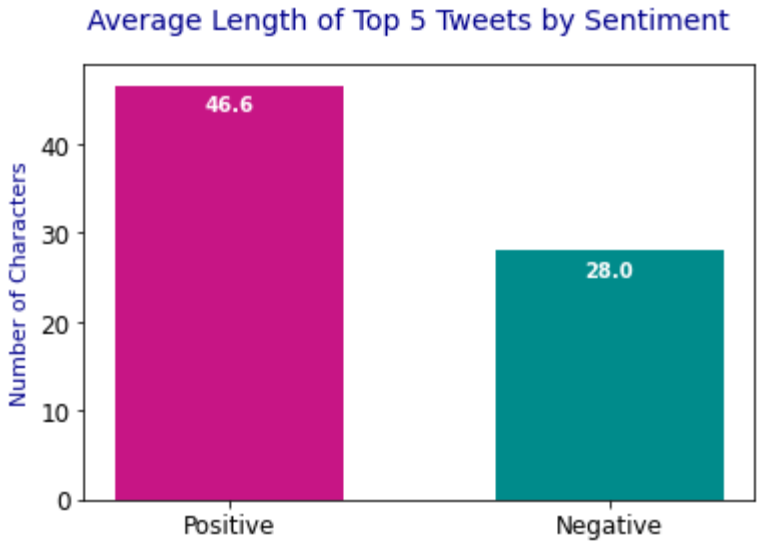


Figure 1.4

Pre-processing

Twitter posts were separated into words. Stop words have been removed. A matrix with a word count has been created. Words with less importance have been assigned smaller weight.

Model evaluation

Three classification models have been deployed. The performance metrics for each model is listed below.

Decision Tree

	precision	recall	f1-score	support
0	0.07	0.89	0.13	12684
4	0.99	0.52	0.68	307316
accuracy			0.53	320000
macro avg	0.53	0.70	0.40	320000
weighted avg	0.95	0.53	0.66	320000

SENTIMENT PREDICTIONS FOR TWITTER POSTS

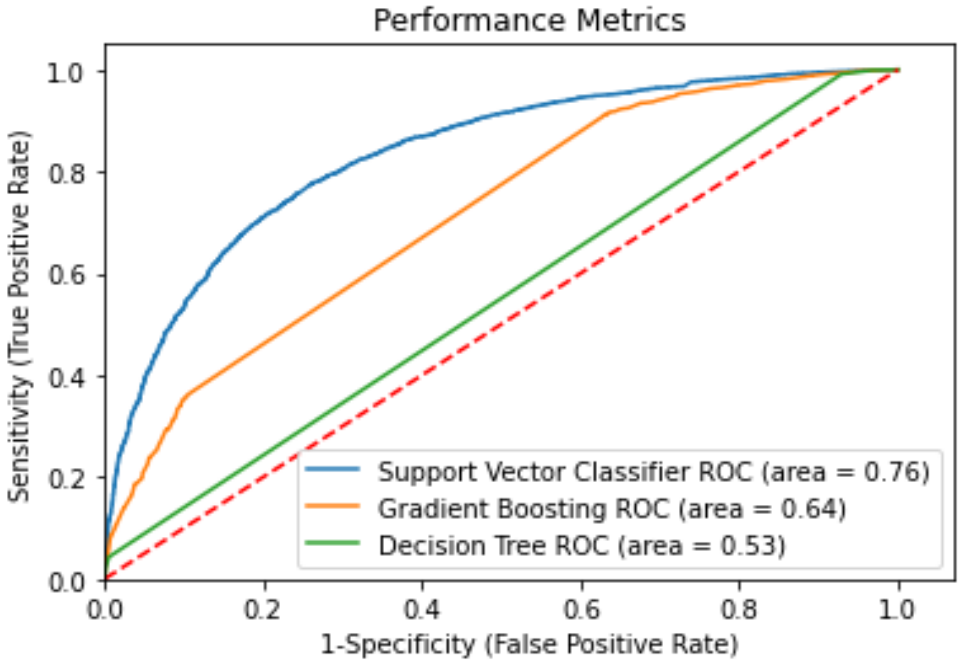
Gradient Boosting

	precision	recall	f1-score	support
0	0.37	0.80	0.50	72979
4	0.91	0.59	0.72	247021
micro avg	0.64	0.64	0.64	320000
macro avg	0.64	0.70	0.61	320000
weighted avg	0.79	0.64	0.67	320000

Support Vector Classifier

	precision	recall	f1-score	support
0	0.55	0.79	0.65	110756
4	0.86	0.65	0.74	209244
accuracy			0.70	320000
macro avg	0.70	0.72	0.69	320000
weighted avg	0.75	0.70	0.71	320000

The chart, comparing the performance metrics of three models, shows that Support Vector Classifier has the best performance.



SENTIMENT PREDICTIONS FOR TWITTER POSTS

Summary

Our findings indicate that a content of a tweet can be classified as having either positive or negative sentiment. The models can be tuned further to increase the accuracy.

Lessons learned

This project involved the use of Apache Spark clusters in Microsoft Azure Databricks. There was a problem uploading the dataset to the Databricks File System (DBFS) via User Interface (UI) due to its size. The support team recommended to use either Databricks CLI or Databricks Utilities. Errors were encountered during the upload via Databricks CLI. Therefore, the original dataset was fragmented into smaller files and successfully uploaded into the DBFS. Clusters tended to freeze while running large tasks. The KNN algorithm on the large dataset was processed within 25 seconds, while Gradient Boosting Classifier took 8 minutes.

In the prior assignment Amazon Web Services EMR was used to collect counts of words with certain sentiment for several types of smartphones from randomly selected web pages using Apache Hadoop clusters. Amazon S3 was used to store the files. Clusters with more than 50 steps tended to freeze. As a result, the smaller clusters were used.

Despite the encountered problems, the information gained from the cloud-based data processing cannot be obtained from local machines.

One of the reasons of slow processing could be the poor performance of local networks. Cloud-based data processing industry is constantly evolving to deliver better performance. The upgrade of the communication infrastructure is the inevitable process.

Please let me know if you have further questions.