## horizontal line



Twitter Coronavirus NLP

Table of Contents

[horizontal line 0](#_Toc99102395)

[Introduction 2](#_Toc99102396)

[Questions 3](#_Toc99102397)

[About Data 3](#_Toc99102398)

[Data cleaning 3](#_Toc99102399)

[Data processing 5](#_Toc99102400)

[Models building 7](#_Toc99102401)

[Techniques/Methodologies 7](#_Toc99102402)

[Multinomial Naive Bayes 7](#_Toc99102403)

[Linear Support Vector Machine (Linear SVC) 9](#_Toc99102404)

[Results 12](#_Toc99102405)

[Perspectives 12](#_Toc99102406)

[Conclusion 12](#_Toc99102407)

# Introduction

Covid-19 was the biggest pandemic the world has seen in awhile. It was a very frightening time in the beginning of the outbreak because no one knew what to expect. Many lives and jobs were lost during this whole tragic event and it is still prevalent in society today. Now this pandemic has impacted a lot of lives negatively, but there were some cases where people had a positive or neutral reaction to this. Certain businesses were able to flourish during covid-19 such as insurance agencies. People were quickly applying to get better insurance and these companies were drastically getting more clients because everyone wanted to make sure that they were insured if anything happened. Another industry that thrived was tech. Since a lot of companies were going remote they needed a way to create a cloud where all of the files can be obtained by the employees or people having the knowledge to fix these systems if they ever go down. There were also neutral opinions about the pandemic mainly during the beginning of 2020. This was because not a lot of people knew about this virus and did not take the effects too seriously.

Since the outbreak of the Coronavirus disease 2019 (COVID-19) in December 2019 and the ensuing pandemic, global opinion has varied about the situation. Many factors have come into play over the last 2+ years: forced or voluntary confinement, mask policies, mandatory or voluntary vaccines, among others. These polarizing factors have led people to express their agreement - and obvious disagreement - on social media, notably Twitter.

The research wondered what the overall sentiment was about the pandemic, how people felt about it and whether there was any noticeable evolution of these feelings over time. The focus was on Twitter feed of 2020, the year the pandemic boomed.

# Questions

What is the overall opinion about the coronavirus?

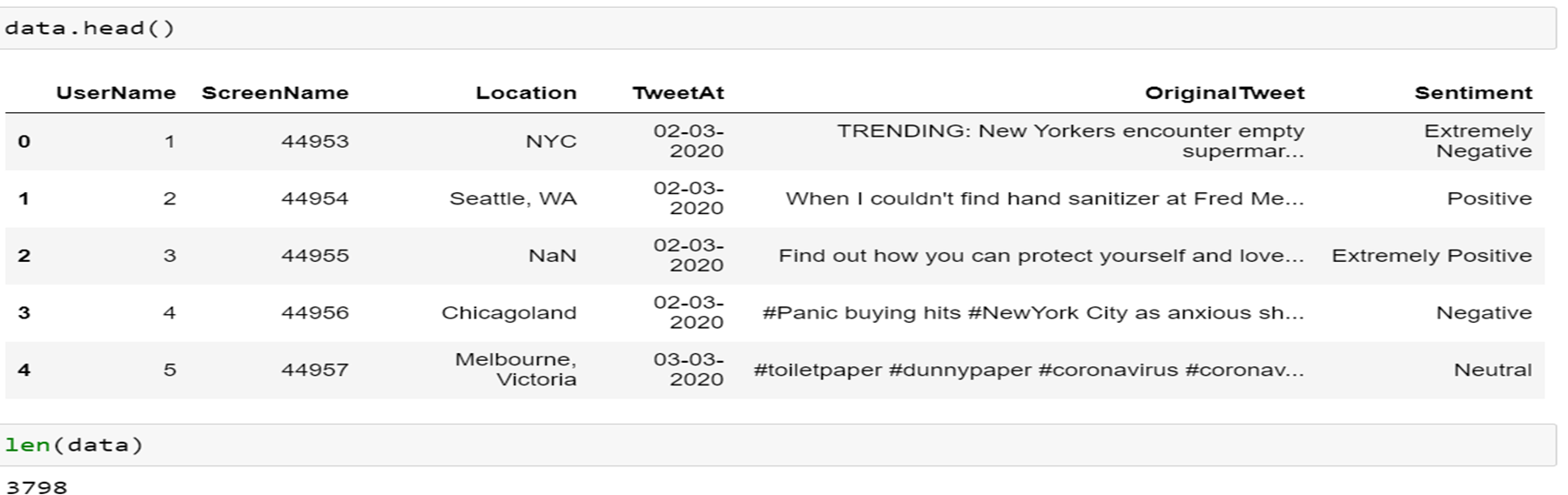
We sought to find out what the consensus was about the coronavirus, whether people were positive, neutral or negative about it by classifying the various tweets through sentiment analysis.

Which tool is better to use for identifying the output?

To accomplish that, we looked at various classifiers and wondered which one was best suited to the task.

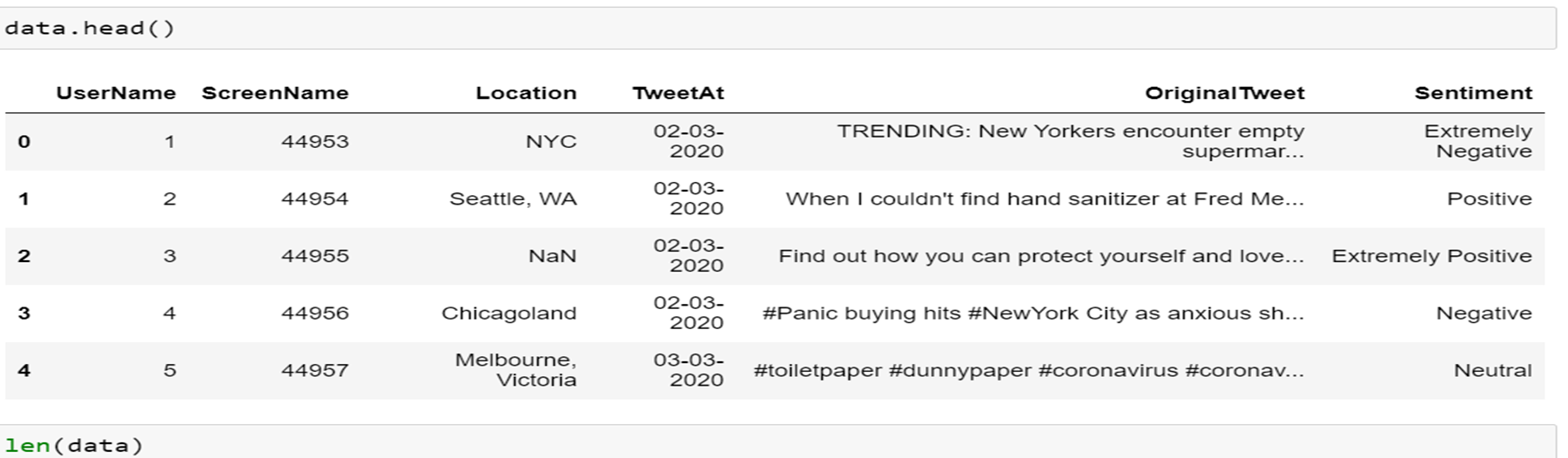
# About Data

The dataset used for this project came from Kaggle API. It contains 6 columns and 3.798 entities. The columns showed the user id, screenname, location, tweet date, tweet, and sentiment. The entities were the Information gathered from 3.798 users in 2020.

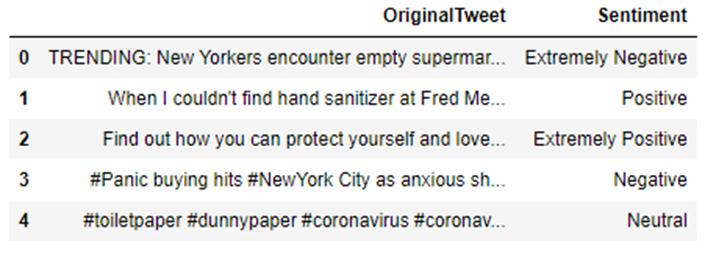


## Data cleaning

The data was scrubbed from Twitter using the #coronavirus hashtag and was limited to the year 2020. The resulting data set had 41,157 entries with 6 variables, ranging from user name to location/date to the tweet content.



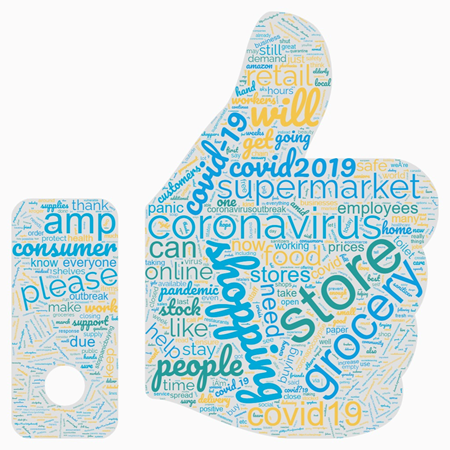
Now twitter data will never be cleaned and the raw data will not be used as is, so we needed to clean it in order to get accurate results. First we wanted to only keep the information that was necessary for what we were looking for. We decided that the Original Tweet and the sentiment were the variables we wanted to use.



Now the sentiment analysis that was provided in this dataset was not 100% accurate because there are a lot of unnecessary words and symbols that make it hard to register what the tweets actually mean. So in the cleaning process we first tokenized the tweets. This helps break down each word and punctuation in a token, so that we can see what words or symbols are being categorized and delete what we don’t need. Then we get rid of words that have no significant value through stop words. These are a lot of filler words, such as “a”, “the”, “at”, etc. Next we use stemming to reduce inflected words to their stem word. Examples would be if a tweet had the words “pandemic” and “pandemics”. The process would only take into account words that are “pandemic” and not repeat any that say “pandemics” because they are basically the same word and do not help with adding sentiment. Then we look at the frequency distribution just to make sure that everything was done right and that there are no unnecessary words or symbols that are still in the tweets.

## Data processing

After cleaning and scrubbing the tweets they were then put into a wordcloud to get a general idea of words people used in order to express their feelings (positive, negative, and neutral).







Some words you can see are in all three categories, mainly due to the context and how the tweets were written. The general consensus was that there were a lot of negative opinions about the coronavirus, but there were still tweets that had neutral and even positive feedback. It is interesting to see what people were saying on Twitter during this time because this was a big outbreak that impacted everyone’s life. Now people live in different circumstances and have been impacted more or less depending on who they are, so it’s interesting to see a wide variety of how people felt.

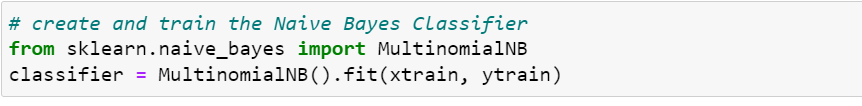
# Models building

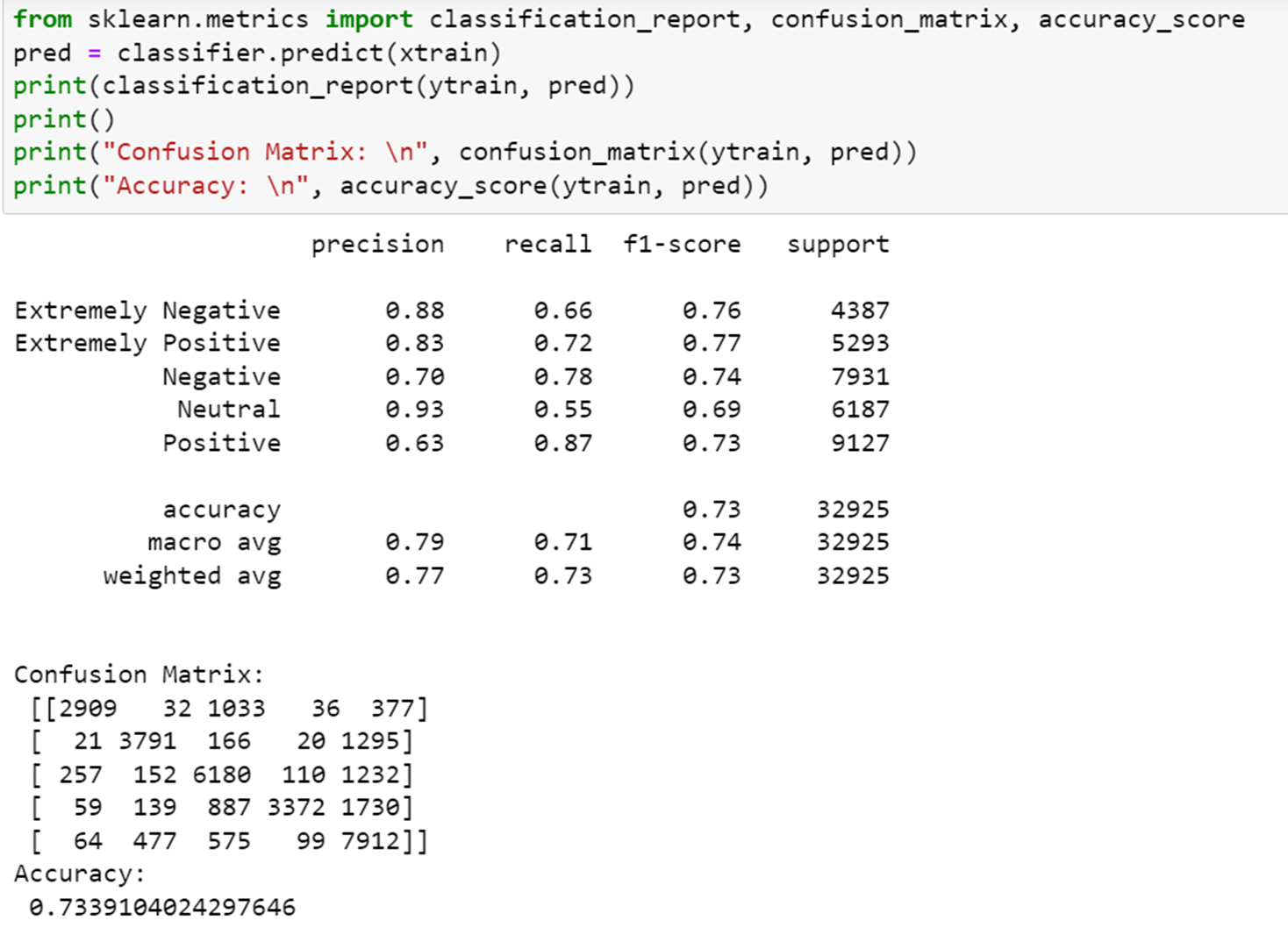
## Techniques/Methodologies

After all the data pre-processing steps, we built two models: Multinomial Naive Bayes and Linear Support vector Machine. These models were built using Sklearn packages. The data were split 80% for training and 20% for testing. Every model was evaluated using cross-validation methods.

## Multinomial Naive Bayes

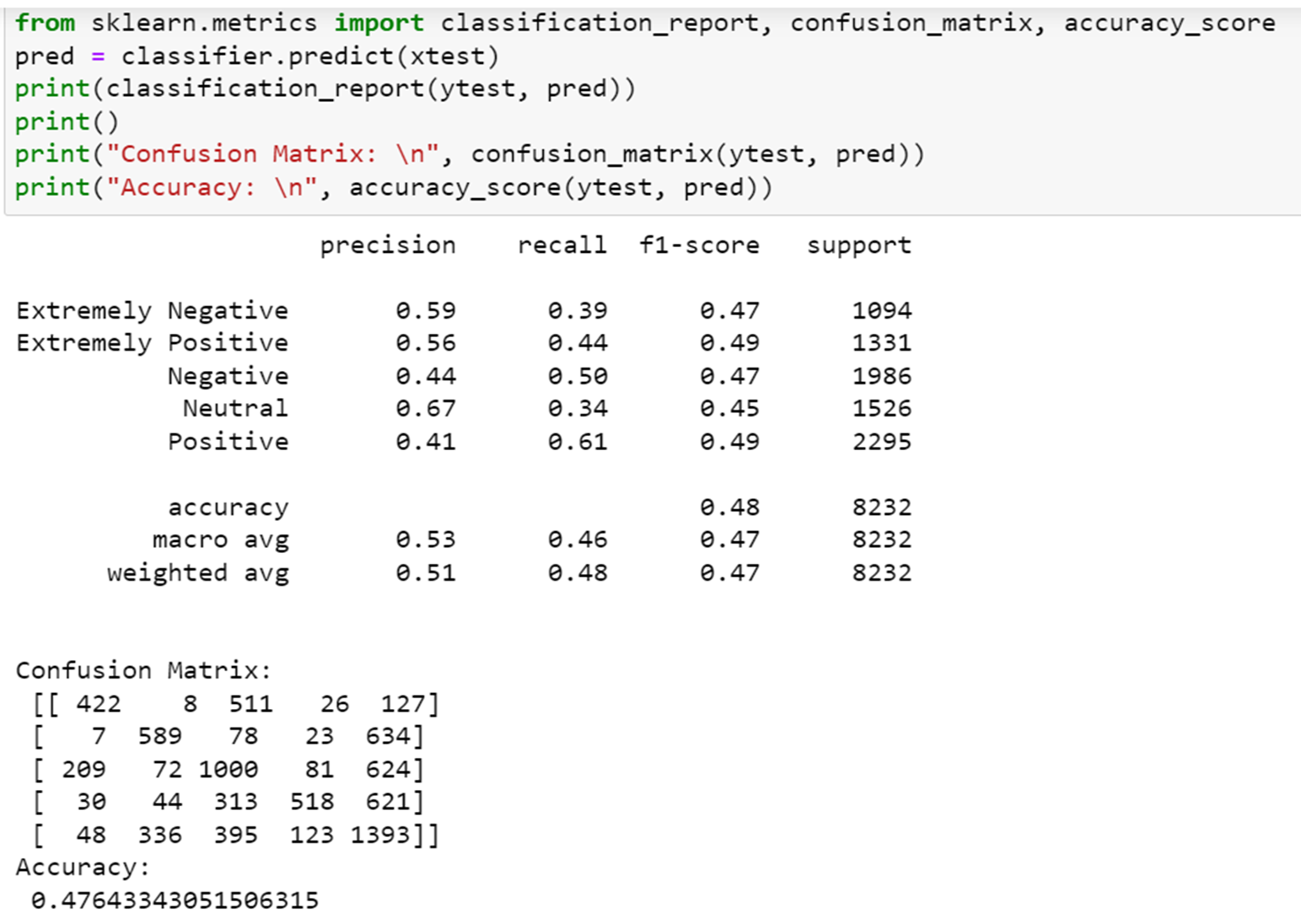
The first Multinomial Naive Bayes of this classification study was run on the training data using 80% of the data. The output came up in a table with the confuse matrix values to evaluate the model.





The confusion matrix of the MNB on the training data came out with overall accuracy of 73%. Besides that we could see that neutral came out as the highest sentiment, 93% followed by “Extremely Negative”, 88%.

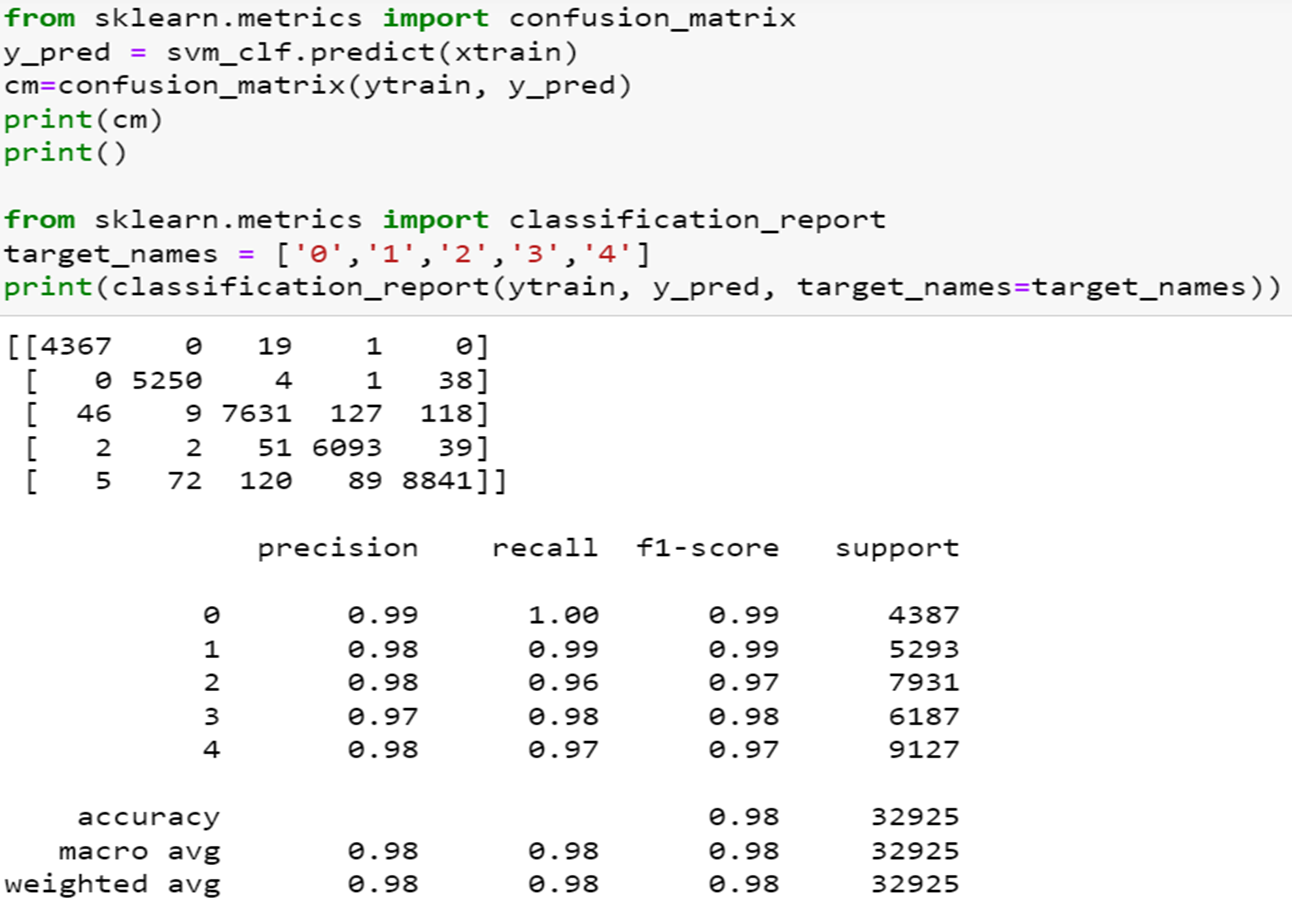
The model was also run on the testing data for the evaluation purpose.

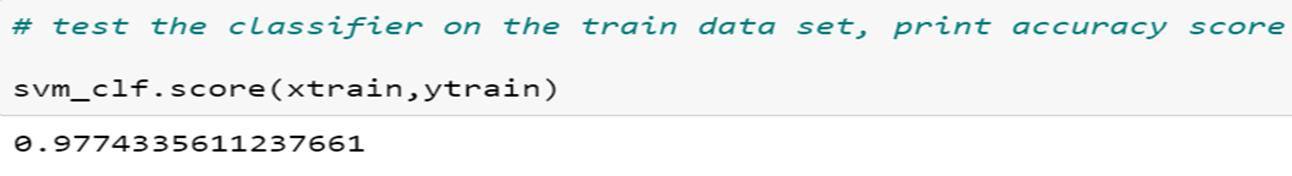


As we could see, the model came up with an accuracy rate lower than it did with training data, 48%. Neutral still be the highest sentiment, followed in this case by Extremely Negative and Extremely Positive.

## Linear Support Vector Machine (Linear SVC)

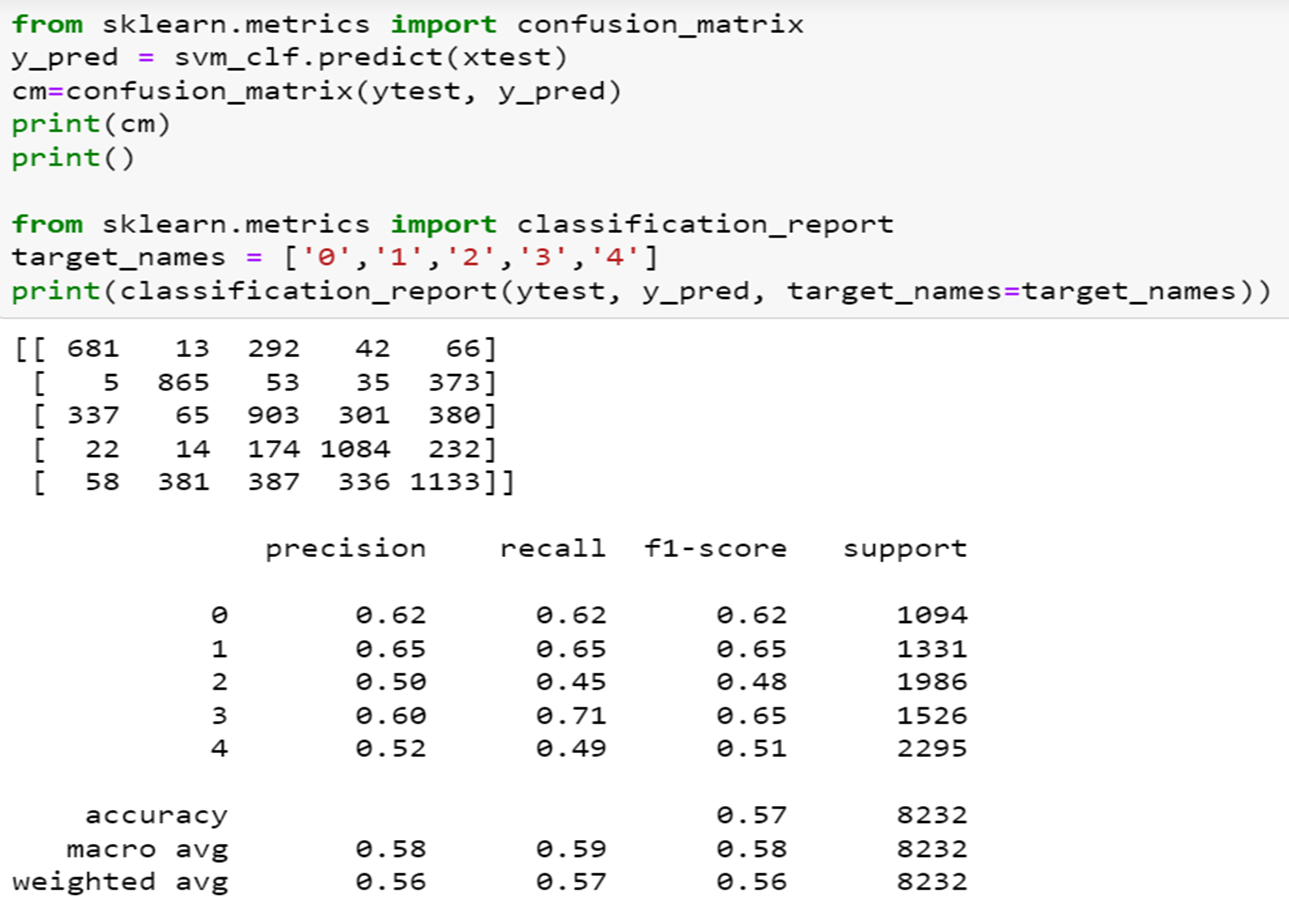
The Linear Support Vector Machine was built using the one provided in the Sklearn package named LinearSVC. Like the same methodology applied with MNB model, we first trained the LinearSVC model on the training data and the outputs in table format with the scores and the confusion matrix.

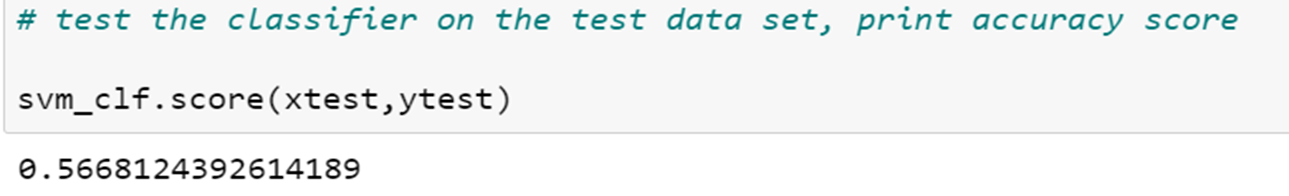




The model performed with 98% overall accuracy rate. Though we could see that the scores were tight between the sentiment annotations (with 1-2 % difference) , Extremely Negative came up with 99%.

We replicated the experiment on the testing data by applying the same model and techniques.





In this time, the model came up with an accuracy rate of 57%. The precision scores showed Extremely Positive as the highest sentiment. This was strange to us and attesting to the bad performance of the model on the testing data.

# Results

* The Support Vector Machine model showed better results (98% of accuracy) than the Multinomial Naive Bayes model did (73%). We definitely recommend SVM as a model to use in the future.

- For all the outputs we got, the models showed better results on the training data than they did on the testing data. With the Multinomial Naive Bayes, the model came up with 73% on training data against 47% on testing data. With the LinearSVC these results were 98% against 58%.

- Overall, the opinions were more neutral and negative about COVID19 than they were positive (to be expected given the nature of the pandemic and the nuisance and problems it caused).

# Perspectives

For the purpose to have more insights about the topic, we thought it would be better to build different models with different features such as Tifdif count-vectorizer or unigram count-vectorizer. That would help for comparative analyses to more understand the topic and the model that fit the best.

# Conclusion

COVID 19 is definitely a ‘hot topic’ and any tweets, internet comments or other publications about it would be interesting to analyze. It has come to permeate the life of billions of people in the past few years and much has been said and written about it. It is definitely an issue that warrants further analysis.

While the results were satisfactory, some of the models could be improved upon.