Stance detection for COVID-19

Drobnitchi Daniel, Suto Robert-Lucian
411

June 2023

1 Introduction

The COVID-19 pandemic has had a profound impact on the world, affecting everything from daily life to global politics and economics. With so much information and misinformation circulating about the virus, it is more important than ever to understand public opinion and sentiment towards the pandemic. Stance detection, the task of automatically determining an author's perspective or attitude towards a particular topic, can provide valuable insights into public opinion, helping us understand what the public needs to hear, in order to be better informed.

2 Data

The data set that we used, Coronavirus tweets NLP, is composed of 44955 unique tweets, by the same amount of users, posted during the pandemic. The data set is split in 3798 values for testing and 41157 for training. The CSV is structured in 6 columns:

- **Location**, which signifies the location from where the tweets was posted;
- **TweetAt**, representing the date on which the tweet was posted;
- **OriginalTweet**, containing the Tweet itself;
- **Sentiment**, which is structured in 5 categories (Extremely Negative/Positive, Negative/Positive and Neutral)
- the remaining **two** consist in data regarding the username of a person, which are all noted as number to respect anonymity.

The scope of the model in this case is to conclude if a tweet is sceptic or not, of the Corona Virus legitimacy. Some examples from the data set are:

1. Extremely Positive: Coronavirus fun fact: if you cough at the grocery store, you get the whole aisle to yourself pretty quickly. #CoronavirusOutbreak #coronavirus #COVID2019
2. Positive: Check out what these folks are up to here in So Cal ? I like this idea ? La Habra supermarket offers special hours for seniors amid COVID-19 crisis https://t.co/ncTXF8TGyf
3. Neutral: ...at this time, our distillery remains in operation, but we will not be offering public tours or hosting functions or events. Our retail store is also closed..

   https://t.co/lYZg2kfsm0"
4. Negative: They should have something where the elderly and the people (who don't stock up a ton of food, essentials) can buy what they need first, then the more greedy and panick buyers can get what they need. ?? #stoppanickbuying #thinkingofothers #coronavirus #COVID19
5. Extremely Negative: New @CSPI consumerÂ's guide examines how 20 largest restaurant chains by sales are handling paid sick leave during COVID-19 pandemic. Results not good: 60

It is important to take into account the relative bias that the social media app has regarding COVID, twitter, baning users which spread misinformation about COVID.

3 Exploratory Data Analysis

The first thing that we looked in the data set for were missing Values. We can observe that besides the Location column, there are no missing values. This is because of the fact that GPS/Location services are optional. Intuitive, this field will not be used for this task, so it's not a seminficiative thing.

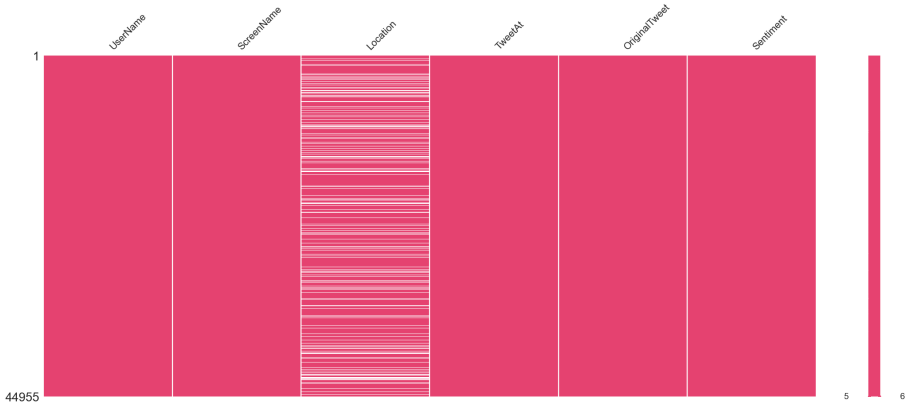


Figure 1: Missing Data

For the simplicity of the Task, we reduced the number of Sentiment classes from five to three, reducing the Extremely labels, to their simple not-extreme sentiment. The distribution is as follows:

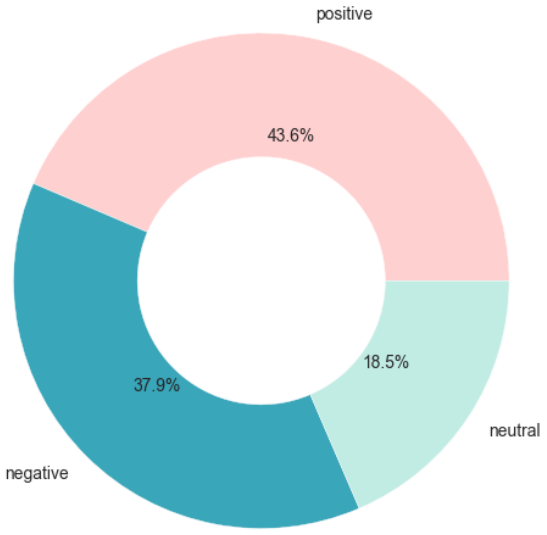


Figure 2: Percentages of sentimental column, in all Data

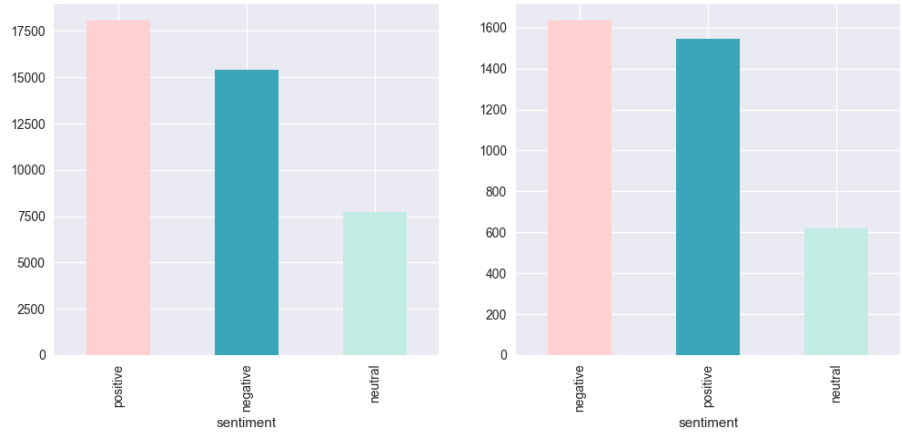


Figure 3: Sentimental column values, in train & test

We also had a look at the number of characters/words related statistics. We can see that regarding length, the positive & negative tweets don't really differ that much. However, the Neutral ones seem to have more words/characters comparing to the others.

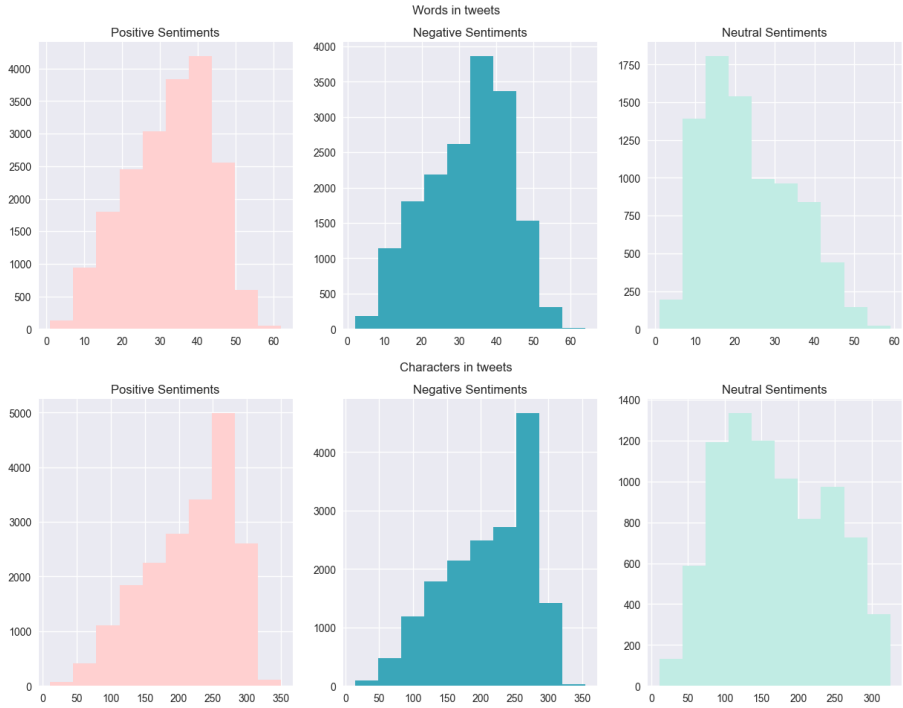


Figure 4: Amount of words and characters based on sentiments

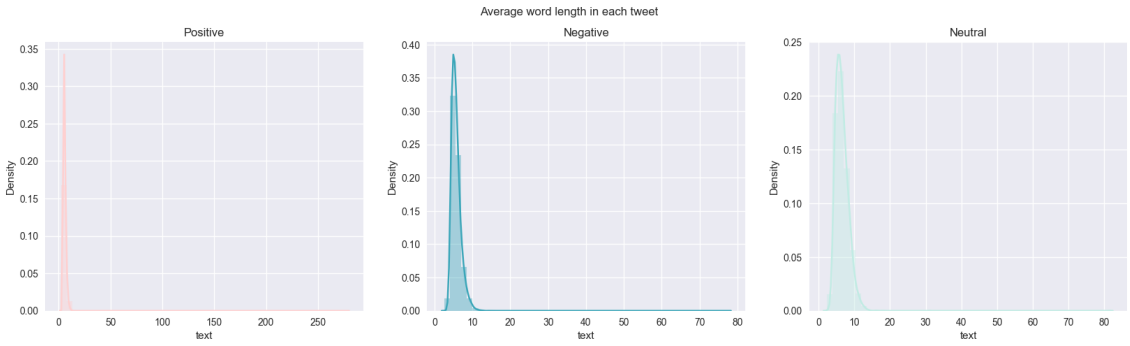


Figure 5: Average words in tweets

We also had a look at the stopwords. For this, we employed the **nltk** library. The most common stop words for each category are as follows:

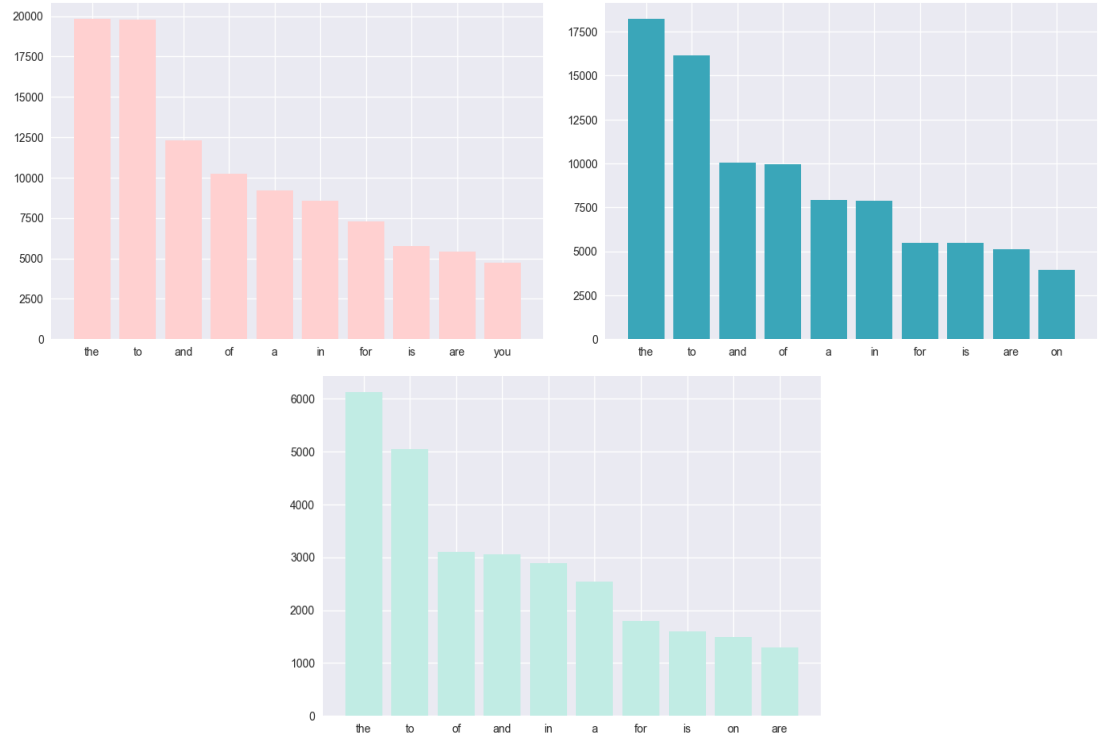


Figure 6: Top 10 most used stopwords across all three sentiments

Special characters are also found in the dataset. Their occurrences histograms are as follows:

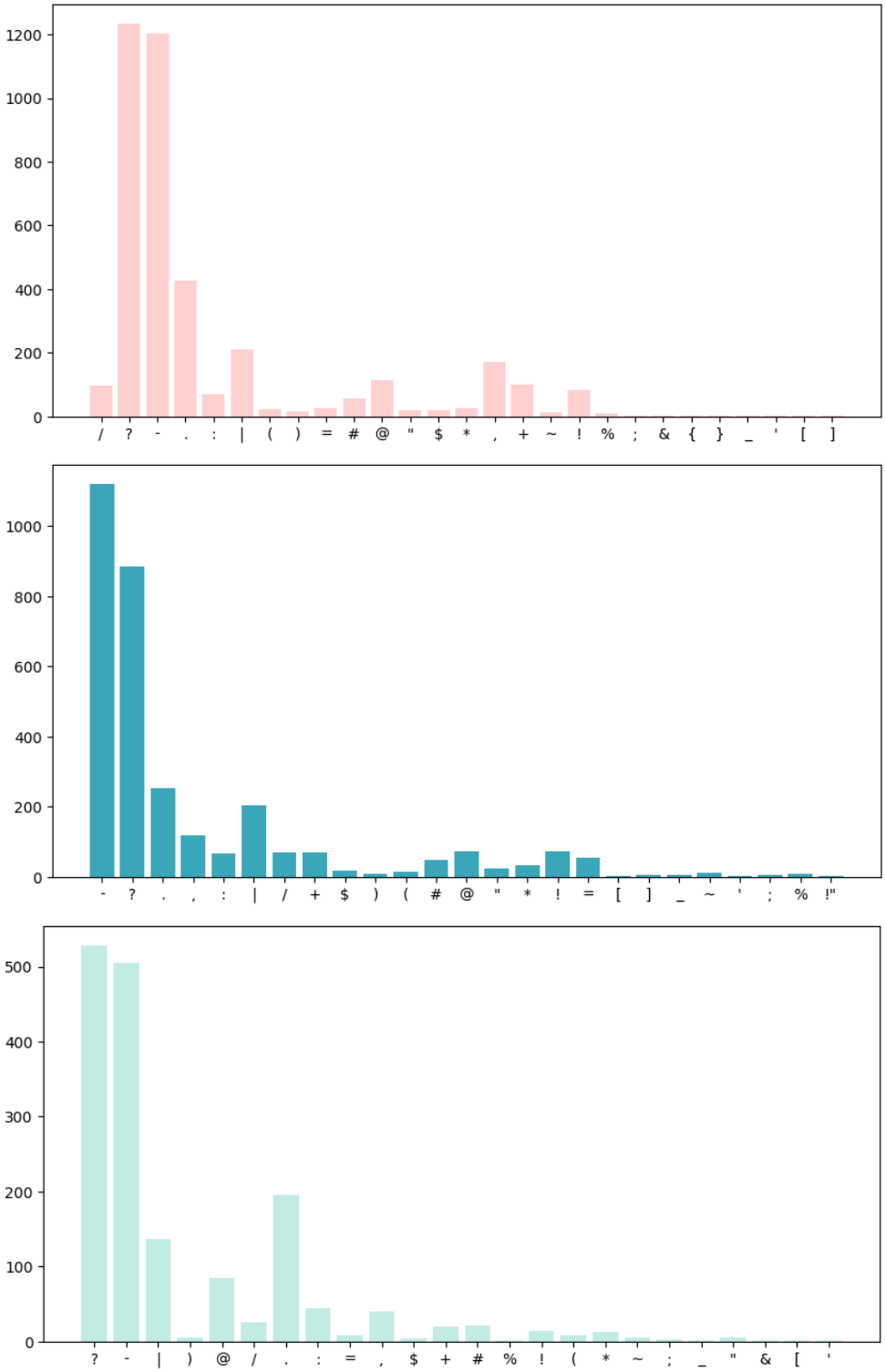


Figure 7: Special Characters across sentiments histogram

We also thought that it would be interesting to have a look at the most used words, mentions and hashtags (beside stop-words) across all of the tweets and mentions.

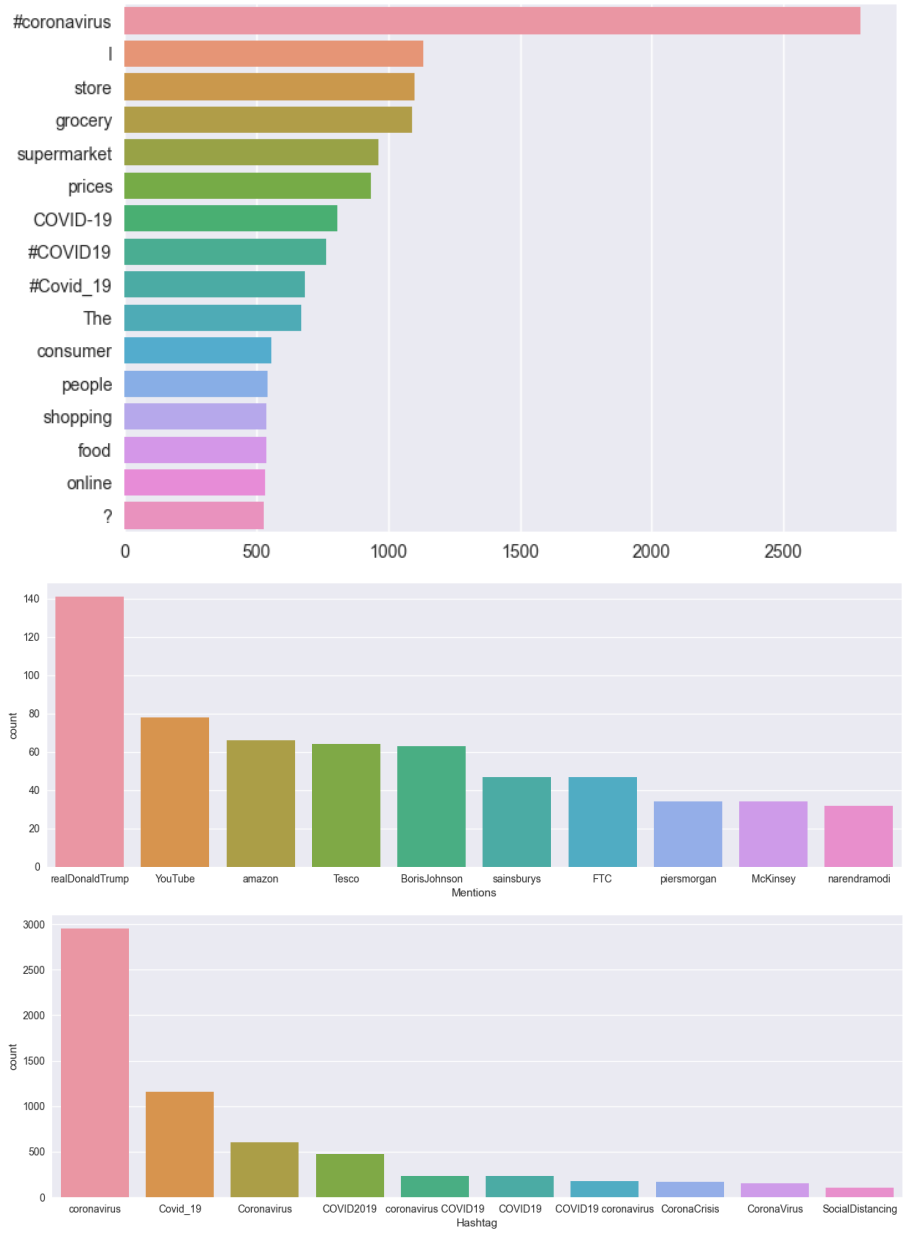


Figure 8: Most common words, mentions and hashtags used.

The fact that the word "**coronavirus**" is the most used on is no surprise.

However, what is surprising is the fact that the most mentioned account is the one belonging to the former President of the United States, Donald Trump. We assume this outcome comes from the false statements he put out during the pandemic. People criticising him but also, his supporters coming to his defence inflating this number by a lot.

Figure 9: Tree of mentions

4 Pre-processing

- We removed unwanted characters, punctuation, and special symbols from the text. This included removing numbers, converting text to lowercase, removing emojis, hashtags and mentions, altogether with links and other special characters.
- We removed common words that did not carry significant meaning, such as articles, pronouns, and conjunctions. These words were removed to reduce noise in the data.
- We split the text into individual tokens, such as words or subwords, using a tokenizer. This step broke the text into meaningful units for further analysis.
- We converted the pre-processed text into numerical features. One-hot encoding represented each token as a binary vector, where each element corre-

sponded to a specific token and was either 1 or 0, indicating its presence or absence.

- We transformed the pre-processed text into numerical features using Term Frequency-Inverse Document Frequency (TF-IDF) encoding. TF-IDF measures the importance of a term in a document relative to a corpus of documents, providing a numerical representation of the text's significance.

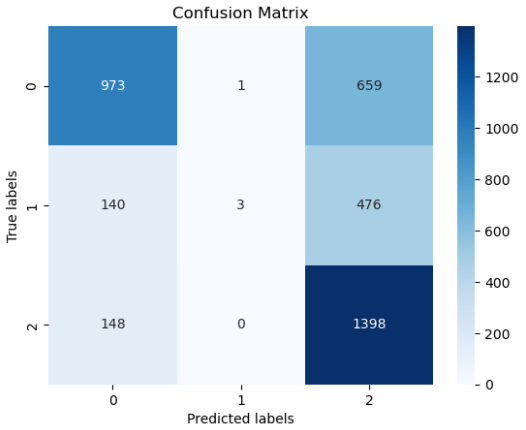
5 Fitting Models

In the model application stage of our project, we utilize the following algorithms:

5.1 Multinomial Naive Bayes

We utilize Multinomial Naive Bayes, a probabilistic classifier, for text classification. With an assumption of feature independence, this algorithm estimates probabilities based on the frequency of occurrence of features, making it suitable for tasks such as document categorization.

The results for this are not ideal, reaching an accuracy of only **0.72**. We used this as a baseline for our next models. The summary of the model is as follows:



Classification Report for Naive Bayes:

	precision	recall	f1-score	support
Negative	0.77	0.60	0.67	1633
Neutral	0.75	0.00	0.01	619
Positive	0.55	0.90	0.69	1546
accuracy			0.63	3798
macro avg	0.69	0.50	0.46	3798
weighted avg	0.68	0.63	0.57	3798

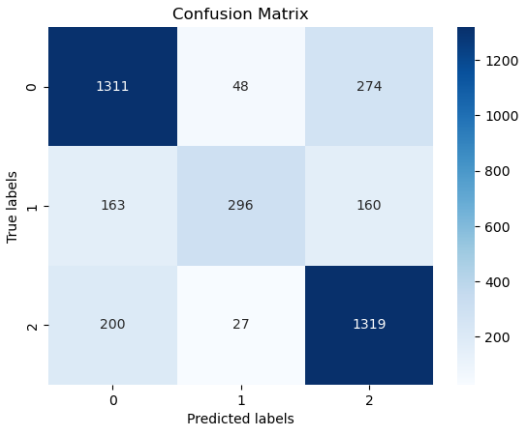
Figure 10: Multinomial Naive Bayes confusion matrix and raport

We can clearly see that our model gets confused at Neutral Tweets. However, this model succeeds at recognising negative tweets, having a bias for this class.

5.2 XGBoost

We apply XGBoost, a powerful gradient boosting algorithm, for enhanced predictive modeling. XGBoost combines weak predictive models, such as decision trees, to create an ensemble model with strong predictive capabilities.

Although simple to implement, this model gave significantly better results than the Multinomial Naive Bayes model, with an accuracy of **0.86**:



Classification Report for XGBoost:

	precision	recall	f1-score	support
Negative	0.78	0.80	0.79	1633
Neutral	0.80	0.48	0.60	619
Positive	0.75	0.85	0.80	1546
accuracy			0.77	3798
macro avg	0.78	0.71	0.73	3798
weighted avg	0.77	0.77	0.76	3798

Figure 11: XGBoost confusion matrix and raport

This model seems to lack the bias towards the negative response, found in the previous model, correctly placing most of the answers.

5.3 BERT

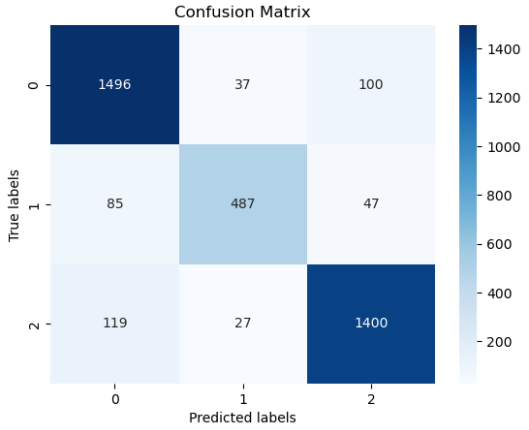
We employ BERT (Bidirectional Encoder Representations from Transformers), a state-of-the-art language model, for language understanding tasks. BERT captures contextual relationships and semantic meaning by leveraging a transformer architecture, making it highly effective in natural language processing tasks.

We used a pre-trained model (Bert Base uncased), with the following architecture:

Layer (type)	Output Shape	Param #	Connected to
input_1 (InputLayer)	[(None, 100)]	0	[]
input_2 (InputLayer)	[(None, 100)]	0	[]
tf_bert_model (TFBertModel)	TFBaseModelOutputWithPoolingAndCrossAttentions(last_hidden_state=(None, 100, 768), pooler_output=(None, 768), past_key_values=None, hidden_states=None, attentions=None, cross_attentions=None)	109482240	['input_1[0][0]', 'input_2[0][0]']
dense (Dense)	(None, 3)	2307	['tf_bert_model[0][1]']

Figure 12: BERT model architecture

This is the model that gave us the best results, sitting at an accuracy of **0.9**:



Classification Report for BERT:

	precision	recall	f1-score	support
Negative	0.88	0.92	0.90	1633
Neutral	0.88	0.79	0.83	619
Positive	0.90	0.91	0.91	1546
micro avg	0.89	0.89	0.89	3798
macro avg	0.89	0.87	0.88	3798
weighted avg	0.89	0.89	0.89	3798
samples avg	0.89	0.89	0.89	3798

Figure 13: BERT confusion matrix and raport

We can see that the number of misplaced positive and negative examples went down by almost a half of what the number was at the XGBoost.

6 Conclusion

In a world where opinions are as diverse as the people who hold them, understanding the public stance on pressing issues such as the coronavirus is crucial. Through the power of Natural Language Processing, our project delved into analysing the many tweets that this virus produced, in order to uncover the stance on this global pandemic. Utilizing three different models - XGBoost, Bert, and Multinomial Naive Bayes - we were able to achieve the following accuracy in our analysis. The Bert model emerged as the leader with an accuracy score of 0.9, followed by the XGBoost model with a score of 0.865 and the Multinomial Naive Bayes model with a score of 0.72. Our findings provide valuable insights into the public opinion on coronavirus and pave the way for further research in this field.

7 References

[1] https://twitter.com/

[2] Stance Detection in COVID-19 Tweets
(https://aclanthology.org/2021.acl-long.127)(Glandtetal.,ACL-IJCNLP2021)

[3] Understanding TF-IDF for Machine Learning (https://www.capitalone.com/tech/machine-learning/understanding-tf-idf/) (Anirudha Simha, October 6, 2021)

[4] "Bert", Huggingface (https://huggingface.co/docs/transformers/model_doc/bert)

[5] Hamad, O.; Hamdi, A.; Hamdi, S.; Shaban, K. StEduCov: An Explored and Benchmarked Dataset on Stance Detection in Tweets towards Online Education during COVID-19 Pandemic. Big Data Cogn. Comput. 2022, 6, 88. https://doi.org/10.3390/bdcc6030088

[6] Stance Detection in COVID-19 Tweets
ACL 2021 · Kyle Glandt, Sarthak Khanal, Yingjie Li, Doina Caragea, Cornelia Caragea