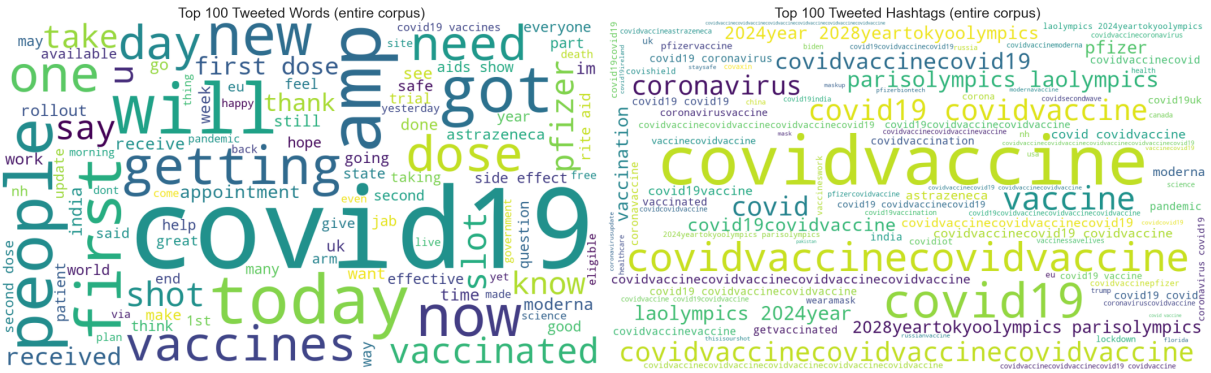
COVID-19 Vaccine Twitter Sentiment Analysis Report

Billie Kim



1 Introduction

Vaccinations are one of the greatest achievements of modern medicine. Partly due to the spread of misinformation, anti-vaccination sentiment has been on the rise even before COVID-19. With the recent outbreak of the novel corona virus, vaccination efforts have become crucial for combating the spread of new infections which in turn, flattens the curve of the COVID-19 crisis. Communities with low vaccination rates face higher risks since vaccines are only effective in populations with high vaccination coverage in order to develop herd immunity against new infections. Thus, it is imperative to understand current sentiment in order to find effective strategies to reduce vaccine hesitancy.

2 Sentiment Analysis

Sentiment analysis is the process of understanding the opinion of a user about a particular subject. It is also referred to as opinion mining and has many practical applications such as in fields of social media monitoring, product analytics, and market research and analysis. Sentiment analysis tasks can be carried out at different levels of granularity— document, sentences and aspect. We will use twitter data to classify and predict sentiment regarding COVID-19 vaccinations expressed in the text of tweets. By understanding current public sentiment, government officials and public health policy-makers can design a more effective approach to educate the public which can further facilitate policy implementation and vaccine deployment strategy.

3 Data Wrangling

3.1 Data Acquisition

A collection of 328,619 tweets is used for analysis in this project. The data was retrieved from a running dataset hosted on Kaggle within the timeline from January 2020 to April 2021[1] . It was pulled

[1]https://kaggle.com/kaushiksuresh147/covidvaccine-tweets

from the Twitter API using the hashtag #covidvaccine. Each record contained the user's name, location, tweet date, text, and hashtags. It also contained some other metadata such as user's description, creation date, followers count, favorites count, verified, text, source, and retweet status.

3.2 Annotations: sample (n=1000)

The data did not come pre-labeled. We sampled 1000 tweets and hand labeled them into three polarities: positive, neutral, and negative. Positive polarity referred to tweets where users expressed enthusiasm for getting the COVID-19 vaccine. Tweets where users displayed willingness or encouraged others to be vaccinated were marked as positive. Tweets where users neither agrees nor disagrees with the vaccination were marked as neutral. Negative polarity was assigned to tweets that disagreed with the COVID-19 vaccination. Tweets that displayed negative reactions and refusal to the vaccine as well as experiencing adverse side effects from vaccination were marked as negative.

The following are examples of tweets are labeled in the form of opinions about covid vaccine in terms of their polarity:

1. "So, in our limited experience, this vaccine is quite clearly more dangerous than covid19 just saying folks be warned be prepared." **Negative**
2. "I don't know what to believe or do anymore about this covidvaccine ppl are saying take the shot it could help not get a sick with deltavarient then ppl saying the vaccine taker will be targeted with delta I am confused." **Neutral**
3. "It is so important that we all receive the covid vaccine if we are to beat or even keep this disease at bay we are so lucky to have our amazing help them to help you get your jab stay safe be covidsafe." **Positive**

4 Exploratory Data Analysis

4.1 Date Trends

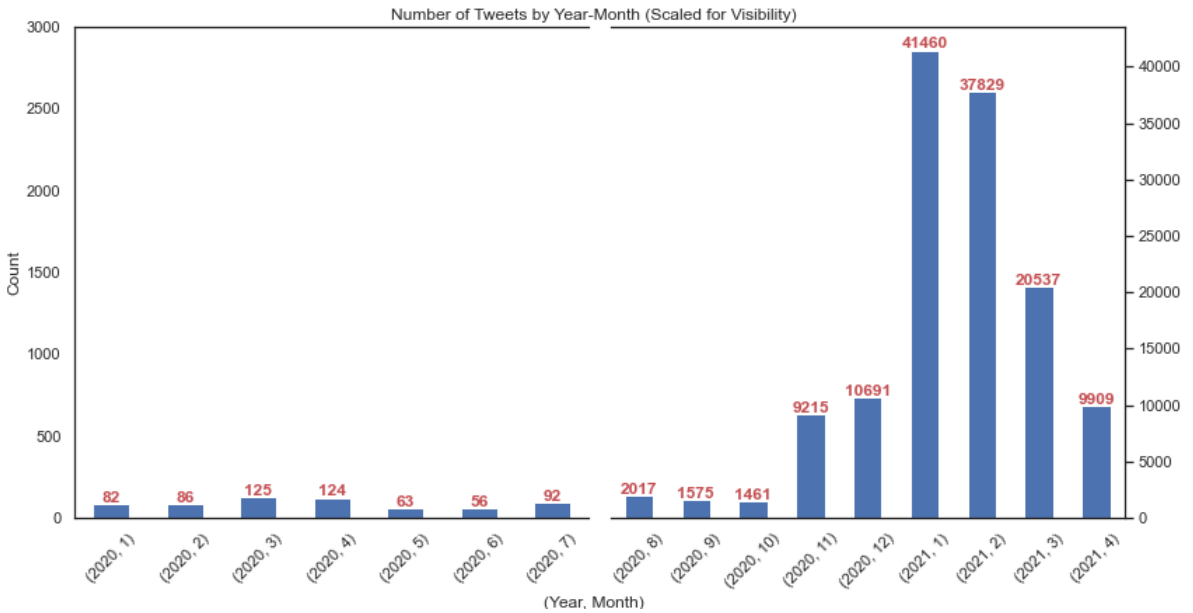


FIGURE 1: Tweets Counts by Year-Month (Scaled for Visibility)

In 2020, there were 25587 tweets and in 2021, there were 109735 tweets. The highest number of tweets were recorded in January 2021 with 41460 tweets. Figure 1 shows distribution of tweets within the data collection timeline.

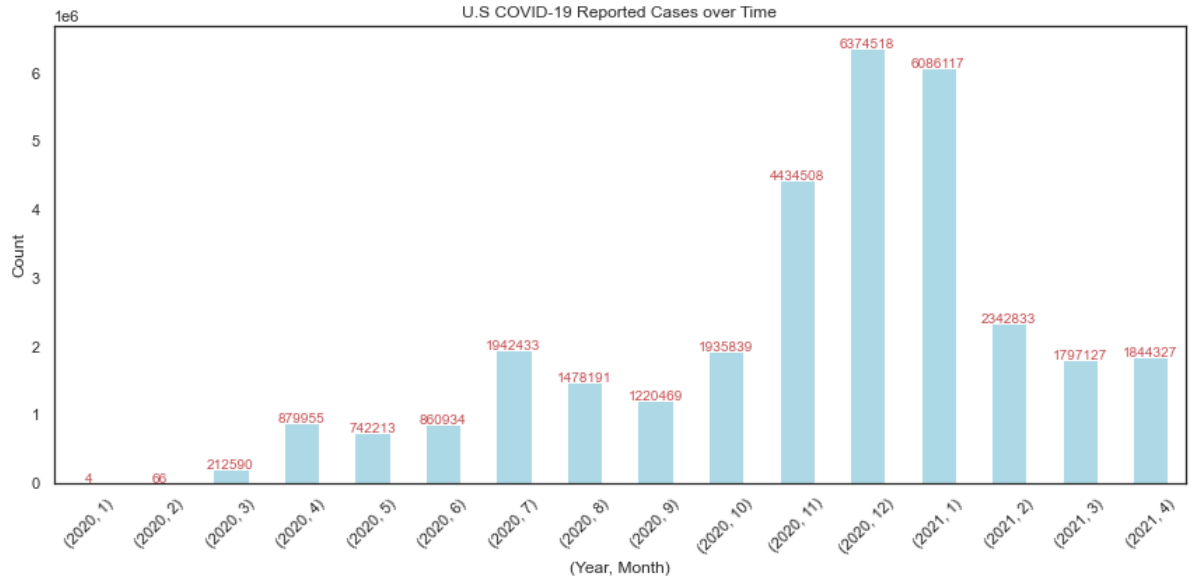


FIGURE 2: U.S COVID-19 Reported Cases (*CDC Data*)

We notice the high increases of tweets in the months of August 2020, November 2020, and January 2021 with percentage changes (from the previous month) of 20.92%, 5.31%, and 2.88%, respectively. These spikes seems to corraborates with the timeline of reported case counts in the U.S.[2]

4.2 WordClouds and Influential Terms

After annotating our sample of 1000, we created word clouds (Figure 2) for each sentiment class which gave us some initial insights and intuition into how our classes are distributed. Although we noticed some overlap between commonly used terms between classes, we do notice some unique terms that are specific to each class. For example, in the negative word cloud, we see words like *protest*, *dying*, *coididiots*, and *side effect*, while in the positive word cloud, we see terms like *vaccinated*, *hopeful*, *grateful*, and *received*. Later on, we were able to see how these key influential terms affected our model by taking a deeper look into their weights which can be seen in Figure 3.

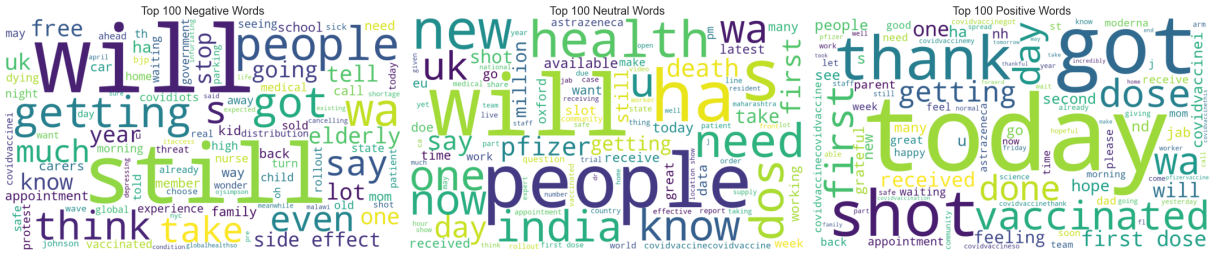


FIGURE 3: WordClouds (sample, n=1000)

[2]https://data.cdc.gov/Case-Surveillance/United-States-COVID-19-Cases-and-Deaths-by-State-o/9mfq-cb36/data

y=Negative top features		**y=Neutral** top features		**y=Positive** top features	
Weight?	**Feature**	**Weight?**	**Feature**	**Weight?**	**Feature**
+1.015	still	+1.006	<BIAS>	+2.350	my
+0.904	not	+0.708	he	+1.757	covidvaccine
+0.899	if	+0.681	by	+1.568	got
+0.820	cannot	+0.658	about	+1.406	dose
+0.731	they	+0.657	show	+1.323	vaccinated
+0.619	only	+0.653	health	+1.215	done
+0.582	because	+0.608	from	+1.123	today
+0.575	sold	+0.607	india	+1.064	first
+0.569	america	+0.529	what	+0.980	hope
+0.561	are	+0.514	open	+0.904	thank
+0.546	elderly	+0.497	million	+0.900	grateful
+0.521	no	+0.484	your	+0.899	so
+0.519	think	+0.463	dr	+0.831	shot
+0.504	maybe	+0.460	on	+0.801	nh
+0.500	wonder	+0.449	are	+0.790	feeling
+0.497	yes	+0.449	national	+0.762	she
+0.489	free	+0.449	oxford	+0.739	waiting
+0.489	dying	+0.437	available	+0.730	thanks
+0.485	covidiots	+0.417	new	+0.727	see
+0.477	most	+0.414	death	+0.717	just
…467 more positive …		*…1553 more positive …*		*…744 more positive …*	
…2078 more negative …		*…992 more negative …*		*…1801 more negative …*	
-0.387	your	-0.862	just	-0.451	show
-0.417	coronavirus	-0.871	am	-0.456	effect
-0.421	my	-0.873	waiting	-0.461	on
-0.427	today	-0.910	still	-0.478	from
-0.449	covid	-1.013	done	-0.488	how
-0.496	at	-1.026	dose	-0.527	health
-0.497	first	-1.091	vaccinated	-0.529	not
-0.516	for	-1.541	got	-0.549	say
-0.600	vaccine	-1.605	covidvaccine	-0.582	about
-0.869	<BIAS>	-1.929	my	-1.010	are

FIGURE 4: Influential Terms and Weights

5 Pre-Processing and Training Data

5.1 Pre-processing Steps

Text data is known to be unstructured data. Thus, it is important to clean/preprocess tweets in order to standardize textual form. This step is crucial task for NLP tasks as it transforms text into a more digestible form so that machine learning algorithms can perform better in the long run by removing unwanted noise.

We list our preprocessing steps below. Instead of aggressively removing stop words, we ran some experiments and found that including English stop words helped our model to perform better.

Preprocessing Steps Outline:

1. HTML decoding — HTML encoding will be decoded into standard English characters
2. Removing twitter handles (@mentions) and URL links
3. Expand contractions
4. Lower-casing
5. Removing numbers and special characters
6. Tokenization
7. Lemmatization

5.2 Feature Extraction

Next, we needed a method to convert text data into a numerical/vectorized representation in order to run machine learning algorithms. We explored 2 ways to go about this. First, is the Bag of Words approach where we use CountVectorizer to convert strings into a frequency representation. This, however, results in biasing the most frequent words and ends up ignoring unique words. The other approach is using TF-IDF (Term Frequency — Inverse Document Frequency) which are word frequency scores. TF-IDF summarizes how often a given word appears within a document and also penalizes words that appear most frequently across all documents.

We ran experiments to see which of the two vectorizers performed better with Logistic Regression at different “n-grams” (unigrams, bigrams, and trigrams) as well as including or removing stop words

in the data. For both vectorizers, we found that including stop words actually helped the performance of the model at all levels of ngrams. Overall, TF-IDF, produced higher test accuracy scores than Count Vectorizer. We found that TF-IDF with bigrams and including stop words yielded the most optimal results with 80.33% validation accuracy using 700 features. Based on these results, we proceeded to modeling using TF-IDF vectors.

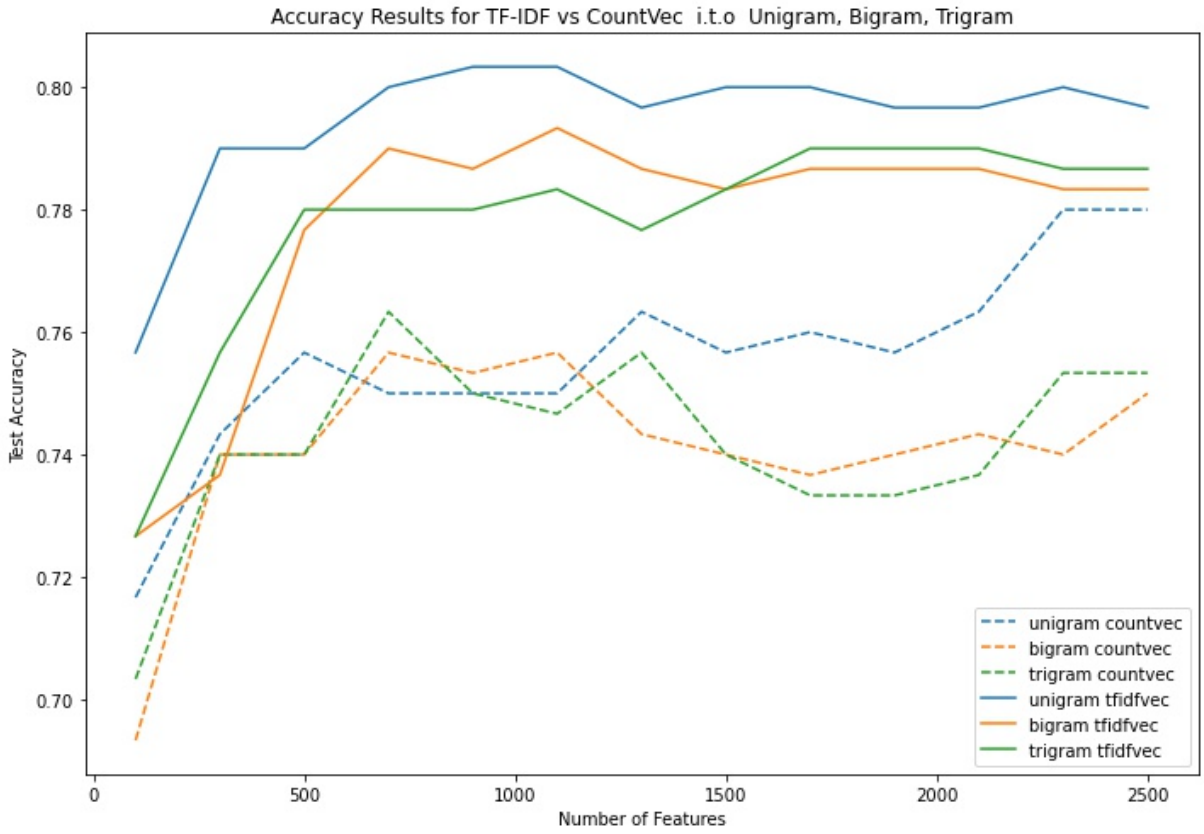


FIGURE 5: Comparing TF-IDF vs CountVec - unigrams, bigrams, trigrams

6 Modeling Approaches

We used three different approaches in our modeling phase. The first modeling approach uses the original sentiment categories (positive, negative, neutral) that we established in our sampling/annotation stage. However, during our analysis, we found a large number of tweets that were news-related announcements that contained factual information rather than subjective opinions of users. In model 1, we ultimately marked these tweets as neutral; however, in model 2, we explore these tweets as a sentiment class on its own and introduced a fact/announcement class. We found support that a separate fact/announcement class was indeed necessary and relevant to understanding COVID-19 vaccine of actual users and that aggregating facts/announcements with neutral polarity tweets would be overall incorrect and distort real users with having neutral sentiment. In model 3, we build upon the sentiment classes established in model 2 and introduce a 2-step model system.

6.1 Model 1: Restricted/Nested Model

The first model is our base Logistic Regression model with a baseline accuracy of 57.9%. There seemed to be a large class imbalance in the distribution of the data. The majority class was the neutral class with 579 records, while positive and negative classes had 327 and 97 occurrences, respectively. We found that

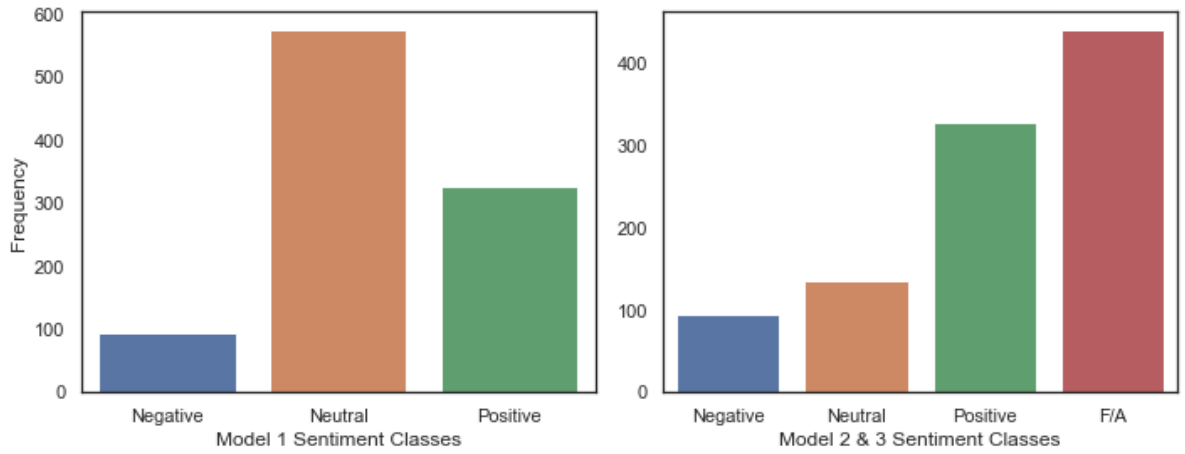


FIGURE 6: Sentiment Classes Distribution

this model with ideal feature parameters had an accuracy of 80%. However, the model was not capturing any of the negative class in the test data.

After hyper parameter tuning, our baseline model achieved a test accuracy of 81% and was able to predict the negative class 12% of the time. At this point in the analysis, this raised several questions such as whether there is a problem with our class labels and if there should be separate class for facts/announcements and news-related tweets.

6.2 Model 2: Full Model

In the 2nd modeling approach, we added a separate class for Fact/Announcements along with the Negative, Neutral, and Positive classes. We found that almost half (427 records) of our sample data belong to the facts/announcements. This shrunk the counts of the Neutral class from 579 to 152 records while the frequency in the negative and positive classes stayed predominately the same.

From tuning hyper parameters, we found a test accuracy of 71%. Although, this modeling approach had a lower test accuracy than the first model, the predictive power of the negative class rose from 12% to 33%.

6.3 Model 3: 2-Step Model

In model 3, we wanted to build upon our full model (model 2) and attempted a 2-step model/pipeline approach.

In step 1, we introduce a subjectivity classifier that classifies if a tweet is a fact or opinion. We tested 3 different machine learning algorithms: Logistic Regression, Ridge Classifier, and a Voting Ensemble method. Logistic Regression achieved a 81.7% test accuracy while the Ridge Classifier achieved a 82% test accuracy. The highest test accuracy for the subjectivity classifier was reached at 84.3% using a Majority Voting Ensemble with the following input base models: Logistic Regression, Decision Tree, and Support Vector Classifier.

In step 2, we fed the predicted output opinions from the subjectivity classifier into the sentiment classifier using our base Logistic Regression model. The most notable increase in test accuracy of 75.3 can be seen from using the Voting Ensemble during the subjectivity classification stage.

(Step 1) Subjectivity Classifier Algorithm	Test Acc.	F1 (macro)	F1 (weighted)	Error-Rate
Logistic Regression	0.817	0.808	0.814	0.183
Ridge Classifier	0.820	0.814	0.819	0.180
Voting Ensemble	0.843	0.842	0.844	0.157

FIGURE 7: Model 3 (Step 1): Subjectivity Classifier Algorithm Comparison Results

(Step 2) Full System Algorithm Combination	Test Acc.	F1 (macro)	F1 (weighted)	Error-Rate
Logistic Regression | LR	0.693	0.588	0.695	0.307
Ridge Classifier | LR	0.700	0.591	0.699	0.300
Voting Ensemble | LR	0.753	0.624	0.738	0.247

FIGURE 8: Model 3 (Step 2): Full 2-Step Model System Results

7 Model Selection & Evaluation

Based on our modeling steps and approaches, we found that model 3 seemed to be the ideal model for our target problem. Although model 1 had the highest test accuracy than the other subsequent models, we determined that it would be grossly inaccurate to use it to predict public sentiment about the vaccine since it does not distinguish facts/announcements as its own category. In addition, in terms of their respected accuracies and their baseline (null) accuracies, both model 2 and model 3 show higher improvements of 27.97% and 32.63%, respectively, than model 1 which has lower improvement of 23.43%.

Model	Algorithm / Method	Test Accuracy	Null Accuracy	Δ Acc.	Error-Rate
Model 1: Reduced Sentiment Model	**Logistic Regression**	0.813	0.579	+23.43%	0.187
Model 2: Full Sentiment Model	**Logistic Regression**	0.707	0.427	+27.97%	0.293
Model 3: 2 Step Model	**Logistic Regression | LR**	0.693	0.427	+26.63%	0.307
	Ridge Classifier | LR	0.700	0.427	+27.3%	0.300
	Ensemble Voting | LR	0.753	0.427	+32.63%	0.247

FIGURE 9: Comparing Metrics & Model Evaluations

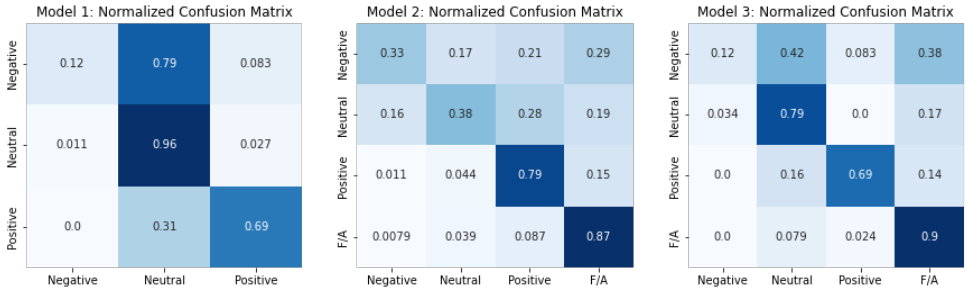


FIGURE 10: Normalized Confusion Matrices for Models 1-3

8 Next Steps

There have been many advances in Natural Language Processing especially in transformer-based language models which uses unsupervised machine learning methods to gain higher accuracies in sentiment analysis tasks. In the future, we would like to build upon our initial work by applying transformer models and transfer learning to increase not only accuracy but also run-time and training.