Sentiment Analysis on Trending COVID-19 Topics

Group 16: Fu Tianyu, Lionel Koh Geng Chye, Oh Kay Yew Patrick, Zhu Yijie

{e0686130,e0503530,e0506467,e0323600}@u.nus.edu

Overview & Approach

COVID-19 has drastically disrupted our way of life. In our study, we attempt to get a greater understanding on some of its impacts by using Natural Language Processing (NLP) to perform ***Sentiment Analysis*** on COVID-19 tweets.

We focus on three controversial 🔥 sub-topics on COVID-19 – **vaccine, masking, and social distancing**. The tweets are gathered from **#USA**, **#Singapore** and **#Malaysia**. 🇺🇸🇸🇬🇲🇾

Our Approach:
- Perform a study on the tweet dataset to understand its **dynamics**
- Design **pre-processing** and manual features that target tweet and sentiment classification
- Train and test out various models
- Select **#bestmodel** 🏅 and perform an ablation study
- Combine our **#bestmodel** with open-source APIs such as **@TextBlob**, **@NLTK-VADER** and Google Natural Language API to form an ensemble for sentiment classification
- Use a Max-Vote on the ensemble to predict the final sentiment label
- Extract the insights on the mentioned trending topics.

Dataset Study and Pre-processing Methods

Initial dataset: 48,948 tweets – COVID tweets extracted from IEEE (Sep20 to Feb21)

Preliminary Observations:
- Mentioned Corona as a city in California
- Mentioned Corona as a beer
- Consisted only of tweet statistics
- Had three or fewer words (excluding the URL)
- Consisted with only a caption saying, "Just posted a photo/video", with media attached

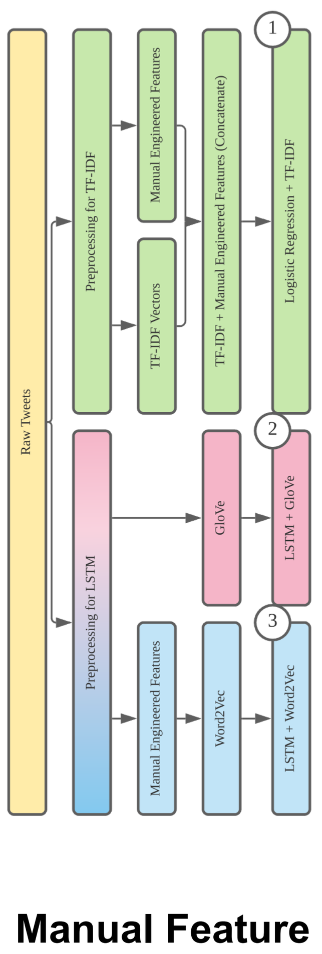
Filtered dataset: 45,551 tweets

Notable tweet dynamics:
- Average word length: 23 words – Tweets are usually consisted of short sentences
- 95.66% of tweets with URLs – Majority of tweets contained URLs that provide no sentiment value (e.g. "https://t.co/rwFB1KXD4g").
- 18.79% of tweets with one of emojis, emoticons, slangs and hashtags

Other observations:
- Large proportion of factual tweets – Sentiment distribution might skew towards neutral sentiments.
- Informal language – Most tweets do not exemplify the proper use of punctuation. E.g., "COVID is here, but it is not going to get me down!" becomes "COVID is here but it is not going to get me down!"
- Evolution in usage of hashtags: E.g., "#COVID VACCINE #BELIEVE THE SCIENCE #PROTECT YOUR FAMILY #PROTECT YOUR COMMUNITY #HAMPTONHIPANDKNEE.COM https://t.co/KG4RxK38ZQ"

No	Pre-processing Step	Description
#1	Removal of URLs, digits, stopwords, punctuation and characters of short length	URLs, digits, stopwords, punctuation and characters of short lengths are removed to keep only the necessary information and remove noise.
#2	Conversion of expressive emojis and emoticons to sentiments	Convert emojis and emoticons to a token that represents their sentiment. E.g., The 😠 emoji is converted to the token "negative", and the =D emoticon is converted to the token "positive".
#3	Conversion of non-expressive emojis to words	Conversion of non-expressive emojis to words E.g. "This place is on 🔥!" becomes "This place is on fire!"
#4	Conversion of slangs/ abbreviations to words	Conversion slangs/acronyms to words. E.g., "lol" is converted to "laugh out loud".
#5	Named entity (NE) concatenation	Entities like "Donald Trump" are identified and concatenated into "Donald_Trump". The intuition is to reduce dimensionality of the vocabulary, and help downstream classifiers to distinguish between similarly named entities (e.g., Donald Trump versus Donald Duck)
#6	Prepending "NOT_" to handle negation	Prepend "NOT_" until the next punctuation token to appropriately reflect sentiments [1]. E.g. "I don't like wearing a mask." becomes "I don't like NOT_wearing NOT_a NOT_mask.". However, due to the informal language structure of tweets, a minor modification is added to consider conjunction words like "but", "unless", "except", etc. Therefore, "not rich but poor" becomes "not NOT_rich but poor", instead of "not NOT_rich but NOT_poor".
#7	Lemmatization	Lemmatization was used to reduce dimensionality of the vocabulary by mapping words back into a single root form (e.g., sings sang sung -> sing)

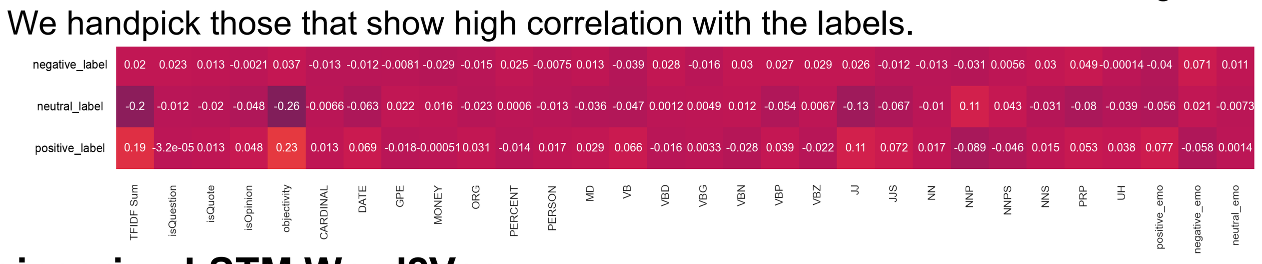
Feature Engineering (1)



Three sets of tweets-to-features pipeline explored:
1. TF-IDF vectorization to generate a bag of words, concatenated with its own manually engineered features
2. Pre-trained GloVe embedding model that was trained on 27B twitter corpus with 200 dimensions
3. Word2Vec with manually engineered features fed into LSTM as an embedding layer

Manual Feature Engineering for TFIDF
- IsX Flags: checks if a tweet contains X, flags 1 if present. E.g., isOpinion, isQuestion
- TF-IDF Sum: Sums the TF-IDF scores of all words in a tweet
- Named Entities (NE): flags if a particular NE is detected e.g., PERSON, ORGANISATION
- POS Tags: flags and normalizes the count of POS Tags in a sentence
- Count of Emoji/Emoticon by Sentiment: E.g., Positive Emoji = 3, Positive Emoticon = 1

Intuition: additional columns serve as additional information for the classifier to segment data on. We handpick those that show high correlation with the labels.



Manual Feature Engineering LSTM Word2Vec
- Additional tags were added in the training corpus – on top of the preprocessing steps performed – as manually tagged features.

Helluva great powder day on 9/23/2020 @vailmtn ❤️ Thanks a looot babe :) #wearamask and FIGHT against COVID!!! https://t.co/WwLx9zT

helluva great <pos_token> powder day <date> <user> <pos_emo> thanks lot <elongated> babe <pos_emo> <hashtag> wear mask </hashtag> fight <allcaps> covid <allcaps> ! ! <url>

sentiment token tag — term tag
elongated token tag — emoji / emoticon tag
hashtag annotation ! special punctuations
all capital token tag

Feature Engineering (2)

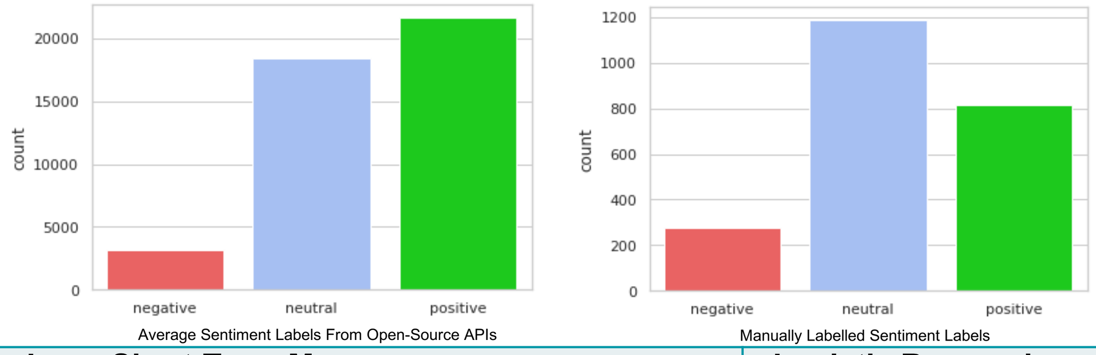
	Remove <allcaps>	Full tagged	Remove !	Remove ?	Remove emo tags
F1-macro score	53.35%	**53.27%**	52.59%	52.40%	52.31%
	Remove term tags	Remove hashtags	Remove all tags	Remove sentiment token tags	
	52.07%	**50.84%**	50.29%	**50.15%**	

Experiments

Datasets for LSTM training

Datasets	Strengths	Limitations
Stanford Sentiment140 dataset 1.6 million tweets	• Large data size and variance	• No "neutral" sentiment labels • No emojis and emoticons • Dataset content has no relevance to COVID
COVID-19 tweets 45,551 tweets – Split into 95% train and 5% test dataset: 43,273 and 2,278 respectively	• Tweets are related to COVID	• No ground truth labels: **Solution:** • Test dataset labels are manually annotated by the team • Train dataset labels are custom engineered by using an average score between TextBlob, NLTK-VADER and Google API.

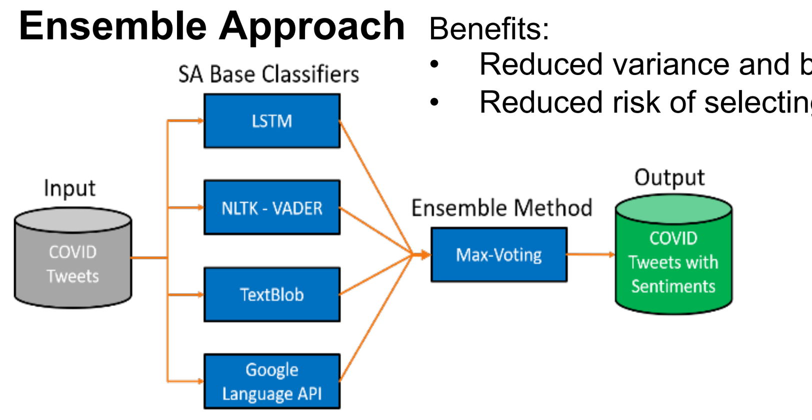
Unbalanced label distribution
- F1-macro score to be used as evaluation metric



Model Comparisons

Model	Long Short-Term Memory			Logistic Regression
Dataset	Stanford Sentiment140	COVID tweets with custom engineered labels		
Variant	Full pre-processing steps		Partial pre-processing steps with annotated tags	Partial pre-processing steps
Features	Pre-trained GloVe	Pre-trained GloVe	Self-trained Word2Vec	TF-IDF
F1-macro score	40.59%	**56.54%**	53.27%	55.70%

Ensemble Approach



Benefits:
- Reduced variance and better generalisation of results [2]
- Reduced risk of selecting a poor classifier

Models	F1-macro Score
LSTM	56.54%
NLTK-VADER	54.75%
TextBlob	54.36%
Google API	58.81%
Ensemble via Max-Vote	**61.26%**

Effectiveness of Pre-processing Steps

Implementation: (10k tweets sampled from the training dataset to retrain LSTM models)
- Start from the best model consisting of ALL pre-processing steps, then remove one step at a time to analyse its contribution to the performance.

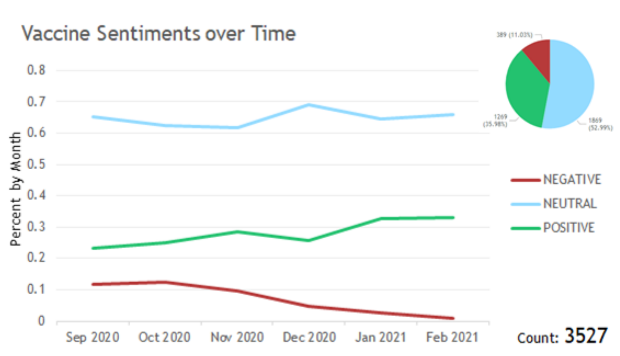
	ALL	#1	#2	#3
F1-macro score	51.63%	46.78%	51.07%	51.80%
	#4	#5	#6	#7
	48.15%	52.47%	51.17%	43.45%

Pre-processing steps with higher effectiveness:
- Removal of steps #1, #4 and #7 resulted in highest drop in F1-macro score.
  - Steps mentioned were to perform text normalization and removal of noise.

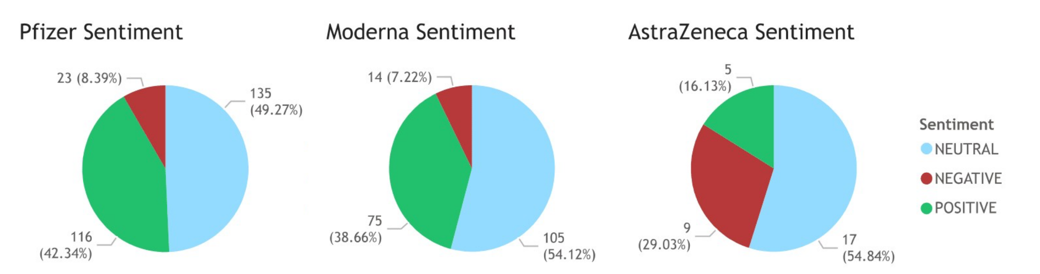
Pre-processing steps with lower effectiveness:
- Removal of steps #2 and #6 resulted in a slight drop in performance.
  - Steps mentioned were to handle emojis, emoticons and negation

Insights

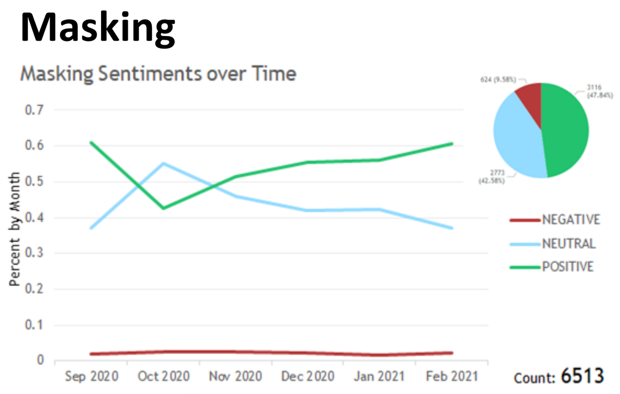
Vaccine



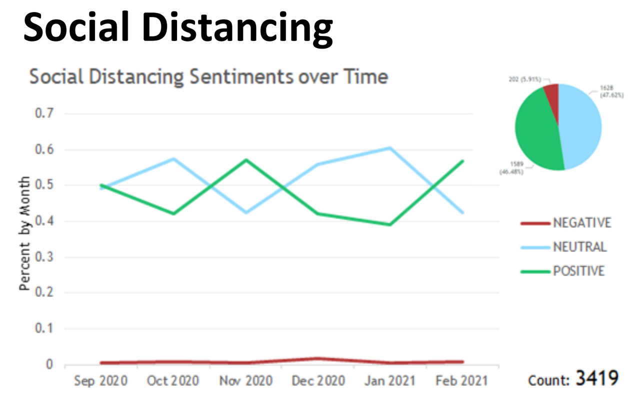
Specific Vaccine Manufacturers



Masking



Social Distancing



Project Challenges

Tweet Labels
- Very subjective -> sometimes requires contextual knowledge to detect sarcasm
  "@realdonaldtrump bottoms up. Drink up, orange!"
- Inter Annotator Agreement needed – model only as good as your labels

Conclusion

- Important to tailor your pre-processing and feature engineering steps to the dataset
- Using relevant datasets for training classifiers will result in better performance
- Future improvements can include the use of contextual word embeddings and attention mechanism to consider the entire context of the tweets.

[1] D. Jurafsky and J.H Martin. 2020. Speech and Language Processing (3rd Ed.). Accessed: Mar 16, 2021.
[2] Da Silva, Nadia FF, Eduardo R. Hruschka, and Estevam R. Hruschka Jr. Tweet sentiment analysis with classifier ensembles. Decision Support Systems 66 (2014): 170-179.