# Temporal analysis of infectious tweets:

Clayton Summitt

## Abstract

Currently, there is no tool to track and measure changes in sentiment in tweets that mention infectious diseases and/or vaccination. From a publicly accessible database of tweet ID's, I built a novel pipeline to simultaneously classify tweets by semantic similarity while also classifying the sentiment of the tweet. The pipeline uses state of the art transformers, sentiment classifiers, and clustering algorithms to detect nine distinct clusters of tweets in a corpus of 495,065 tweets, trained on 1,832,669 tweets. Sentiment with the highest count recorded was neutral.

One shortcoming of the models employed today for classification on COVID19 related data is a lack of consistent twitter data on infectious disease preceding the outbreak in 2019. The current conversations regarding vaccines on twitter are not new, and my project is different in that it takes a longer term temporal look at the conversation on Twitter.

## 1 Introduction

The conversation around microorganisms is not new nor is it narrow in scope. For example, there are preserved written communications, from 1763, between the British General Jeffery Amherst and Colonel Henry Bouquet, devising a plan for the British army to spread smallpox to the indigenous community with blankets(Ranlet 2000). In 1793, a yellow fever outbreak in Philadelphia caused many of the residents to flee the nation's capital, including George Washington, who wrote to his colleagues about the disease outbreak and the impact it had on the fledgling government(Smith 1996; Foster et al. 1998). There are numerous writings around the first vaccine from both proponents, such as Dr. Jenner, the inventor of vaccination, and detractors in the established medical community(Smith 2011). The medical profession, infectious disease treatments, vaccines, and technology have evolved. Although vaccines are now heralded as having saved more lives than any other treatment, the public discourse about the way to treat infectious diseases has changed little since the time of Jenner(Vetter et al. 2018). Some rhetoric remains the same, "more people are now killed by vaccination than by ...." , a statement that has been promoted in 2021 as it was 1920, and still as wrong then as it is now.(Bloom et al. 2017). The history of written communication regarding infectious diseases and more recently, vaccination, provides a compelling reason to monitor trends in the public discourse and note changes in attitudes in real time.

At the end of 2019, there were discussions amongst virologists about a viral outbreak in the Wuhan province of China. The 2019 outbreak was linked to a novel coronavirus related to the SARS-CoV-1 virus that killed 495 people in 2003, though with novel mutations that allowed for easier human-to-human transmission (File et al. 2003; Harrison et al. 2020). The explosive spread of SARS-CoV-2 around the globe also had a parallel growth in the number of tweets from around the world related to COVID-19, and the effect the pandemic has had on the world. With almost 2 years having passed since the outbreak began, there is now an extremely large corpus of tweets related to the subject of infectious diseases and vaccination to work with.

Due to Twitter's large user base, and ease of tweet gathering based on simple keyword matches through the twitter API, multiple large databases have been constructed containing the unique twitter IDs(Banda et al. 2020). I was interested in examining one large corpus of 50 million tweets that was previously used in network analysis, to gain an insight into the temporal trends on the broader subject of infectious disease and ultimately the sentiment expressed by people around the world on this subject, and more narrowly on vaccines(He et al. 2021).

Tweets also provide unique challenges to Natural Language Processing (NLP). For example, datasets of tweets are largely unlabeled, the text is often short and often does not conform to grammar rules found in books, blogs or other social media posts. These challenges with the ability to filter and subset the global corpus to specific keyword terms provides an opportunity to monitor temporal trends and use transfer learning from state of the art models to classify tweets using unsupervised learning and sentiment analysis in concert with classification. I also chose to use an unsupervised approach to this problem as creating robust labeled datasets is noted by many researchers, expensive and time consuming(Gencoglu 2020; Schonfeld et al. 2020; Rosenthal et al. 2017). Using an unsupervised approach with autoencoder decoders may alleviate that bottleneck.

## 2 Background

Analysis of tweets for infectious disease using unsupervised, semi supervised or supervised classification techniques have been performed on single diseases/viruses such as human papillomavirus (HPV) , COVID-19 and influenza(Gencoglu 2020; Luo et al. 2019; Wakamiya et al. 2018; Park et al. 2020). Since the emergence of SARS-CoV-2, there has been an explosion of research on using social media texts, combined with NLP tasks to understand a wide range of topics. Few if any have attempted a temporal approach to understand the public's sentiment to infectious disease severity and vaccination.

## 3 Methods

### 3.1 Data

He et al provide a publicly downloadable archive of Twitter tweet ids, in weekly archives that have been gathered between October 30th, 2016, to April 24th, 2021 on the subject of infectious diseases and vaccines(He et al. 2021). These tweets had been collected from the Twitter streaming api using the following keywords: Measles, MMR, pertussis, DTaP, TDap, chicken pox, (contains:vaccin), varicella, whooping cough, influenza, (contains:polio), rotavirus, pneumococcal, pneu-c-13, hepatitis, meningococcal, HPV, flu, (contains:immuniz), (contains:immunis), cholera, ebola, papillomavirus, diphtheria, H1N1, H5N1, H7N9, H3N2, HIV, malaria, mumps, chickenpox, rubella, RSV, typhoid, (autism (jab, (contains:shot), needle, (contains:vacc), (contains:immuni))). Extraction of tweet text and metadata from tweet ids was performed by me using the tweepy package and a modified script from the panacea lab git repo[1]. To download and hydrate a large corpus of tweets, a twitter student researcher developer account was applied for and granted access with a limit of 5 million tweet downloads per month. Before downloading tweets, each weekly file was randomly shuffled and split into four. This was done to provide equal representation of their respective weekly tweets in the dataset and to allow for future pulls from theTwitter API. Retweets were excluded from the dataset, providing ~3.2 million unique tweets from the 10 million downloaded. Tweets were cleaned to remove usernames, urls, newline characters and pound symbols.

### 3.2 Tweet Embeddings

My goal is to group semantically similar tweets together and perform sentiment analysis on individual tweets. I chose SBERT, a pretrained state of the art transformer network that exceeds BERT on Semantic Textual SImilarity and outperforms BERT networks using the SentEval network, as most tweets resemble a single sentence(Reimers and Gurevych 2019). Using a multi anguage model, SBERT was pre-trained with 1,832,669 tweets using a masked language model, with 15% masked word rate, for 3 epochs. Once pre-trained, 495,065 tweets, a selected subset of tweets that contained either the root vacc or vax (to capture tweets that include vaccine, vaccinated, vaxxed etc) were embedded. The max sequence length for all embeddings was equivalent to the longest tweet, which was 43 words.

### 3.3 Auto Encoder Decoder

Recent work by Wang provides state of the art results for labeling unlabeled data in an unsupervised approach using a Auto-Encoder(Wang et al. 2021). The SBERT embedder produces embeddings that are a flat vector with a dimension of 768. To reduce the dimensionality of the vector space I used an Autoencoder decoder (AE). The AE uses the same vector as both the input and target during training with the capacity to

[1] https://github.com/thepanacealab/covid19_twitter

[2] https://www.kaggle.com/datasciencetool/covid19-vaccine-tweets-with-sentiment-annotation.

learn how to predict their inputs from a reduced vector space.Once triained, the model can be used to extract the middle layer from new inputs, reducing the dimension of the input (768) to the middle layer output (32), avoiding the pitfalls present in PCA analysis, maintaining pairwise relationships during dimensionality reduction(McInnes et al. 2018). The AE processes through two deep layers (128 units then 32 units) followed by two deep layers (128 units and 768) (figure 1). Each layer used a tanh activation function with a Stochastic Gradient Descent optimizer with an initial learning rate = 0.001, employing a decay rate 0f 0.96 for every 10000 steps, which helps with both optimization and generalization by having the network avoid noisy data in the early phase, resulting in improved pattern learning(You et al. 2019). Binary cross entropy was the loss function used to train the model. This function minimizes the mean error (as a probability) between the target and the predicted label for each dimension in the vector. Isplit the training data 80:20 and used ~400000 tweets as a validation set. After training the AE, I extract the middle layer of size 32 on our test data for grouping.

### 3.4 Sentiment

Parallel to the sentence embeddings, I used a pre-trained BERT model, tweeteval, which was trained on 200 million tweets for sentiment classification tasks(Rosenthal et al. 2017). I fine tuned the pretrained model using a labeled training set of 5603 tweets[2] and then used the pretrained model to predict sentiment of 450K tweet test set. Outputs from the model include the probability of a negative (0), neutral (1) or positive (2) sentiment where the max probability is assigned as the predicted sentiment score.

### 3.5 Grouping

After AE and sentiment analysis, I utilize the Accelerated Hierarchical Density Based Clustering algorithm to group tweets together based on their reduced similarity (HDBscan). This algorithm finds subsets of the data which group together but does not force points into clusters for increased accuracy of clustering compared to other unsupervised methods, specifically K-means(McInnes and Healy 2017). Advantages of this algorithm include that it does not require selecting the number of clusters ahead of time, does not assume gaussian distribution, and does not assign outliers to clusters, which K-means clustering does. To visualize the groupings UMAP was used to reduce the 32 dimension space to 2 dimension, coloring of points was done by cluster and again by sentiment score(McInnes et al. 2018). I randomly chose 80000 AE outputs and fed them into a HDBscan model for fitting, after fitting, a the rest of the points were assigned predicted groupings. The grouping output was used by UMAP to color points by cluster assignment. The HDBscan model was initialized to a minimum cluster size of 20, a metric of manhattan distance. Outliers were removed for clarity See figure 2.

## 4 Results

The main results from this project can be seen in figure 2. Using 1.8 million tweets as a training set, with 495,065 tweets as a test set, without filtering for language, I was able to cluster a random sample of 80000 tweets into nine distinct clusters. Sentiment scores are more frequently evaluated as being neutral (figure 3).

Though randomly sampled, there may is a chance of bias in the data, as the number of tweets related to Covid-19 and the debates surrounding vaccines has swamped out the trends of the previous 4 years. Moving forward, the next logical step to this project would be to build a topic modeling algorithm to quickly identify topics and name topics to provide testable labels.

## 4.1 Discussion

Sentiment analysis that is limited to only a positive, neutral or negative, in my opinion, may not capture relevant information in the tweets that is needed to measure changing attitudes towards vaccines. Additionally it is challenging to measure model performance on unlabeled datasets. Many broad scope models use evaluation sets to test the performance, there are few large datasets that with such narrow focus as infectious disease.

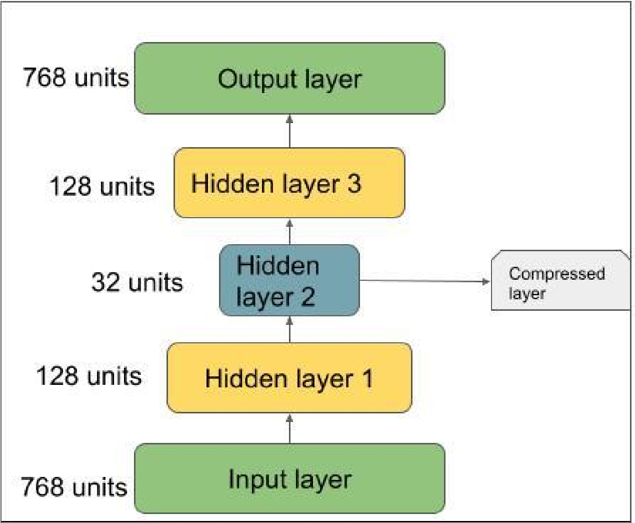

Figure 1: Stacked auto encoder layout.

[1] https://github.com/thepanacealab/covid19_twitter

[2] https://www.kaggle.com/datasciencetool/covid19-vaccine-tweets-with-sentiment-annotation.

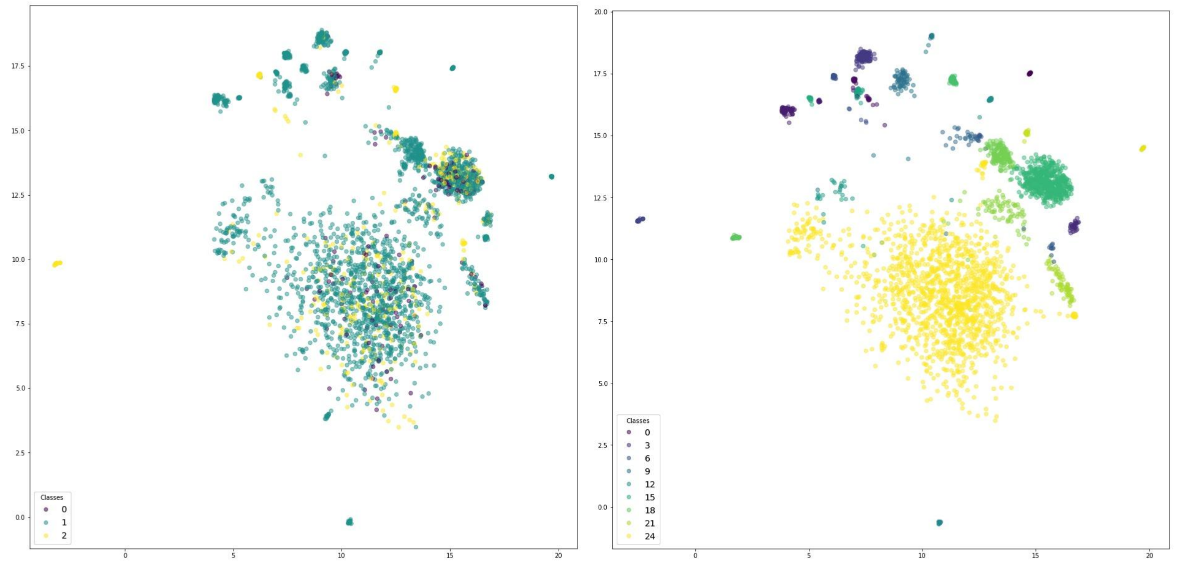

**Figure 2**(l-r): (**l)**Sentiment of clustered tweets, purple is negative, green is neutral, yellow is positive. The majority of tweets are identified as having neutral sentiment.
(**r**) UMAP 2D representation of tweet clusters projected down from 32D space, tweets that belong to no cluster have been removed for clarity. Each color represents a distinct cluster.

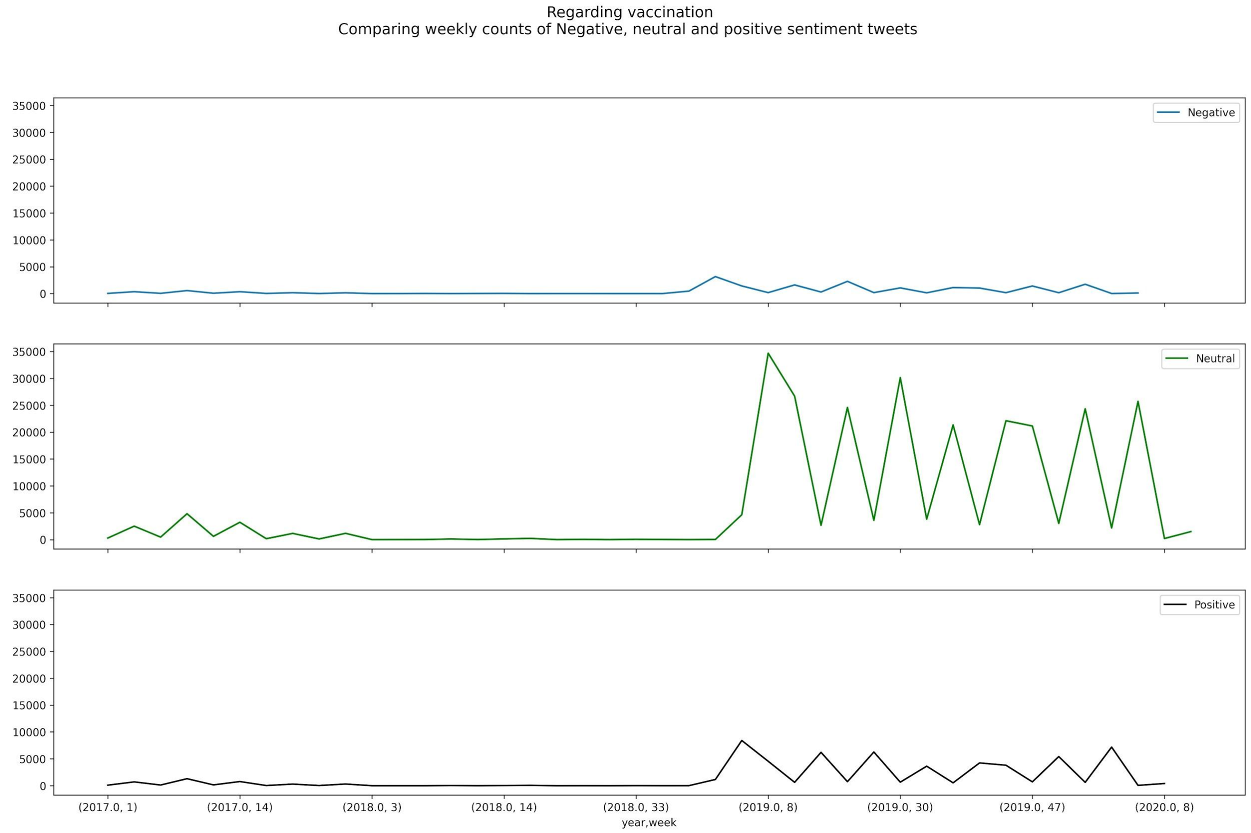

Figure 3: Tweets collected and categorized, were predominantly neutral in sentiment.

[1] https://github.com/thepanacealab/covid19_twitter

[2] https://www.kaggle.com/datasciencetool/covid19-vaccine-tweets-with-sentiment-annotation.

# Works Cited

Banda, Juan M., et al. “A Large-Scale COVID-19 Twitter Chatter Dataset for Open Scientific Research—An International Collaboration.” *Epidemiologia*, vol. 2, no. 3, 2021, pp. 315–324., doi:10.3390/epidemiologia2030024.

Cinelli, Matteo, et al. “The COVID-19 Social Media Infodemic.” *Scientific Reports*, Nature Publishing Group UK, 6 Oct. 2020, dx.doi.org/10.1038/s41598-020-73510-5.

Foster, Kenneth R., et al. “The Philadelphia Yellow Fever Epidemic of 1793.” *Scientific American*, vol. 279, no. 2, 1998, pp. 88–93., doi:10.1038/scientificamerican0898-88.

Gencoglu, Oguzhan, and Mathias Gruber. “Causal Modeling of Twitter Activity During COVID-19.” *ArXiv.org*, 23 Sept. 2020, arxiv.org/abs/2005.07952.

Gencoglu, Oguzhan. “Large-Scale, Language-Agnostic Discourse Classification of Tweets During COVID-19.” *Machine Learning and Knowledge Extraction*, vol. 2, no. 4, 2020, pp. 603–616., doi:10.3390/make2040032.

Harrison, Andrew G, et al. “Mechanisms of SARS-CoV-2 Transmission and Pathogenesis.” *Trends in Immunology*, Elsevier Ltd., Dec. 2020, www.ncbi.nlm.nih.gov/pmc/articles/pmc7556779/.

He, Zitao, et al. “A Collection of Tweets Related to Climate Change/Infectious Diseases and Vaccines, Broken down by Week.” *Dryad Data -- A Collection of Tweets Related to Climate Change/Infectious Diseases and Vaccines, Broken down by Week*, Dryad, datadryad.org/stash/dataset/doi:10.5061/dryad.djh9w0w05.

Luo, Xiao, et al. “A Natural Language Processing Framework to Analyse the Opinions on HPV Vaccination Reflected in Twitter over 10 Years (2008 - 2017).” *Human Vaccines &amp; Immunotherapeutics*, vol. 15, no. 7-8, 2019, pp. 1496–1504., doi:10.1080/21645515.2019.1627821.

Mcinnes, Leland, and John Healy. “Accelerated Hierarchical Density Based Clustering.” *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, 2017, doi:10.1109/icdmw.2017.12.

Mcinnes, Leland, et al. “UMAP: Uniform Manifold Approximation and Projection.” *Journal of Open Source Software*, vol. 3, no. 29, 2018, p. 861., doi:10.21105/joss.00861.

Nistor, Sergiu Cosmin, et al. “Building a Twitter Sentiment Analysis System with Recurrent Neural Networks.” *Sensors*, vol. 21, no. 7, 2021, p. 2266., doi:10.3390/s21072266.

Park, Han Woo, et al. “Conversations and Medical News Frames on Twitter: Infodemiological Study on COVID-19 in South Korea.” *Journal of Medical Internet Research*, vol. 22, no. 5, 2020, doi:10.2196/18897.

Reimers, Nils, and Iryna Gurevych. “Making Monolingual Sentence Embeddings Multilingual Using Knowledge Distillation.” *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, doi:10.18653/v1/2020.emnlp-main.365.

Reimers, Nils, and Iryna Gurevych. “Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks.” *ArXiv.org*, 27 Aug. 2019, arxiv.org/abs/1908.10084.

Reimers, Nils, and Iryna Gurevych. “Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks.” *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, doi:10.18653/v1/d19-1410.

Rosenthal, Sara, et al. “SemEval-2017 Task 4: Sentiment Analysis in Twitter.” *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, 2017, doi:10.18653/v1/s17-2088.

Rosenthal, Sara, et al. “SemEval-2017 Task 4: Sentiment Analysis in Twitter.” *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, 2017, doi:10.18653/v1/s17-2088.

Schonfeld, Justin, et al. “Debates about Vaccines and Climate Change on Social Media Networks: a Study in Contrasts.” 2020, doi:10.1101/2020.11.24.396226.

Smith, Kendall A. “Edward Jenner and the Small Pox Vaccine.” *Frontiers in Immunology*, vol. 2, 2011, doi:10.3389/fimmu.2011.00021.

Smith, Mark A. “Andrew Brown’s ‘Earnest Endeavor’: The ‘Federal Gazette’’s Role in Philadelphia’s Yellow Fever Epidemic of 1793.” *The Pennsylvania Magazine of History and Biography*, vol. 120, no. 4, Historical Society of Pennsylvania, 1996, pp. 321–42, http://www.jstor.org/stable/20093070.

Stanik, Christoph, et al. “Unsupervised Topic Discovery in User Comments.” *ArXiv.org*, 19 Aug. 2021, arxiv.org/abs/2108.08543v1.

Stanik, Christoph, et al. “Unsupervised Topic Discovery in User Comments.” *2021 IEEE 29th International Requirements Engineering Conference (RE)*, 2021, doi:10.1109/re51729.2021.00021.

Stevens, Philip. *Fighting the Diseases of Poverty*. Routledge, 2017.

Vetter, Volker, et al. “Understanding Modern-Day Vaccines: What You Need to Know.” *Annals of Medicine*, vol. 50, no. 2, 2017, pp. 110–120., doi:10.1080/07853890.2017.1407035.

Wakamiya1*, Shoko, et al. “Twitter-Based Influenza Detection After Flu Peak via Tweets With Indirect Information: Text Mining Study.” *JMIR Public Health and Surveillance*, JMIR Publications Inc., Toronto, Canada, publichealth.jmir.org/2018/3/e65/.

Wang, Kexin, et al. “TSDAE: Using Transformer-Based Sequential Denoising Auto-Encoder for Unsupervised Sentence Embedding Learning.” *ArXiv.org*, 10 Sept. 2021, arxiv.org/abs/2104.06979.

[1] https://github.com/thepanacealab/covid19_twitter

[2] https://www.kaggle.com/datasciencetool/covid19-vaccine-tweets-with-sentiment-annotation.

Wang, Kexin, et al. “TSDAE: Using Transformer-Based Sequential Denoising Auto-Encoder for Unsupervised Sentence Embedding Learning.” *ArXiv.org*, 10 Sept. 2021, arxiv.org/abs/2104.06979.

Yang, Yinfei, et al. “Multilingual Universal Sentence Encoder for Semantic Retrieval.” *ArXiv.org*, 9 July 2019, arxiv.org/abs/1907.04307.

You, Kaichao, et al. “How Does Learning Rate Decay Help Modern Neural Networks?” *ArXiv.org*, 26 Sept. 2019, arxiv.org/abs/1908.01878.