Spring 6-1-2021

Exploring the Use of Social Media to Infer Relationships Between Demographics, Psychographics and Vaccine Hesitancy

Abhimanyu Kapur
Dartmouth College, Abhimanyu.Kapur.21@Dartmouth.edu

EXPLORING THE USE OF SOCIAL MEDIA TO INFER RELATIONSHIPS BETWEEN DEMOGRAPHICS, PSYCHOGRAPHICS AND VACCINE HESITANCY

A Thesis

Submitted to the Faculty

in partial fulfillment of the requirements for the

degree of

Bachelor of Arts

in

Computer Science

by

Abhimanyu Kapur

Advised by

Professor Soroush Vosoughi, Department of Computer Science

DARTMOUTH COLLEGE

Hanover, New Hampshire

June 2021

Abstract

The growing popularity of social media as a platform to obtain information and share one's opinions on various topics makes it a rich source of information for research. In this study, we aimed to develop a framework to infer relationships between demographic and psychographic characteristics of a user and their opinion on a specific narrative - in this case, their stance on taking the COVID-19 vaccine. Twitter was the chosen platform due to the large USA user base and easily available data. Demographic traits included Race, Age, Gender, and Human-vs-Organization Status. Psychographic traits included the Big Five personality traits (Conscientiousness, Neuroticism, Openness, Agreeableness, Extraversion), Risk Seeking, Risk Aversion, Inward Focus, and Outward Focus. Our pipeline involved preprocessing the data, labeling tweets as vaccine-hesitant using distant supervision, training a vaccine hesitancy classifier to classify a second dataset, obtaining demographic and psychographic inferences for each user, and finally running a logistic regression with vaccine hesitancy as the dependent variable and sets of demographic and psychographic characteristics as the independent variable. We achieved an F1 score of 0.947 for our classifier and found statistically significant trends in vaccine hesitancy for race, age, gender, and human-vs- organization status. On the other hand, there were no significant relationships between any of the psychographic traits and vaccine hesitancy. It should be noted that this study was not pre-registered and the values for all variables (dependent and independent) come from noisy classifiers. As such, these results should only

be viewed as a preliminary analysis of the demographic and psychographic factors correlated with vaccine hesitancy. We conclude that such a framework is a useful tool to identify the relations between different demographics and popular narratives. Further work and better data are necessary to improve the framework to the point where the strength of the correlations can be considered and not just the overall relationships. Furthermore, while psychographic traits yielded no significant results, there were several limitations in their inference, and focusing on improving psychographic trait inference is an important avenue for future studies.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Section 1.1

Problem Statement

Social media has become an integral part of the lives of people worldwide. With more than 3.5 billion people across the globe using some form of social media[34], it is fast proving to be an important medium for communication and a tool to understand the opinions and preferences of different populations. Studies have shown the importance of social media for consumption and diffusion of news[29][28], as well as understanding and shaping public opinion[16][18][7]. We are particularly interested in the latter, that is understanding public opinion on various topics, and going a step beyond to see whether we can understand how different populations perceive different issues. Therefore, not only are we interested in ascertaining a user group's views on a topic but also in identifying their demographics and/or psychographics from their social media interactions.

Given the outbreak of SARS-CoV-2 (henceforth referred to as COVID-19) in 2020 and the anti-vaccine movement against the COVID-19 vaccines released in 2020-21, we felt this served as an opportunity to explore whether it would be possible to infer relationships between demographics/psychographics and hesitancy to take the vaccine. We aim to build a framework to enable such an exploration that could then be expanded and further used.

The merits of such an exploration are clear in the current climate. If we can identify populations through social media who might be more hesitant to take vaccines, this can aid policymakers as well as public health officials to target these populations and optimize the allocation of resources for maximum impact. Furthermore, such techniques could have wide-ranging applications beyond public health including policymaking, election prediction, advertising, etc. Our aim was to find and establish a framework that would allow us to find relationships within the social media studies and hence the results for specific trends and relationships should be viewed as a preliminary analysis.

Section 1.2

Prior Work

During the study, we did not find there had been prior research in building a framework for using social media to analyze differences in response to a narrative as this study aims to; however, the individual components of the framework would not be possible without the work of many researchers before us.

There are numerous studies on understanding ideological positions via Twitter and other social media. We were particularly inspired by Barberá's 2015 study on estimating Twitter users' ideologies[5], as well as Kosinski et al's work on predicting private traits and attributes from Facebook likes[21]. When searching for existing research on the possibility of demographic and psychographic inference, Ito et al's work on demographic and psychographic estimation of Twitter users using social structures was particularly helpful and showed that the user profile tends to be more important than their tweets for such estimation[17]. Furthermore, applying distant supervision techniques to Twitter is not a novel approach. Go, Bhayani & Huang, introduced the approach of using distant supervision for initial labelling (using emoticons for sentiment classification) and then classification on Twitter[13]. We also found it useful to follow Marchetti-Bowick and Chambers' efforts in harvesting key terms and then using those for initial tagging, beyond a binary approach[26].

Chapter 2

Data

Section 2.1

Twitter Data

Twitter was chosen as the social media to look at as it has a sizable American user base (22% of the US population[37]) as well as a wealth of datasets available and an easy to use API for querying. The primary data object interacted with is the Tweet object which can be found in Appendix A. Two Twitter datasets were used in the study as described below.

2.1.1. Dataset 1

The first dataset was retrieved from Kaggle and consists of tweets with the hashtag (#CovidVaccine) which allowed us to focus on COVID-19 related vaccine tweets [20]. As a user collected dataset it did not include all the features of the Twitter Tweet object, instead it consisted of the following features:

- user_name
- user_location
- user_description

- user_created
- user_followers
- user_friends
- user_favourites
- user_verified
- date
- text

This dataset was used to train and test the vaccine hesitancy classifier.

Date Range	Number of Tweets	Number of Unique Users
2020-07-31 : 2020-09-16	102347	54751

Table 2.1: Descriptive Statistics, Dataset 1

2.1.2. Dataset 2

The second dataset was a daily updated dataset of geo-tagged tweets about COVID-19 [22]. Following the new Twitter content redistribution policy, it only consisted of tweet IDs which were then hydrated. The geo-tagging allowed us to use geographic features in our inferences, and the recency of the data, as compared to Dataset 1, provided us with the latest tweets as vaccines began rolling out across the world. The broader range of hashtags and topics queried to form the dataset provided us with a broader base of tweets to run inferences on, compared to Dataset 1.

Date Range	Number of Tweets	Number of Unique Users
2020-10-01 : 2021-04-15	106126	42280

Table 2.2: Descriptive Statistics, Dataset 2

Section 2.2

Geographic Data

Additionally, a geographic dataset was used to map user locations to US states. We retrieved a dataset from the US Census Bureau which consisted of a list of all US states and territories as well as their geographic boundaries as polygons of lists of (latitude, longitude) pairs[8].

Chapter 3

Methodology

Once the datasets (dataset 1 and 2) had been identified and hydrated, we began by labelling each Tweet object as pro-vaccine or anti-vaccine using distant supervision. In this process, we identified hashtags that served as signals for either pro/anti-vaccine. This information is stored in the *is_vaccine_hesitant* field. Next, the tweet text and user description data was preprocessed, following which two separate encodings were generated using sentence-BERT and scikit-learn's TF-IDF vectorizer. We then trained a variety of classifiers (as described below) to predict the vaccine hesitancy of a given tweet object using either feature set (a) the encoded tweet text; or (b) the encoded tweet text, encoded user description, follower count (raw + log), following count (log + count). After identifying feature set (a) with TF-IDF encodings and AdaBoost Classifier as the best combination, we trained the classifier on the entire dataset 1 and then used this classifier to label dataset 2. The next task was to obtain the demographic and psychographic information, for which we used the following packages: *Ethnicolr*, *M3 Inference*, *Receptiviti* API. For geographic features, we used a dataset from the US Census Bureau to map user locations to the state they were in and then to the corresponding geopolitical region. Finally, we ran a logistic regression with vaccine hesitancy as the dependent variable and the demographic and psychographic data as independent variables.

Section 3.1

Pipeline

The pipeline for this study can be seen in Figure 3.1

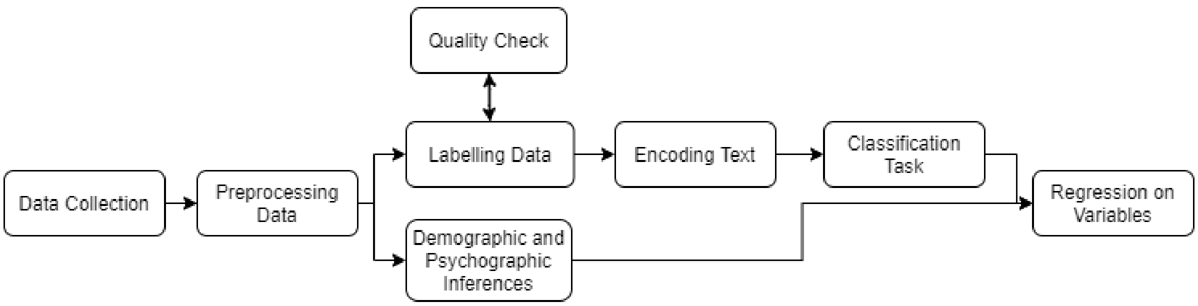

Figure 3.1: Pipeline

Section 3.2

Hydrating Tweets

Dataset 2 is a list of Twitter IDs that require hydration. Hydration is the process of querying Twitter's API with a tweet ID to retrieve the entire tweet object. The program Hydrator was used for the process.

Section 3.3

Preprocessing

Our preprocessing algorithm did the following to the data:

- Removed duplicate tweets
- Removed stopwords from the tweet text

- Removed URLs from the tweet text
- and replaced them with a text token - “URL”
- Removed extra white spaces and punctuation, and applied lower casing to the tweet text
- Lemmatized the tweet text

As an example, a tweet text before and after preprocessing is seen below:

Before Preprocessing:

#DNA zooms up charts in 1st week; hear #vaccines episode: https://t.co/oDrayhi7zN . #pandemic , #COVID19 , #CovidVaccine

After Preprocessing:

“*dna zoom chart 1st week hear vaccine episode URL pandemic covid 19 covidvaccine*”

While we remove hashtags from the text, hashtag information is not lost as it is stored as a separate property of the Tweet object.

Section 3.4

Encoding

For the classification text, we needed to vectorize the text features: the tweet text and user descriptions. Two methods were tried, TF-IDF and BERT.

3.4.1. TF-IDF

TF-IDF (Term Frequency - Inverse Document Frequency) are scores given to all words in a set of documents. Term Frequency refers to the frequency of a word within a given document and Inverse Document Frequency refers to the down-weighting of words that appear in multiple documents. Thus when we fit the TF-IDF vectorizer (as implemented in *scikit_learn*) to our dataset, it tokenizes the set of documents (in our case the individual tweets) and builds a weighted vocabulary. Then we can transform (vectorize) the training and testing sets and build the classifier.

3.4.2. BERT

BERT (Bidirectional Encoder Representations from Transformers), is a state of the art language model[9] that encodes the contextual relations between words and can learn the context of a word based on the surrounding words and allows for powerful semantic encodings. In this study, we used SentenceTransformers to encode our tweet text. SentenceTransformers is a package that provides a number of pre-trained models based on Sentence-Bert[30].

Section 3.5

Classifier

3.5.1. Getting Initial Vaccine Hesitancy

In the absence of existing datasets and models, a rule-based approach was used to tag tweets as vaccine hesitant to form training data for the classification task. Lists of pro-vaccine and anti-vaccine hashtags were first assembled using previous research and first-hand browsing of anti-vaxxer twitter[19]. A study on vaccine opposition on Twitter by Bonnevie et al[6] was particularly useful, as it provided a list of stand-alone and co-occurring words related to vaccine opposition.

As there were a significant number of pro-vaccine posts that also used anti-vaccine terminology to discuss or berate the ideology, tweets were first checked for “pro-vax” hashtags and tagged as non-vaccine hesitant (Class 0). Then, the remaining tweets were tagged as vaccine hesitant (Class 1) if they contained anti-vaccine sentiment, else not.

Attempts to augment the initial labelling with sentiment analysis were considered but discarded as it did not make a significant difference to results.

Pro-Vaccine Hashtags

- getvax
- immunization
- vaccinate
- covidiots
- vaccinesforall
- igottheshot
- vaccinateyourkids
- vaccination
- vaccinations
- vaccine
- vaccinessavelives
- vaccineswork

Anti-Vaccine Hashtags

- cdcfraud
- autism
- cdcwhistleblower
- CovidHoax
- IWillNOTComply
- 5g
- markofthebeast
- idonotconsent
- microchip
- b1less
- breakabillion
- hearthiswell
- vaccineinjury
- firefauci
- antivaxmovement
- learntherisk
- NoForcedVaccination
- StopForcedVaccination
- fuckvaccines
- exposed
- vaccinechoice
- vaccineskill
- antivax
- VaXXedII
- Vaccinesuncovered
- vaccinefailure
- vaccineprotection
- VaccinesCauseAIDs
- VaccinesCauseAutism
- vaccinescausesids
- vaccinesdangers
- vaccineskillandmaim
- vaccinetruth
- anti-flu-vaccine
- anti-vacc
- anti-vaccination
- anti-vaccine
- anti-vaccines
- anti-vax
- anti-vaxer
- anti-vaxers
- antivaccination
- antivaccine
- antivaccines
- unvaccinated
- unvax
- unvaxed
- StopMandatoryVaccination
- VaccinesCauseRegressiveAutism

After the labelling task, Dataset 1 was then sampled to reduce class imbalance which left us with a dataset of 600 tweets - 200 in Class 1 and 400 in Class 0.

3.5.2. Making Predictions

We tested the following classifiers (implemented using the standard *sklearn* library) :

- Multi-Layer Perceptron (MLP)
- Support Vector Machines (SVM)
- Random Forest
- Gaussian Naive Bayes

With the following feature sets:

- Encoded tweet text
- Encoded tweet text, encoded user description, follower count (raw and log), following count (raw and log), verified status

Note: Due to time constraints we were only able to test TF-IDF encoded classifiers with the first feature set.

After classifying the dataset using the vaccine hesitancy classifier trained on Dataset 1, we then removed all tweets from outside the United States of America. The US-only dataset contained 47327 tweets and 21727 unique users. All further work was conducted on this dataset.

Section 3.6

Obtaining Demographic Data

To obtain demographic data for the Twitter users, we utilized external libraries which predicted race, age and gender[1] information as well as whether a user is an individual

[1]In this thesis, we use the terms gender and sex interchangeably

human or an organization.

3.6.1. Race

To predict the race of a user, we used a package, *Ethnicolr*[23] which utilizes a user's last name and census data to predict the probability of them belonging to the following races: black, white, asian or pacific islander (api), hispanic.

3.6.2. Age, Gender & Human-vs-Organization Status

To obtain age, gender and human-vs-organization status information from user tweets we used an implementation of the M3 (Multimodal, Multilingual, and Multi-attribute) system, which is a deep learning system for demographic inference from Twitter data [36]. The model uses user profile pictures as well as the following fields to predict user demographic information:

- Name
- Screen Name
- Language
- Description

We used the text-only model as not all users had profile pictures in our datasets.

3.6.3. Location

All users in Dataset 2 had coordinates for their tweet texts. We followed the assumption that users would be tweeting from their home state and using a US Census bureau dataset, were able to map each user's tweet location to the corresponding American state. We then divided the US into four geopolitical regions based on a Gallup poll demarcation [35] :

- The West/East Coast Alliance
- The Heartland/New South Alliance
- The Blue Collar Midwest
- Other

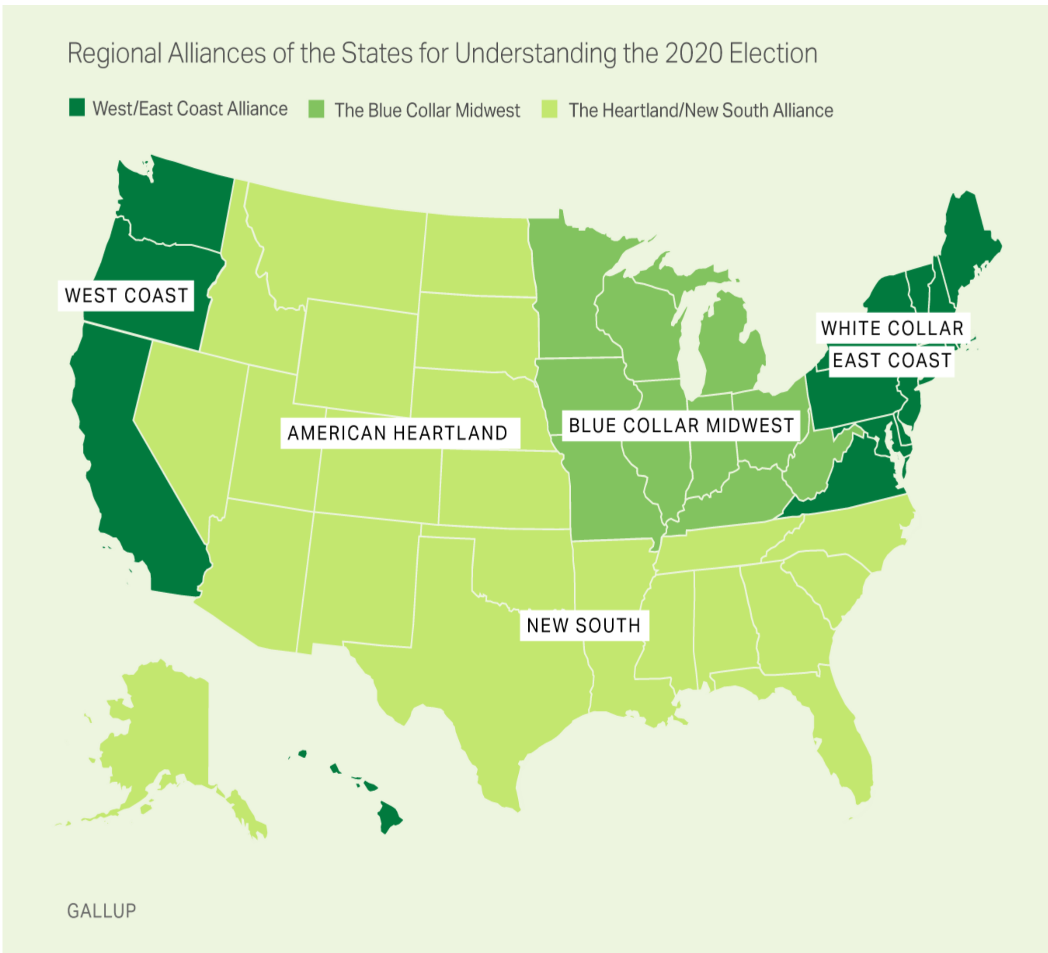

Figure 3.2: Map of US States Classified According to Gallup's Demarcation[35]

The Gallup classification only included the first three categories. As US territories such as Puerto Rico were left out of this classification system, we added an 'Other' category to capture their data.

Section 3.7

Obtaining Psychographic Data

The psychographic metrics used (as listed below) were obtained using an external service API, Receptiviti[1].

- Big Five Personality Traits
  - Conscientiousness
  - Neuroticism
  - Openness
  - Agreeableness
  - Extraversion
- Risk Seeking
- Risk Aversion
- Inward Focus
- Outward Focus

Traditionally such traits would be measured through a self-assessment inventory or clinically administered test. As that would be challenging to gather over social media, we rely on Receptiviti's alternate approach which they term a "Language-Based Personality" framework. A clinical personality assessment reflects a person's self-perception, while Receptiviti claims that their approach better reflects how a person's personality is perceived by others and thus is a more objective evaluation [2].

To obtain the scores for the psychographic measures, the tweet texts were passed to the API for analysis, and user scores were averaged across their individual tweets' scores. To understand how Receptiviti does the scoring, we turn to the documentation:

> Measures in the Language-Based Personality framework are always in the range of 0 to 100. Our measures are baselined against our proprietary personality datasets, which are comprised of hundreds of thousands of personality-labelled language samples that exceed 350 words. A language sample that generates a score of 80 on any Personality measure implies that 80% of all samples in our curated baseline dataset have scores that are less than the score of language sample being analyzed[2].

Section 3.8

Regression Task

The final step was the regression task. After demographic and psychographic inferences were made, they were averaged per user and we were left with a dataset of 21727 tweets (1 per user) with all the inferences. This dataset was used in the regression task.

We used the *statsmodel* package to run a logistic regression with vaccine hesitancy as the dependent variable and the following sets of psychographic and demographic factors as the independent variables:

(a) Age Only

(b) Location Only

(c) Race Only

(d) All Psychographic factors (Big 5 Personality traits, Risk Seeking, Risk Aversion, Inward Focus, Outward Focus)

(e) All Demographic Factors (Race, Age, Geographic Location, Human-vs- Organization Status)

(f) All Psychographic and Demographic Factors

For each of the results, the confidence intervals and odds ratio were then calculated. The results and our interpretation is provided in the next sections. Additionally, we also attempted to make feature data (for race, age, inward/outward focus and risk seeking/aversion) binary and run logistic regression. Given the noisiness of the classifier and the accuracy of the demographic and psychographic identifiers, we felt this set of results were not an accurate representation of the data.

Chapter 4

Results

It should be noted that this study was not pre-registered and the values for all variables (dependent and independent) come from noisy classifiers. As such, these results should only be viewed as a preliminary analysis of the demographic and psychographic factors correlated with vaccine hesitancy.

Section 4.1

Interpreting Results

In this section, we provide a brief overview of how to understand and interpret the results provided.

Section 4.2

F1 Score and ROC Curve

For our classification task, we use the F1 score as our primary metric; it incorporates both precision and recall scores into one score between 0.0 and 1.0, as seen in the formula below.

$$F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

We also provide the ROC (Receiver Operating Characteristic) curves for the classifiers. These graphs have the False Positive Rate on the X-axis and True Positive Rate on the Y-axis, showing the performance of the classifier at all thresholds.

4.2.1. p-Value

In the final regression task, we provide the p-Value for each relation. The p-Value is a value between 0 and 1 which can be considered a measure of significance of the results. Typically a p-Value less than 0.05 indicates statistically significant results.

4.2.2. Odds Ratio

In the final regression task, we provide the odds ratios for all the variables. An odds ratio is a "measure of association between an exposure and an outcome"[32]. An odds ratio greater than 1 indicates that for a given outcome the odds of the event are decreased, while less than 1 indicates odds are increased [31].

Section 4.3

Classification Results

The results for all classifier tested are listed in Table 4.1.

	Bert: Tweets Only	Bert: All Features	TF-IDF: Tweets Only
Dummy Classifier	0.366	0.347	0.342
Logistic Regression	0.658	0.667	0.769
Random Forest	0.519	0.453	0.919
AdaBoost	0.684	0.684	0.947
MLP	0.711	0.638	0.868
SVC	0.687	0.000	0.812
Gaussian Naive Bayes	0.675	0.494	0.845

Table 4.1: Table of Classification Results

As seen above the best classifier was the AdaBoost Classifier using the TF-IDF vectorizer with the tweet text being the only feature. We present the ROC curve (Figure 4.1), heatmap (Figure 4.2) and class prediction error graph (Figure 4.3) for the same, below.

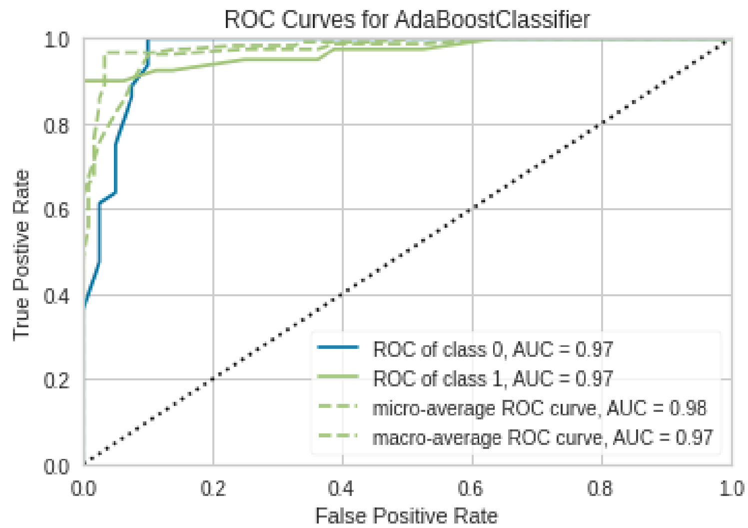

Figure 4.1: ROC Curve

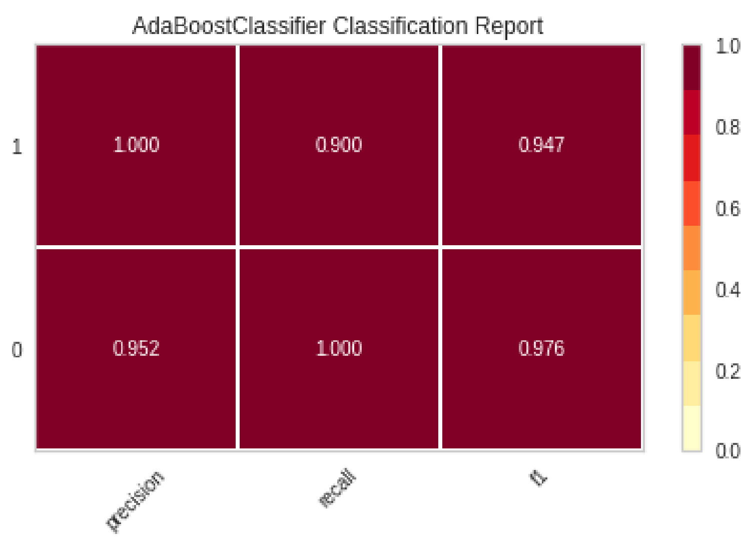

Figure 4.2: Heatmap

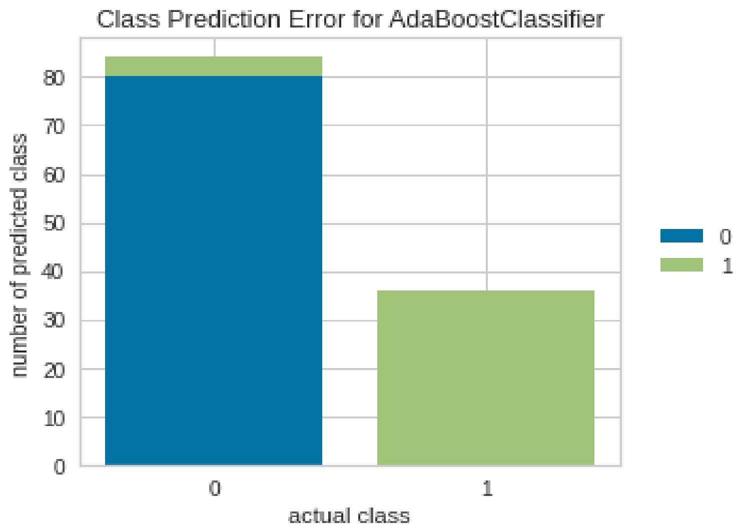

Figure 4.3: Class Prediction Error

Section 4.4

Regression Results

Below we present the results and corresponding odd ratios table from running the regression.

4.4.1. Age

Dep. Variable:	is_vaccine_hesitant	**No. Observations:**	21727
Model:	Logit	**Df Residuals:**	21724
Method:	MLE	**Df Model:**	2
Log-Likelihood:	-8629.9	**Pseudo R-squ.:**	-0.1839
converged:	True	**LL-Null:**	-7289.5
Covariance Type:	nonrobust	**LLR p-value:**	1.000

Table 4.2: Age Results

	coef	**std err**	**z**	**P¿|z|**	**[0.025**	**0.975]**
age_<=18	-3.1637	0.152	-20.795	0.000	-3.462	-2.866
age_19-29	-0.9403	0.125	-7.527	0.000	-1.185	-0.695
age_30-39	-4.1010	0.071	-57.592	0.000	-4.241	-3.961

Table 4.3: Age Results

	5%	**95%**	**Odds Ratio**
age_<=18	0.031371	0.056954	0.042270
age_19-29	0.305694	0.498850	0.390506
age_30-39	0.014400	0.019036	0.016557

Table 4.4: Age: Odds Ratio and Confidence Intervals

4.4.2. Location

Dep. Variable:	is_vaccine_hesitant	**No. Observations:**	21727
Model:	Logit	**Df Residuals:**	21724
Method:	MLE	**Df Model:**	2
Log-Likelihood:	-11507.	**Pseudo R-squ.:**	-0.5785
converged:	True	**LL-Null:**	-7289.5
Covariance Type:	nonrobust	**LLR p-value:**	1.000

Table 4.5: Location Results

	coef	**std err**	**z**	**P¿|z|**	**[0.025**	**0.975]**
blue_collar_midwest	-2.6951	0.076	-35.574	0.000	-2.844	-2.547
heartland_new_south	-1.8172	0.034	-53.748	0.000	-1.883	-1.751
other	-2.1239	0.184	-11.531	0.000	-2.485	-1.763

Table 4.6: Location Results

	5%	**95%**	**Odds Ratio**
blue_collar_midwest	0.058219	0.078349	0.067538
heartland_new_south	0.152068	0.173618	0.162486
other	0.083334	0.171549	0.119565

Table 4.7: Location: Odds Ratio and Confidence Intervals

4.4.3. Race

Dep. Variable:	is_vaccine_hesitant	**No. Observations:**	21727
Model:	Logit	**Df Residuals:**	21724
Method:	MLE	**Df Model:**	2
Log-Likelihood:	-9812.3	**Pseudo R-squ.:**	-0.3461
converged:	True	**LL-Null:**	-7289.5
Covariance Type:	nonrobust	**LLR p-value:**	1.000

Table 4.8: Race Results

	coef	**std err**	**z**	**P¿|z|**	**[0.025**	**0.975]**
black	-16.0026	0.286	-55.920	0.000	-16.563	-15.442
hispanic	-3.1586	0.135	-23.468	0.000	-3.422	-2.895
api	-2.8332	0.173	-16.363	0.000	-3.173	-2.494

Table 4.9: Race Results

	5%	**95%**	**Odds Ratio**
black	6.405850e-08	1.966753e-07	1.122440e-07
hispanic	3.263366e-02	5.530960e-02	4.248476e-02
api	4.189399e-02	8.258870e-02	5.882151e-02

Table 4.10: Race: Odds Ratio and Confidence Intervals

4.4.4. All Psychographics

Dep. Variable:	is_vaccine_hesitant	**No. Observations:**	21727
Model:	Logit	**Df Residuals:**	21718
Method:	MLE	**Df Model:**	8
Log-Likelihood:	-6951.7	**Pseudo R-squ.:**	0.04634
converged:	True	**LL-Null:**	-7289.5
Covariance Type:	nonrobust	**LLR p-value:**	1.272e-140

Table 4.11: All Psychographic Results

	coef	**std err**	**z**	**P¿|z|**	**[0.025**	**0.975]**
risk_aversion	-0.0059	0.001	-5.109	0.000	-0.008	-0.004
risk_seeking	-0.0023	0.001	-1.893	0.058	-0.005	8.26e-05
openness	0.0011	0.004	0.282	0.778	-0.006	0.008
extraversion	-0.0050	0.004	-1.259	0.208	-0.013	0.003
neuroticism	0.0183	0.003	6.780	0.000	0.013	0.024
agreeableness	0.0071	0.004	1.879	0.060	-0.000	0.015
conscientiousness	-0.0590	0.005	-13.066	0.000	-0.068	-0.050
inward_focus	-0.0067	0.001	-6.189	0.000	-0.009	-0.005
outward_focus	-0.0022	0.001	-3.190	0.001	-0.004	-0.001

Table 4.12: All Psychographic Results

	5%	95%	Odds Ratio
risk_aversion	0.991860	0.996365	0.994110
risk_seeking	0.995262	1.000083	0.997669
openness	0.993692	1.008486	1.001062
extraversion	0.987283	1.002791	0.995006
neuroticism	1.013076	1.023836	1.018442
agreeableness	0.999693	1.014691	1.007164
conscientiousness	0.934415	0.951098	0.942720
inward_focus	0.991279	0.995464	0.993369
outward_focus	0.996430	0.999146	0.997787

Table 4.13: All Psychographics: Odds Ratio and Confidence Intervals

4.4.5. All Demographics

Dep. Variable:	is_vaccine_hesitant	**No. Observations:**	21727
Model:	Logit	**Df Residuals:**	21716
Method:	MLE	**Df Model:**	10
Log-Likelihood:	-7630.9	**Pseudo R-squ.:**	-0.04684
converged:	True	**LL-Null:**	-7289.5
Covariance Type:	nonrobust	**LLR p-value:**	1.000

Table 4.14: All Demographics Results

	coef	**std err**	**z**	**P¿|z|**	**[0.025**	**0.975]**
blue_collar_midwest	-0.7843	0.081	-9.686	0.000	-0.943	-0.626
heartland_new_south	0.1371	0.044	3.121	0.002	0.051	0.223
other	-0.1951	0.190	-1.029	0.304	-0.567	0.177
gender_male	-0.4978	0.047	-10.620	0.000	-0.590	-0.406
black	-2.3094	0.266	-8.682	0.000	-2.831	-1.788
hispanic	-0.4095	0.115	-3.565	0.000	-0.635	-0.184
api	-0.3557	0.151	-2.351	0.019	-0.652	-0.059
age_<=18	-0.5243	0.146	-3.585	0.000	-0.811	-0.238
age_19-29	-1.9174	0.146	-13.124	0.000	-2.204	-1.631
age_30-39	-2.1129	0.085	-24.885	0.000	-2.279	-1.946
org_is-org	-1.6141	0.074	-21.939	0.000	-1.758	-1.470

Table 4.15: All Demographics Results

	5%	95%	Odds Ratio
blue_collar_midwest	0.389460	0.534940	0.456440
heartland_new_south	1.052331	1.250147	1.146982
other	0.567404	1.193078	0.822774
gender_male	0.554518	0.666368	0.607876
black	0.058970	0.167287	0.099322
hispanic	0.530109	0.831654	0.663978
api	0.520900	0.942597	0.700713
age_<=18	0.444424	0.788426	0.591942
age_19-29	0.110394	0.195732	0.146995
age_30-39	0.102358	0.142779	0.120891
org_is-org	0.172335	0.229946	0.199067

Table 4.16: All Demographics: Odds Ratio and Confidence Intervals

4.4.6. All Demographics and Psychographics

Dep. Variable:	is_vaccine_hesitant	**No. Observations:**	21727
Model:	Logit	**Df Residuals:**	21707
Method:	MLE	**Df Model:**	19
Log-Likelihood:	-6842.7	**Pseudo R-squ.:**	0.06130
converged:	True	**LL-Null:**	-7289.5
Covariance Type:	nonrobust	**LLR p-value:**	2.540e-177

Table 4.17: All Demographics and Psychographics Results

	coef	std err	z	P¿|z|	[0.025	0.975]
blue_collar_midwest	-0.3233	0.084	-3.868	0.000	-0.487	-0.159
heartland_new_south	0.5519	0.048	11.457	0.000	0.458	0.646
other	0.1243	0.190	0.654	0.513	-0.248	0.497
gender_male	0.1630	0.059	2.783	0.005	0.048	0.278
black	-0.2446	0.261	-0.937	0.349	-0.756	0.267
hispanic	0.0429	0.115	0.373	0.709	-0.183	0.268
api	-0.1145	0.152	-0.752	0.452	-0.413	0.184
age_<=18	0.5309	0.160	3.321	0.001	0.218	0.844
age_19-29	0.2858	0.168	1.700	0.089	-0.044	0.615
age_30-39	0.4180	0.121	3.448	0.001	0.180	0.656
org_is-org	0.1116	0.092	1.209	0.227	-0.069	0.293
risk_aversion	-0.0062	0.001	-5.316	0.000	-0.009	-0.004
risk_seeking	-0.0026	0.001	-2.123	0.034	-0.005	-0.000
openness	-0.0003	0.004	-0.087	0.931	-0.008	0.007
extraversion	-0.0091	0.004	-2.264	0.024	-0.017	-0.001
neuroticism	0.0137	0.003	4.862	0.000	0.008	0.019
agreeableness	0.0045	0.004	1.175	0.240	-0.003	0.012
conscientiousness	-0.0586	0.005	-12.893	0.000	-0.068	-0.050
inward_focus	-0.0070	0.001	-6.442	0.000	-0.009	-0.005
outward_focus	-0.0019	0.001	-2.634	0.008	-0.003	-0.000

continued on next page

continued from previous page

Table 4.18: All Demographics and Psychographics Results

	5%	95%	Odds Ratio
blue_collar_midwest	0.614406	0.852611	0.723775
heartland_new_south	1.580158	1.908604	1.736634
other	0.780348	1.643127	1.132347
gender_male	1.049356	1.320180	1.177004
black	0.469337	1.306276	0.782997
hispanic	0.833157	1.307779	1.043832
api	0.661706	1.201867	0.891786
age_<=18	1.243010	2.326106	1.700404
age_19-29	0.957279	1.850173	1.330839
age_30-39	1.197667	1.926409	1.518946
org_is-org	0.933025	1.339803	1.118065
risk_aversion	0.991512	0.996076	0.993791
risk_seeking	0.994937	0.999797	0.997364
openness	0.992196	1.007195	0.999667
extraversion	0.983077	0.998774	0.990894
neuroticism	1.008189	1.019356	1.013757

continued on next page

continued from previous page

agreeableness	0.996979	1.012165	1.004543
conscientiousness	0.934700	0.951508	0.943066
inward_focus	0.990889	0.995129	0.993007
outward_focus	0.996770	0.999525	0.998147

Table 4.19: All Demographics and Psychographics: Odds Ratio and Confidence Intervals

Chapter 5

Discussion

Section 5.1

Classification Task

Given the limited amount of training data, we obtained a surprisingly high F1 score with the AdaBoost classifier and TF-IDF vectorization. For our use case TF-IDF was the clear choice over BERT, not only did it provide higher F1 scores across most classifiers, but also had a significantly shorter time to encode the text: on one sample SentenceTransformers took 9+ hours, while TF-IDF vectorizer took less than 10 minutes.

It should also be noted that this classifier has not been tested for robustness over different time periods. Based on first-hand exploration, we find that while the amount of discourse on vaccines and COVID-19 on Twitter fluctuates over time, the type of discourse doesn't seem to differ significantly. Hence we hypothesize that the classifier should be relatively robust over time. A simple experiment that could help test such a claim would be to split our training data in half chronologically and train separate classifiers on each half. If there are large differences in the F1 scores between classifiers this would indicate a lack of robustness over time.

Section 5.2

Regression Task

Despite the noise in the data and classifier, the final results showed statistically significant trends in the demographic data.

5.2.1. Demographics

Significant relationships with vaccine hesitancy were observed for all demographic factors: race, age, geopolitical region, human-vs-organization status. Given the noise in the data and the limitations in the study as described in the next section, the exact strength of the relationships is likely flawed and not particularly accurate; however, the results are promising in highlighting general trends.

Starting with Human-vs-Organization status, humans on Twitter are far more likely to be vaccine hesitant as compared to organizations. This seems intuitive as organizations are less likely to tweet on political, medical, or controversial issues. The difference in genders also aligns with research regarding COVID-19 vaccine hesitancy[12]. With regards to age, we find that Twitter users aged 40 and older in the sample to be most likely to be vaccine hesitant, as compared to the other age groups. These findings are not supported by vaccine uptake surveys as of April 2021, which indicate higher vaccination rates within the 40+ demographic as compared to those younger[4]. However, the age and co-morbidity based priority system for COVID-19 vaccines established by US state governments complicates interpretations of vaccination reports, as older age groups had greater access to vaccines. This was not accounted for. The trends for vaccine hesitancy with race are interesting, as the racial majority seemed most likely to be vaccine hesitant. This is incongruent with public perception and multiple studies that indicate that racial minorities tend to be more vaccine hesitant than Caucasian populations in the United States and the

United Kingdom[14][38]. While the difference in results could be highlighting a new trend, the more likely explanation is that we had a biased sample consisting mostly of Caucasian (or white) users due to either a biased dataset or incorrect inference. It is also worth noting that given the difference in demographics between Twitter users and the general US population, these results may also indicate a difference in vaccine hesitancy trends for those on Twitter versus the general population. Furthermore, given that 10% of adult Twitter users create 80% of the tweets[37], such findings could also be a result of certain demographics being more vocal about their beliefs leading to an overestimation. Finally, we also find strong relation for vaccine hesitancy with geopolitical region, with Heartland and New South States seeming to be the most vaccine hesitant geopolitical region. Our findings align with recent surveys from the U.S. Department of Health & Human Services which identify states such as Montana, Wyoming, North Dakota, South Dakota and Idaho as the most vaccine hesitant[3], all of which lie in the Heartland and New South Alliance. Furthermore, as of April 23rd, 2021, many of the Heartland and New South states were reported to have the lowest vaccination rates making up seven of the bottom ten states[10].

Such trends have significant implications on public health campaigns. If we can use existing social media to identify vaccine hesitant populations, governments and public health experts can focus time, effort and resources to target these vulnerable populations. Furthermore, such a framework requires minimal resources and time as compared to existing methods such as polling or longitudinal studies which can lead to early identification of vulnerable populations.

5.2.2. Psychographics

The psychographic traits seem to have little to no correlation to vaccine hesitancy, with the absolute values of all but Neuroticism falling below 0.01, both when compared in isolation and with demographic traits. Even the highest trait, Neuroticism, scored

0.0183 (not accounting for demographic traits) which is not a significant correlation. While this result seems to indicate that psychographic traits do no correlate with vaccine hesitancy, such a hypothesis seems counter-intuitive and is inconsistent with prior studies. Prior research has shown that the Big Five personality traits influence vaccine attitudes, specifically that people high in agreeableness, conscientiousness and neuroticism are more likely to consider vaccination as beneficial[25]. Given the significant limitations in the inference of psychographic traits (as discussed in the next section), we find that psychographic traits would be challenging to accurately identify and work with, in regards to Twitter media. As mentioned below, the large amount of text per sample required per user exceeds the number of characters a tweet allows. We hypothesise that with a better set of data to score on we would likely find stronger correlations.

Section 5.3

Privacy and Ethics

This study brings up significant concerns regarding privacy online. It displays the extent to which publicly available social media information can be used to infer demographic and psychographic information about users which they may not be comfortable sharing and would not consent to if asked. While this study is not meant to be an examination of ethics of exploring social media data, we find it important to make note of the ethical side of such explorations as ethical considerations must be taken in any future applications of our framework. Especially when dealing with any identifying and demographic data, care must be taken to respect the privacy of individual users. Although all data used is in the public domain, we choose not to display any names/usernames or identifying data as part of the study.

Section 5.4

Limitations

5.4.1. Data

A big limitation in this study was the lack of good training data for the classifier. We could only find a small set of anti-vaccine data (through labelling). There are 4 potential factors for this:

- Both the datasets we used were deficient in anti-vaccine tweets.
- Our labelling methods needed improvement.
- In 2019 Twitter cracked down on anti-vaxxer content and removed many tweets and accounts[11]. This means they were not hydrated when we obtained the datasets.
- There are a dearth of anti-vaxxer accounts/data on Twitter as a platform [15].

Another limitation of our data is that it is specific to Twitter. A mix of social media data would have been ideal. Certain sub-reddits and Facebook groups are organizing points and echo chambers for anti-vaccine movements which would have been helpful in learning signals for vaccine hesitancy. It is also important to note that Twitter user demographics are not representative of US demographics[37] which could influence results as our dataset samples weren't adjusted to be representative of the US populations.

Finally, as a result of the noise in the data and the small training dataset, our vaccine hesitancy predictions are quite noisy and contain a significant number of false positives. We focused on reducing the number of false negatives over false positives due to the class imbalance and the small number of anti-vaccine content.

5.4.2. Demographic and Psychographic Inference

There were several limitations in our demographic and psychographic inferences as well.

Psychographic Inference. For the psychographic inferences, we were limited by the free tier of the Receptiviti service, which restricted how much text we could process. In our case removing this constraint wouldn't have affected our results as due to the lack of data, we had very few tweets per user, with an average of fewer than 3 tweets per user. Furthermore, due to the character constraint imposed by Twitter, tweets are small with the largest tweets in our dataset being 59 words in length. This poses a problem as the API documentation recommends providing text samples of at least 350 words for the highest accuracy. It must also be noted that, as mentioned in the methodology, the scores for psychographic labels are not the same as those that would be given after a clinical test, they imply that those many percent of the samples in the Receptiviti database have scores less than the score of the text sample being analyzed for the given psychographic trait. Hence, there is also the added limitation of not knowing what the "proprietary datasets" are and the internal accuracy measures.

Finally, even if psychographic scores are accurate, there may be differences in observed online and offline personality traits [27], or even in online personalities on different social media [33] which would make results harder to interpret.

Demographic Inference. With regards to the race inference, the census data used by Ethnicolr is from 2010, while the latest US Census was conducted in 2020. Demographics have changed in the last decade, and this could adversely affect results. Furthermore, some users added prefixes and suffixes to their names, or added fake or joke names to their profile. Our preprocessing method could not fully account for these factors which likely reduced the accuracy of the results. Additionally, Ethnicolr's methods are not perfect. In one application, they observed that the last name model (which we use in this study) had a .9 AUC and 83% accuracy [23]. In another evaluation it was also found that Ethnicolr had significantly lower accuracy for black, asian and hispanic individuals (33%, 60% and 59%) as compared to white individuals (96%)[24].

With regards to the age, gender and human-vs-organization inferences, our largest shortcoming is that we could not utilize the full strength of the M3 inference package due to issues in retrieving profile pictures and a number of fake or missing profile pictures.

With regards to the location features, there is no perfect geopolitical divide as the matter is subjective. It is possible a different arrangement of US states would produce more illuminating results. Additionally, the 'Other' category is not a particularly useful grouping as the US territories are not necessarily geographically or politically similar.

Chapter 6

Conclusion

Section 6.1

Findings

Our primary goal was to establish a framework that can be used to find relationships between demographics/psychographics and vaccine hesitancy through social media, that can be easily extended to other topics. In this regard, we succeeded. Our model showed statistically significant trends in demographic populations which seem to mirror what other studies have found while showing no significant trends for psychographic characteristics. As mentioned before, it is important to note that given the noise in the data and classifier, the strength of the correlations is likely not accurate and is not strong enough to be applied elsewhere. Still, a framework that is able to find trends and relationships without a massive set of training data and using easily available social media data provides a promising path for examining a wide variety of narratives, and doing quick preliminary explorations.

Section 6.2

Future Work in the Area

As aforementioned, this project was undertaken as a proof of process, and it would be interesting to see how results would differ on different social media. We'd be interested in using data from Facebook and Reddit and see whether similar relations could be observed. The purpose of the study is to establish a framework, and thus we'd be interested in applying this framework to other topics and areas such as election polling, understanding different populations' reactions towards various policy issues such as gun control.

Given the limitations surrounding the lack of valid data, collecting more training data to improve the classifier, as well as increasing the number of tweets per user collected to get more accurate psychographic and demographic results would help yield more significant results. Hyper-parameter tuning might lead to improvements in both the classification and regression, but the key bottlenecks in this project lie in the noisy data and classifiers. Furthermore, it would be interesting to see the effect of adding other demographic characteristics such as nationality and income level to the set of variables. For the purposes of applying the results to the general population, we would also be interested in running a version of this study where we sample the data to match US demographics and see the effect on the trends. We acknowledge this approach would still have limitations.

Finally, given the limitations of our approach towards getting the demographic and psychographic data, it is possible that our labelling is inaccurate and flawed. Hence using alternative approaches or manual labelling of users to generate more accurate labellings would help verify the validity of the results.

Appendix A

Section A.1

Tweet Object

Attribute	Type	Description
created_at	String	UTC time when this Tweet was created. Example:
		"created_at": "Wed Oct 10 20:19:24 +0000 2018"
id	Int64	The integer representation of the unique identifier for this Tv
		"id":1050118621198921728
id_str	String	The string representation of the unique identifier for this Tw
		"id_str":"1050118621198921728"
text	String	The actual UTF-8 text of the status update. See twitter-tex
		"text":"To make room for more expression, we will now cour
source	String	Utility used to post the Tweet, as an HTML-formatted strin
		Example:
		"source":"Twitter Web Client"
truncated	Boolean	Indicates whether the value of the text parameter was trunca
		"truncated":true
in_reply_to_status_id	Int64	Nullable. If the represented Tweet is a reply, this field will co
		"in_reply_to_status_id":1051222721923756032
in_reply_to_status_id_str	String	Nullable. If the represented Tweet is a reply, this field will co
		"in_reply_to_status_id_str":"1051222721923756032"
in_reply_to_user_id	Int64	Nullable. If the represented Tweet is a reply, this field will co
		"in_reply_to_user_id":6253282
in_reply_to_user_id_str	String	Nullable. If the represented Tweet is a reply, this field will co
		"in_reply_to_user_id_str":"6253282"
in_reply_to_screen_name	String	Nullable. If the represented Tweet is a reply, this field will co
		"in_reply_to_screen_name":"twitterapi"

user	User object	The user who posted this Tweet. See User data dictionary for complete list c
		Example highlighting select attributes:
		{ "user": {
		"id": 6253282,
		"id_str": "6253282",
		"name": "Twitter API",
		"screen_name": "TwitterAPI",
		"location": "San Francisco, CA",
		"url": "https://developer.twitter.com",
		"description": "The Real Twitter API. Tweets about API changes, servic
		"verified": true,
		"followers_count": 6129794,
		"friends_count": 12,
		"listed_count": 12899,
		"favourites_count": 31,
		"statuses_count": 3658,
		"created_at": "Wed May 23 06:01:13 +0000 2007",
		"utc_offset": null,
		"time_zone": null,
		"geo_enabled": false,
		"lang": "en",
		"contributors_enabled": false,
		"is_translator": false,
		"profile_background_color": "null",
		"profile_background_image_url": "null",
		"profile_background_image_url_https": "null",
		"profile_background_tile": null,
		"profile_link_color": "null",
		"profile_sidebar_border_color": "null",
		"profile_sidebar_fill_color": "null",
		"profile_text_color": "null",
		"profile_use_background_image": null,
		"profile_image_url": "null",
		"profile_image_url_https": "https://pbs.twimg.com/profile_images/942858
		"profile_banner_url": "https://pbs.twimg.com/profile_banners/6253282/14
		"default_profile": false,
		"default_profile_image": false,
		"following": null,
		"follow_request_sent": null,
		"notifications": null
		}
		}

coordinates	Coordinates	Nullable. Represents the geographic location of this Tweet a
		"coordinates":
		{
		"coordinates":
		[
		-75.14310264,
		40.05701649
		],
		"type":"Point"
		}
place	Places	Nullable When present, indicates that the tweet is associated
		"place":
		{
		"attributes":{},
		"bounding_box":
		{
		"coordinates":
		[[
		[-77.119759,38.791645],
		[-76.909393,38.791645],
		[-76.909393,38.995548],
		[-77.119759,38.995548]
		]],
		"type":"Polygon"
		},
		"country":"United States",
		"country_code":"US",
		"full_name":"Washington, DC",
		"id":"01fbe706f872cb32",
		"name":"Washington",
		"place_type":"city",
		"url":"http://api.twitter.com/1/geo/id/0172cb32.json"
		}
quoted_status_id	Int64	This field only surfaces when the Tweet is a quote Tweet. T
		"quoted_status_id":1050119905717055488
quoted_status_id_str	String	This field only surfaces when the Tweet is a quote Tweet. T
		"quoted_status_id_str":"1050119905717055488"
is_quote_status	Boolean	Indicates whether this is a Quoted Tweet. Example:
		"is_quote_status":false
quoted_status	Tweet	This field only surfaces when the Tweet is a quote Tweet. T
retweeted_status	Tweet	Users can amplify the broadcast of Tweets authored by othe

quote_count	Integer	Nullable. Indicates approximately how many times t
		"quote_count":33
		Note: This object is only available with the Premium
reply_count	Int	Number of times this Tweet has been replied to. Exa
		"reply_count":30
		Note: This object is only available with the Premium
retweet_count	Int	Number of times this Tweet has been retweeted. Exa
		"retweet_count":160
favorite_count	Integer	Nullable. Indicates approximately how many times t
		"favorite_count":295
entities	Entities	Entities which have been parsed out of the text of th
		"entities":
		{
		"hashtags":[],
		"urls":[],
		"user_mentions":[],
		"media":[],
		"symbols":[]
		"polls":[]
		}
extended_entities	Extended Entities	When between one and four native photos or one vid
		"entities":
		{
		"media":[]
		}
favorited	Boolean	Nullable. Indicates whether this Tweet has been like
		"favorited":true
retweeted	Boolean	Indicates whether this Tweet has been Retweeted by
		"retweeted":false
possibly_sensitive	Boolean	Nullable. This field only surfaces when a Tweet cont
		"possibly_sensitive":false
filter_level	String	Indicates the maximum value of the filter_level paran
		Example:
		"filter_level": "low"
lang	String	Nullable. When present, indicates a BCP 47 languag
		"lang": "en"
matching_rules	Array of Rule Objects	Present in filtered products such as Twitter Search a
		"matching_rules": " [{
		"tag": "twitterapi emojis",
		"id": 1050118621198921728,
		"id_str": "1050118621198921728"
		}]"

Bibliography

[1]

[2] *Receptiviti documentation.*

[3] *Vaccine hesitancy for covid-19: State, county, and local estimates*, May 2021.

[4] LYDIA ANDERSON, Thomas File, Joey Marshall, Kevin Mcelrath, and Zachary Scherer, *How do covid-19 vaccination and vaccine hesitancy rates vary over time?*, Apr 2021.

[5] Pablo Barberá, *Birds of the same feather tweet together: Bayesian ideal point estimation using twitter data*, Political Analysis **23** (2015), no. 1, 76–91.

[6] Erika Bonnevie, Jaclyn Goldbarg, Allison K. Gallegos-Jeffrey, Sarah D. Rosenberg, Ellen Wartella, and Joe Smyser, *Content themes and influential voices within vaccine opposition on twitter, 2019*, American Journal of Public Health **110** (2020), no. S3, S326–S330, PMID: 33001733.

[7] Samantha Bradshaw and Philip N Howard, *Challenging truth and trust: A global inventory of organized social media manipulation*, The Computational Propaganda Project **1** (2018).

[8] US Census Bureau, *Cartographic boundary files*, May 2020.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, *BERT: Pre-training of deep bidirectional transformers for language understanding*, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (Minneapolis, Minnesota), Association for Computational Linguistics, June 2019, pp. 4171–4186.

[10] Alison Durkee, *Here are the states with the greatest covid-19 vaccine hesitancy, poll says*, Apr 2021.

[11] Daniel Funke and Susan Benkelman, *Twitter has finally gone after anti-vaccine misinformation. but the results are inconsistent.*, May 2019.

[12] Vincenzo Galasso, Paola Profeta, Martial Foucault, and Vincent Pons, *Covid-19 vaccine's gender paradox*, 03 2021.

[13] Alec Go, Richa Bhayani, and Lei Huang, *Twitter sentiment classification using distant supervision*, Processing **150** (2009).

[14] Kevin Grumbach, Timothy Judson, Manisha Desai, Vivek Jain, Christina Lindan, Sarah B. Doernberg, and Marisa Holubar, *Association of Race/Ethnicity With Likeliness of COVID-19 Vaccine Uptake Among Health Workers and the General Population in the San Francisco Bay Area*, JAMA Internal Medicine (2021).

[15] Keith Gunaratne, Eric A. Coomes, and Hourmazd Haghbayan, *Temporal trends in anti-vaccine discourse on twitter*, Vaccine **37** (2019), no. 35, 4867–4871.

[16] Xuehua Han, Juanle Wang, Min Zhang, and Xiaojie Wang, *Using social media to mine and analyze public opinion related to covid-19 in china*, International Journal of Environmental Research and Public Health **17** (2020), no. 8.

[17] Jun Ito, Kyosuke Nishida, Takahide Hoshide, Hiroyuki Toda, and Tadasu Uchiyama, *Demographic and psychographic estimation of twitter users using social structures*, pp. 27–46, Springer International Publishing, Cham, 2014.

[18] Kokil Jaidka, Saifuddin Ahmed, Marko Skoric, and Martin Hilbert, *Predicting elections from social media: a three-country, three-method comparative study*, Asian Journal of Communication **29** (2019), no. 3, 252–273.

[19] Amelia Jamison, David A. Broniatowski, Michael C. Smith, Kajal S. Parikh, Adeena Malik, Mark Dredze, and Sandra C. Quinn, *Adapting and extending a typology to identify vaccine misinformation on twitter*, American Journal of Public Health **110** (2020), no. S3, S331–S339, PMID: 33001737.

[20] Kaushik Kash, *Covid vaccine tweets*, Mar 2021.

[21] Michal Kosinski, David Stillwell, and Thore Graepel, *Private traits and attributes are predictable from digital records of human behavior*, Proceedings of the National Academy of Sciences **110** (2013), no. 15, 5802–5805.

[22] Rabindra Lamsal, *Coronavirus (covid-19) geo-tagged tweets dataset*, 2020.

[23] Suriyan Laohaprapanon and Gaurav Sood, *Ethnicolr*.

[24] ———, *Evaluating ethnicolr*, Nov 2020.

[25] Fang-Yu Lin and Ching-Hsing Wang, *Personality and individual attitudes toward vaccination: a nationally representative survey in the united states*, BMC Public Health **20** (2020), no. 1.

[26] Micol Marchetti-Bowick and Nathanael Chambers, *Learning for microblogs with distant supervision: Political forecasting with Twitter*, Proceedings of the 13th

Conference of the European Chapter of the Association for Computational Linguistics (Avignon, France), Association for Computational Linguistics, April 2012, pp. 603–612.

[27] Tamsin C. Marriott and Tom Buchanan, *The true self online: Personality correlates of preference for self-expression online, and observer ratings of personality online and offline*, Computers in Human Behavior **32** (2014), 171–177.

[28] Shannon C McGregor, *Social media as public opinion: How journalists use social media to represent public opinion*, Journalism **20** (2019), no. 8, 1070–1086.

[29] Rasmus Kleis Nielsen and Kim Christian Schrøder, *The relative importance of social media for accessing, finding, and engaging with news*, Digital Journalism **2** (2014), no. 4, 472–489.

[30] Nils Reimers and Iryna Gurevych, *Sentence-bert: Sentence embeddings using siamese bert-networks*, Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 11 2019.

[31] Clay Smith, *Idiot's guide to odds ratios*, Dec 2018.

[32] Magdalena Szumilas, *Measuring vaccine hesitancy: The development of a survey tool*, Journal of the Canadian Academy of Child and Adolescent Psychiatry = Journal de l'Academie canadienne de psychiatrie de l'enfant et de l'adolescent **19** (2010), no. 3, 227–229.

[33] Lee Taber and Steve Whittaker, *Personality depends on the medium: Differences in self-perception on snapchat, facebook and offline*, Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (New York, NY, USA), CHI '18, Association for Computing Machinery, 2018, p. 1–13.

[34] H. Tankovska, *Number of social media users 2025*, Jan 2021.

[35] V. Lance Tarrance, *A new regional paradigm for following u.s. elections*, May 2021.

[36] Zijian Wang, Scott Hale, David Adelani, Przemyslaw Grabowicz, Timo Hartmann, Fabian FlÃ¶ck, and David Jurgens, *Demographic inference and representative population estimates from multilingual social media data*, 05 2019, pp. 2056–2067.

[37] Stefan Wojcik and Adam Hughes, *How twitter users compare to the general public*, Jan 2021.

[38] Chris Martin Christopher A Nellums Laura B Guyatt Anna L Melbourne Carl Bryant Luke Gogoi Mayuri Wobi Fatimah Al-Oraibi Amani Hassan Osama Gupta Amit John Catherine Tobin Martin D Carr Sue Simpson Sandra Gregary Bindu Aujayeb Avinash Zingwe Stephen Reza Rubina Gray Laura J Khunti Kamlesh Pareek Woolf, Katherine McManus and Manish, *Ethnic differences in sars-cov-2 vaccine hesitancy in united kingdom healthcare workers: Results from the uk-reach prospective nationwide cohort study*, medRxiv (2021).