University of Montana

ScholarWorks at University of Montana

Graduate Student Theses, Dissertations, & Professional Papers

Graduate School

2020

MODELING TWITTER SENTIMENT AS A FUNCTION OF PARTICULATE MATTER 2.5 FOR COMMUNITIES IMPACTED BY WILDFIRE ACROSS MONTANA AND IDAHO

Matthew Kelly

MODELING TWITTER SENTIMENT AS A FUNCTION OF PARTICULATE MATTER 2.5

FOR COMMUNITIES IMPACTED BY WILDFIRE ACROSS MONTANA AND IDAHO

By

Matthew Kelly

Bachelor of Science in Biology, Minor in Chemistry, University of Missouri-Kansas City,
Kansas City, MO, 2011

Bachelor of Arts in Psychology, University of Missouri-Kansas City,
Kansas City, MO, 2011

Thesis

presented in partial fulfillment of the requirements
for the degree of

Master of Science
in Computer Science

The University of Montana
Missoula, MT

Winter 2020

Approved by:

Ashby Kinch PhD, Dean
Graduate School

Douglas Brinkerhoff PhD, Chair
Department of Computer Science

Erin Landguth PhD
School of Public and Community Health Sciences

Jesse Johnson PhD
Department of Computer Science

Kelly, Matthew, Master of Science, Winter 2020 Computer Science

Modeling Twitter Sentiment as a Function of Particulate Matter 2.5 for Communities Impacted by Wildfire across Montana and Idaho

Chairperson: Douglas Brinkerhoff PhD

Fine particulate matter (PM2.5) is a known pollutant with clinically detrimental physiological and behavioral effects. We consider Twitter sentiment as a potential indicator for well-being in communities impacted by wildfire-associated PM2.5 across Montana and Idaho spanning 5 years (2014-2018). From these geospatial air quality data and geo-tagged tweets, we trained county level models to examine the power of Twitter sentiment as a function of PM2.5. For all 24 counties sampled, we found between 1 and 8 affective dimensions where a positive r^2 was detected with a significant F-statistic ($p < 0.05$). Specifically, we show that sentiment for anticipation in the wildfire-prone county of Missoula, MT yielded respective training/test set r^2 of 0.0958 and 0.0686 with a p-value for the F-statistic of 3.09E-07. These analyses support social media sentiment as a potential public health metric by showing one of the first observations of a relationship between PM2.5 and Twitter sentiment.

CONTENTS

1 INTRODUCTION

This master's thesis utilizes techniques in data science to explore the relationship between air quality and distress expressed in social media. It is organized into sections, the first of which includes previous studies on Twitter sentiment analysis and research goals. Section 2 provides background on big data, including a description of our tweet and air quality data. Frameworks for accessing and analyzing data motivate many decisions we make in our section 3 methods. Results are presented in section 4 with a comprehensive review for Missoula County, MT in 4.1, and all counties in 4.2. Finally, we state our conclusions in section 5 by addressing our research goals with our findings.

PM2.5 is a known pollutant with clinically detrimental physiological and behavioral effects [1]. The inspiration for then using PM2.5 as a predictor for Twitter sentiment came from a review of previous research where tweet sentiment was explored as a health proxy. A review of those studies follow here. For each, we consider how tweet sentiment has been used, how we may use existing methods, and where our research diverges.

Social media has become a powerful tool allowing researchers to survey a population before, during, and after an event of interest, enabling the study of unpredictable events. Lin et al. (2017) leveraged this capability to create computational focus groups of spatiotemporally similar users affected by the November 2015 Paris attacks [2]. The dimensions of the response were defined as 3 primary negative emotions: anger, anxiety, and sadness. Tweet sentiment for these attributes were detected using the bag of words model Linguistic Inquiry and Word Count (LIWC) [3]. An increase and subsequent decrease in sentiment intensity was recorded as a period of distress and recovery. Their findings were (1) that a spike in anger, anxiety, and sadness was observed on the day of the attack, and (2) proximity to the attack correlated strongly with the magnitude of distress. Our approach is also to decompose tweets into affective dimensions, but we are interested in modeling sentiment not for a single event, but as correlated across several years to explore the time-invariant relationship to PM2.5.

In addition to a retrospective analysis, Twitter sentiment has also been explored for the early detection of an acute outbreak in thunderstorm induced asthma [4]. Joshi et al. (2020) were able to predict these outbreaks up to 9 hours prior to hospital records of the event in 3/18 of their experiments, and before news reports of the outbreak in 5/18. This was accomplished by first identifying tweets which included personal health keywords like “cough”. They hypothesized that multiple tweets with pathologically relevant references, separated by short time intervals, were rare events and therefore could be signals of distress. We also aimed to create a predictive model to illustrate correlations between environmental stimuli and sentiment. However, we could not find a significant number of tweets containing personal health keywords that were collocated with our PM2.5 data. As a result, we expanded our data scraping to encompass all possible tweets.

1.1 Research Goals

1. Discover the best method for obtaining tweets by moderating cost and search power.
2. Compare the performance of multiple language models through their impact on predicting sentiment from PM2.5.
3. Determine if the predictive models are under or overfit.
4. Identify the biggest sources of error.
5. Propose ways to decrease error.
6. Discuss the relationship between PM2.5 and Twitter sentiment.

2 BIG DATA

Data can become so-called big data when it is too large for a monolithic database [5]. It is therefore stored in distributed systems which often include functionality for capture, retrieval, as well as analysis. Sources of big data can be classified as social, transactional, or machine. Social data are generated by people on services alike social media. Transactional information encompasses primarily business and stock market data. And machine data are created through industrial processes, scientific research, and anywhere else an Internet of Things (IoT) device can be found. Using these definitions, our Twitter data is considered social, and our PM2.5 data is considered machine.

Analytically, these distinctions become relevant when considering how data types are monetized differently, how queries impose bias, and how the signal-to-noise ratio is affected by a composition of data types. For example, regardless of where in the United States we sampled, a manual review showed that the Twitter users posting the most tweets were usually bots uploading information such as stock price and weather. These numerical tweets would have introduced low variance, neutral sentiment into our models, but were controlled for by ignoring tweets and users which were affectively null. Tweet acquisition is covered in 3.2, further methods on filtering and normalization can be found in 3.3, and sources of error are discussed in our section 5 conclusions.

2.1 The Data Economy

Large and valuable datasets compose the data economy and are aggregated by a wide range of services. Consumption of these data is unlike the consumption of other goods and services in that its use does not reduce its supply, and its abundance does not reduce its value [6]. In the case of big data, an increase in supply leads to an increase in value. This is because larger datasets can yield emergent properties of exponentially greater utility, inflating cost equivalently. Such dynamics have resulted in specialized data producers and consumers with complex strategies for monetization and acquisition, compounded by tensions between the demand for more data privacy and more powerful analytics.

Data science as a tool to influence behavior carries with it the potential for insight as well as abuse. Self-regulation of these concerns has been empowered by the dominance and novelty of big tech like Google in search, Amazon in e-commerce, Instagram in photos, and Twitter in free speech blogging. Some companies, such as Instagram [7] and Facebook [8], enforce privacy by forbidding any form of direct sale of their data. This is contrast to Twitter which offers tailored products to business, science, and government. Public tweets are even archived in the Library of Congress [9].

Policy within information technology companies evolves rapidly because of the global nature of their products, inconsistent or absent government regulation, the arms race between security and bad actors, and the capriciousness of public opinion. Consequentially, data scientists must adapt with equal pace. We have found that our own scrapers have required modification as frequently as daily. Directly communicating with Twitter on our project has been helpful in overcoming some of these hurdles.

Successful participation in the digital economy can be reduced to dealing with the difficulty of assigning value to data [10]. This is especially true as the number of free-to-use services and data warehouses increases. New players in a digital space can be disadvantaged by not being able to compete on price when incumbent services are free. However, companies that charge content creators like Social Blade, pay creators like advertisers, and broker data like Brand24 constitute an ecosystem stabilized by competitive fees. Some of these data firms offer great value for acquiring tweets, but in our experience did not offer historical data or were opaque in terms of how their tools worked. We therefore were challenged to weigh the monetary cost of acquiring data directly from Twitter with the time cost of developing our own scraper in section 2.3.

2.2 PM2.5

Our PM2.5 air quality data is geospatial in nature, stored as GeoTIFF rasters [11]. Readings were obtained from previous work which combined measurements from ground monitoring stations and satellite imagery. These data were input into a model which yielded 32-bit floating point values with a spatial resolution of 16x13 km and temporal resolution of 1 day spanning 2009/06/01 to 2019/03/31. This range of dates would later be restricted to the 5 years between 2014 and 2018 to match the years for which we had tweets for all counties. Models which included all dates in each year at first struggled to predict tweet sentiment from PM2.5. We hypothesized that the exogenous variable of seasonality was suppressing any potential relationship between the two. This seasonal variation was controlled for by further restricting the range of PM2.5 and Twitter data to only the summers of 2014-2018. Ultimately, we would discover that selecting only summer data would become vital for detecting significant correlations.

2.3 Twitter

Twitter is a platform for creating, consuming, aggregating, and analyzing real-time information. Its strategies for monetization include advertising to users and selling publicly posted tweets [12]. A tweet can contain rich multimedia including text up to 280 characters, photos, videos, polls, mentions, hashtags, moments, reposts of other tweets, replies, and reactions in the form of likes, emojis, and gifs [13]. Because tweets are minimally censored [14] and their text terse, Twitter has differentiated itself as the free speech microblogging service.

Tweets are posted or scheduled in chronological order to a user's Profile timeline and are automatically geotagged by Foursquare's location services, unless the user opts out [15]. Tweet geo-tagging is explored in greater detail in section 3.2.3. Following hashtags, topics, and other users, populates the Home timeline with a synthesis of highly ranked tweets [16]. And the Explore timeline curates trending tweets and users [17]. This suggested content encourages engagement and is a primary success metric for social media platforms [18].

Engagement is measured in terms of likes, retweets, follows, replies, and clicks. Business [19] and non-business [20] tools have been developed by Twitter to track the performance of previous posts with the purpose of guiding the creation of new content. Custom analytics can also be performed by first downloading tweets through one of Twitter's data products, or through a data firm specializing in social media, or by scraping its website directly. For all 3 of these strategies, special care must be taken when submitting a query because in many contexts Twitter enables quality filters by default. These filters are informed by engagement and therefore may return non-representative samples. Section 3.2.1 outlines our search strategy.

First-party access to Twitter data is partitioned into 2 Representational State Transfer (REST) Application Programming Interfaces (APIs) each with several data products, and each with rate limits and quotas [21]. API v2.0 documentation suggests that scientific research may make use of expanded caps [22][23]. At present though, researchers are limited to 300 API requests per 15 minutes, returning at most 500 tweets per request no older than 30 days, totaling no more than 500,000 per month. Older tweets require a premium request and cap total monthly tweets to 25,000. Beyond these limits are the Custom or Enterprise Tracks, rates for which are not posted and are subject to change, but we were quoted at greater than $250,000 to retrieve all tweets from 50 of the most populous cities across Montana and Idaho, 25 each, spanning the same 10-year range as our PM2.5 data (2009-2019). As we await the finalization of Twitter’s academic research policies, we have in the meantime found it prudent to scrape our own data as described in 3.2.

3 METHODS

Our methods center on data scraping, sentiment analysis, modeling the relationship between PM2.5 and sentiment, and aggregating model output. Methods for scraping in section 3.2 describe how we query tweets, the attributes of a tweet, and how the tweet data structure is used in scraping. Natural Language Processing (NLP) is used in 3.3 for data cleaning and implementing 2 types of unsupervised sentiment analysis models. In 3.4 we explain why we chose to linearly model sentiment as a function of PM2.5, then carefully consider the assumptions of linear regression. Finally, in 3.5 we preempt our results with an explanation of our success metrics and how our analyses will address research goals.

3.1 Working with PM2.5 and County Level Data

We extracted daily PM2.5 rasters for Montana and Idaho and projected them to the WGS 84 coordinate system. These data were then masked with county shape files, also WGS 84, taken from the US Census Bureau [24] using the Shapely [25] and RasterIO [26] libraries. The shapefile-cropped raster data were averaged to find the daily PM2.5 value per county and served as our explanatory variable.

Both PM2.5 and Twitter data were aggregated by county because neighboring cities in the same county were assumed to be under the same environmental conditions. Also, through trial-and-error we found that Twitter automatically geo-tags tweets to the nearest city even when significantly outside jurisdictional boarders. This means that even when a user is outside the city they are tagged in, it is likely we still assign them to the correct county. Our choice of querying the 25 most populous cities in Montana and the 25 most populous cities in Idaho resulted in 24 counties between them. Keeping our data at the county level, as opposed to state, allows us to gage the location dependence of sentiment on PM2.5.

3.2 Creating a Custom Twitter Scraper

Limitations in existing Twitter mining libraries reduce to either quotas or obsolete APIs. Our process for surveying scrapers began by selecting Python and JavaScript as languages which account for 63% and 15% of Twitter scraping repositories on GitHub. Of those, all projects which have been starred and updated within 6 months were manually screened for quality by code review. The 2 types of implementations we found were either web scrapers or calls to the official Twitter API. Having already decided to scrape Twitter, we observed that no such repositories on GitHub consistently adapted to Twitter's frequently evolving search result interface. It was therefore judicious to create our own scraper from the ground up.

3.2.1 Building Twitter Search Queries

An advanced Twitter search URL has the general form:

https://twitter.com/search?
q=[search string] [operator key]:[operator value]&[filter key]=[filter value]

These queries have space delimited operators and ampersand delimited filters. We used the date range operators "since" and "until" and the geo-tag operator "place" [27]. In our experience the behavior of operators is less stable than filters. It is important to manually test that every operator and filter in a search has the desired effect because the only indication a query is bad is the return of arbitrary tweets.

Search filters can be set by either specifying them in the query URL or by toggling them from within a user's notification settings [28]. This means search results will vary depending on whether a user is logged in. By default, spam filters are enabled and performed well in practice. The only filter we added was "f=live", which disabled the filter for tweets with only high-quality engagement, in order to collect the most representative sample.

3.2.2 Tweet Attributes & Obfuscation

The primary features of a scraped tweet are profile image, display name, username, verification status, time, tweet text, tweet media, location, replies, retweets, and likes. We extracted username, date of post, and text alone. Within the text we ignored emojis because we did not have emoji training data or an obvious method for mapping emojis to sentiment using our language models in 3.3.

These features are typically stored in the attributes of HTML elements, or tags. Elements can be uniquely referenced with the "id" attribute. On un-obfuscated websites, the value of the "id" attribute typically describes the type of data inside that element. Other attributes such as "class" specify the template a tag conforms to and can therefore also be used to deduce data type. But on obfuscated sites, the values of these attributes are randomized or encrypted so that they are no longer human readable. We were however able to overcome these challenges in 3.2.6.

3.2.3 Twitter Location Services

We make use of the premium search operator “place” which accepts either an ID or string name. Using location operators with strings often returns erroneous tweets for small to moderate sized cities. Other location operators exist, such as bounding boxes drawn with longitude and latitude, but because Foursquare has become the default location service for anyone posting a new tweet, we decided to begin our scraping by obtaining a table of place IDs for all our 50 cities of interest. Since June 2019 precise geolocation has been removed [29] but may still be accessed for some services like marketing [30].

Each tweet contains only a single place ID and therefore it is not possible to first search on the state level and then segregate those tweets more precisely later. Searching for all tweets in Montana and Idaho yields a different dataset than searching for all cities in both states. While it is faster and less expensive to make fewer API calls with a larger geographic scope, we chose to scrape on the city level for higher precision.

3.2.4 Dates and Times

Time stored as an HTML tag (<time>) was the only attribute which remained uniquely identifiable after Twitter began obfuscating the attributes of its web search results. Without a <time> element enumerating each tweet in the results, scraping would have been more difficult. Other social media such as Snapchat do replace the <time> tag with a generic <div> element. This indicates that our method for identifying tweets should not be considered stable.

Twitter assigns every tweet a Unix timestamp which makes controlling for time zones simple. We binned our tweet sentiment at the same 1-day temporal resolution as our PM2.5 data by indexing dates using the Skyfield library [31]. We set the winter equinox as the start of each year, then calculated an offset in days for each tweet and PM measurement. This was the easiest and most precise way to specifically sample summer-only data for reasons mentioned in 2.2.

3.2.5 Scraping Un-obfuscated Tweets

We make use of the Beautiful Soup library for our un-obfuscated scraper [32]. It functions as a parser and does not use the interactive Document Object Model (DOM). This text-only approach has the advantage of speed over a DOM-aware library like Selenium [33]. Beautiful Soup can also be parallelized to overcome network I/O bottlenecks, while Selenium cannot. We could reliably run 2 parallel threads without being rate-limited by Twitter, achieving 30 tweets/second on a single IP address. Successive pages of tweets were loaded by using the “data-min-position” HTML attribute found in the last tweet of a given page. Adding the min-position attribute to our search query and refreshing was all that was necessary to retrieve the next batch of data.

3.2.6 Scraping Obfuscated Tweets

Without the ability to find the min-position attribute we had to retrieve additional pages of tweets by simulating scrolling. This meant using Selenium, an industry standard tool in quality control. Each tweet was found using the <time> tag. From this point we traversed up the HTML tree to locate the username, identified as a string prepended with an “@”. We then traversed down the tree to find the next string element which was always the text body of the tweet. DOM traversal was accomplished by generating XPath [34] language formatted queries and submitting them using the XPath interface within Selenium.

JavaScript calls to the browser triggered scrolling and the loading of additional tweets. However, parsing time and RAM usage increased as more tweets were loaded. We reused the same tree traversal method centered around <time> tags to identify already scraped tweets and delete them with JavaScript calls. These methods improved performance, but even so we were not able to parse faster than 5 tweets per second. Because Selenium was not CPU parallelizable, scraping in the cloud was the next best optimization.

3.2.7 Cloud Scraping

We chose DigitalOcean and Docker as our primary cloud tools because of their tight integration. DigitalOcean specializes in creating virtual Linux environments, and Docker enables containerized execution of code. The Docker Machine API was used to dynamically create DigitalOcean compute instances called Droplets [35]. Each containerized scraper was parameterized with a search query targeting a specific city, and data was then sent to a dedicated database Droplet for final download and analysis. Distributing in this manner allowed us to use many instances of slower scrapers, while also not exceeding Twitter's per IP rate limit.

3.3 Natural Language Processing

NLP is broadly the discipline of giving computers the ability to understand human language. This most commonly takes the form of machine learning with methods for annotating meaningful linguistic structures such as parts-of-speech. The classical NLP tools we make use of are tokenizers, which parse words from non-words, and WordNet, a graph of meaningfully related words found the Natural Language Tool Kit (NLTK) library [36].

3.3.1 Cleaning Tweets

Each tweet was cleaned by first using the NLTK word tokenizer to generate a sequence of words and non-words found in the text body. We discard URLs, mentions, and usernames prepended with “@”, all of which are assumed not to contribute to the meaning of a tweet. Hashtags are identified as words or phrases joined without spaces and prepended with “#”. These tags are designed to be meaningful, so care must be taken to parse them back into individual words. All other words were passed through a custom spellchecker using the SymSpell [37] and Wordninja [38] libraries, both cited in contemporary NLP literature as found in Banthia et al. [39] and Tekumalla et al. [40].

3.3.2 VADER Sentiment Model

Existing sentiment models such as Valence Aware Dictionary for Sentiment Reasoning (VADER) and its predecessor LIWC are widely adopted in academia as well as industry with proven performance [42]. VADER builds on the bag of words approach by using heuristics, accounting for such things as negation words. Studies have shown VADER (F1=0.96) beating humans (F1=0.84) in a task classifying text as positive, neutral, or negative.

Our early results could only predict VADER sentiment as a function of both lagged sentiment and PM2.5, but failed as a function of PM alone. Improving the quality of our data cleaning allowed us to weakly detect VADER sentiment as a function of just PM. We then rationed that we could further explore this relationship by creating an even more sensitive language model. We reviewed VADER sentiment for a sample of tweets and identified several opportunities for advancement. Peer reviewed training data for tweet sentiment analysis was unfortunately not available, so we focused on an unsupervised strategy.

3.3.3 Creating a Custom Sentiment Model

Affective mining, or affective computing is the process of measuring the emotional dimensions in natural language. Our implementation synthesized the National Research Council Canada (NRC) Word-Emotion Association Lexicon (EmoLex) [43] with Google Book's Ngram project [44] to create our own bag of words model. EmoLex provides a human annotated dictionary of words categorized into the 8 bi-polar Plutchik emotions: joy/sadness, trust/disgust, anger/fear, and anticipation/surprise along with the polar dimensions positive/negative. This lexicon was filtered down to the 6468 words with non-neutral affect. [Table 2] shows output for the word "advance".

Emotion	Sentiment
Joy	0.25
Sadness	0.0
Trust	0.0
Disgust	0.0
Anger	0.0
Fear	0.25
Anticipation	0.25
Surprise	0.25

[Table 1.1] The affective response for the word “advance” as scored by EmoLex. A mapping is shown between the emotional dimensions and their respective sentiment scores. Data are normalized to 1.

Polarity	Sentiment
Positive	1.0
Negative	0.0

[Table 1.2] A mapping of polar dimensions and sentiment scores from EmoLex for the word “advance”.

This 6468-word lexicon was expanded by mapping words with known sentiment in their synonyms a 4-step process. First (1), if an input word can be matched exactly in EmoLex, its affective score is returned. Failing this, secondly (2) an exact match is searched for in a lexicon of synonyms to EmoLex. Failing this, thirdly (3) synonyms of the input word are generated, and each are searched for as exact matches in EmoLex, then normalized to a single score. Fourth (4), having no prior match, synonyms of the input word are matched against the lexicon of synonyms to EmoLex. And lastly, failing all else neutral affect is returned.

To create a lexicon of synonyms to EmoLex, each word in EmoLex is input into NTLK's WordNet and synsets for that word are found. Synsets are alternative meanings to word. For example, the word "bank" can mean a guarantee, or the bank of a river, etc. Then for each alternative meaning, or synset, we consider its lemmas. Lemmas are words of the same synset which have the same meaning. The sentiment for each word in EmoLex is then distributed to of its lemmas to a new synonym lexicon. When multiple words from EmoLex map to the same synonym, the new sentiment is normalized to 1. This synonym lexicon is what is used in steps 2 and 4 above.

In steps 3 and 4, synonyms to the input word are generated through a similar method to generating synonyms to EmoLex. First the input world's synsets is found, then all its lemmas listed. These lemmas are used to find exact matches to sentiment in either EmoLex, step 3, or its synonym lexicon, step 4.

The expansion of sentiment in EmoLex to its synonyms, and the expansion of input words to their synonyms are weighted by word frequency as found the Google Ngram project. By weighting synonyms in this way, we anticipate the most likely meaning amongst competing lemmas. Expanding on the example word "advance", we discovered that the synonym "forefront" was not found in either EmoLex or its synonyms. This means "forefront" does not apply to conditions 1 or 2. Its lemmas "vanguard" and "cutting edge" do exist in either EmoLex or the synonym lexicon. Since "vanguard" was found in EmoLex, matching condition 3, the sentiment for "forefront" was assigned to that of "vanguard" seen in [table 4].

Polarity	Forefront	Vanguard
Positive	1.0	1.0
Negative	0.0	0.0

[Table 1.3] The word “forefront” was not found in EmoLex, but a mapping was possible through its synonyms. While “advance” had sentiment all 9 affective dimensions, its synonym “forefront” only contains polar dimensions.

As stated above all sentiment for individual words are normalized to 1 when added together for the purposes of expanding our lexicon. This normalization continues similarly for the addition of sentiment from multiple words in a tweet. However, when words are affectively null, or an entire tweet is affectively null, those records are discarded. Not discarding neutral sentiment would make normalizing to 1 impossible, resulting in low precision when a lot of neutral affect is aggregated. This became a significant problem because of the discovery of machine and financial data in our tweet dataset mentioned in section 2. Discarding neutral, unnormalized affect had the added benefit of improving the predictive power of our forecast models.

Modeling sentiment as a function of PM2.5 expects as input a single value for the response variable. Consolidating multiple tweets for a single day required 2 rounds of averaging and normalizing. First, each user’s own tweets on a given day were averaged and normalized. Then, affect for all users on a given day were averaged and normalized. Each of the 9 affective dimensions were split into separate timeseries and modeled independently with shared daily values for PM2.5.

3.4 Modeling Sentiment as a Function of PM2.5

Our model choices were driven by (1) a desire to examine if a relationship between sentiment and PM2.5 could be supported, as opposed to building the best forecast tool. And (2) the observation that both sentiment and PM2.5 are excellent candidates for autoregression. Sections 3.4.1 and 3.4.3 describe the types of models we use, while 3.4.2 and 3.4.4 follow with considerations. 3.5 integrates our choice of success metrics and how relevant results are generated in section 4.

3.4.1 Choosing a Linear Model

We chose linear least squares estimators because their interpretability support the goal of conservatively exploring a possible relationship between tweet sentiment and PM2.5. Further research may expand on social media as a measure of public health by leveraging the power of non-linearity, but first we must establish if such a connection is reasonable. Runtime was also important for this preliminary study. Even simple non-linear models were orders of magnitude slower, compounding extensive cross-validation and the challenge of consolidating insight from permutations on lag order, polynomial degree, multiple hypothesis testing, and language model for 24 counties. A linear model is not without caveats. Each assumption in 3.4.2 includes consideration for where violations may occur.

3.4.2 Gauss-Markov Assumptions

The Gauss-Markov theorem asserts that least squares estimators are the best linear unbiased estimators when (1) the model is linear, (2) error variance is constant, (3) errors are independent, (4) errors are normal, (5) there is no perfect multicollinearity, and (6) there is no omitted variable bias.

(1) A model is linear when its parameters are linear, meaning its explanatory variables are linearly related to the response variable. A linear model with non-linear coefficients cannot be guaranteed to be unbiased. However, non-linear transformations are acceptable on independent variables. In 3.4.3 we consider linear and quadratic polynomials which model sentiment as a function of PM2.5 alone and as a function of PM2.5 with lagged sentiment.

(2) Stationarity or homoskedasticity is a property of stochastic processes and is present when the mean and variance of a dataset are constant. Regressing on non-stationary data can result in a spurious correlation even with highly confident models. Additionally, the accuracy, or variance, of the model may fluctuate through time despite coefficients which remain unbiased. This is because high variance data may adjust model parameters too much, while low variance data may not adjust them enough. We found that all our timeseries became stationary by simply applying their first difference. This was measured with a confidence interval of 95% using an augmented Dickey-Fuller test implemented in the Stats Models package [47].

(3) Error terms must be independent and not autocorrelated. When errors can be used to predict other errors, this suggests that our independent variables cannot fully explain our dependent variable. This can be remedied in an autoregressive model by increasing the lag order. Coefficients

remain unbiased, failing condition 3, while patterns in the residuals manifest themselves with variance in the output as a function of time.

It is also possible that autocorrelated error terms could be the result of modeling the wrong relationship. A model more or less complex, or even non-linear, may have been necessary. But if we successfully find independence in the error terms this supports the hypothesis our model and variables were appropriately chosen.

(4) Normality is assumed for errors. When random errors are not from a normal distribution, we can no longer make assertions about our confidence intervals. This is only true for small sample sizes, however. As sample size increases, the Central Limit Theorem guarantees that the expected values, or sample means, of our random variables will approach a normal distribution.

(5) No perfect multicollinearity requires that none of our independent variables are linear transformations of each other. We do not expect there to be perfect correlation between PM2.5 and tweet sentiment. Nor do we expect PM2.5 to be perfectly collinear with its lags. Nor do we expect sentiment to be perfectly collinear with its lags. These assumptions are reasonable because atmospheric processes and human behavior are prototypically non-deterministic, and therefore their lags are unlikely to be linear transformations of each other. In cases where multicollinearity is significant, confidence in the parameters decreases as multiple equally valid solutions emerge.

(6) No omitted variable bias assumes that no variables which were excluded drive variables which were included. If this were violated, the model would try to explain the omitted variable in terms of the included variables. The explanatory variables would depend in part on their error terms, which are no longer independent themselves in violation of condition 3. We do

expect violation of exogeneity here. It is impossible to account for and measure all known influencers of sentiment and PM2.5.

It is possible that the influence of temperature on PM2.5 and sentiment is the true relationship. Or, that seasonality, which drives temperature, accounts for the influence on both variables. This effect of season and temperature on affect and wildfire is why we chose to model only summer seasons. While still in violation of condition 6, our model remains a valid tool for making predictions. The most significant consequence is that it is only possible to draw conclusions on correlation, not causation.

3.4.3 An Autoregressive Process

An autoregressive model (AR) was appropriate because we expect that future sentiment and PM2.5 are both significantly influenced by their previous values. Shih et al. suggest that day-of-the-week Twitter sentiment follows a 7-day cycle with a low on Monday and high on Friday. We estimate then that the length of ascension or descension to be about 3 days for a 7-day sinusoid. Ensuring our lag window is smaller than the period of the weekly sentiment cycle, we first modeled an AR(3) process but consider lags from 0-14 as well.

Beginning with an AR(1) process:

$$y_t = \phi y_{t-1} + \varepsilon_t$$
$$= \phi(\phi y_{t-2} + \varepsilon_{t-1}) + \varepsilon_t$$
$$y_t = \phi y_0 + \textstyle\sum_{k=0}^{t-1} \phi^k \ \varepsilon_{t-k}$$

y: response, ϕ: parameter, ε: error, t: timestep

AR(1) variance:

$$var(a) = 0, \ var(ax) = a^2 var(x)$$
$$var(y_t) = \sigma^2 \sum_{k=0}^{t-1} \phi^{2k}$$

AR(1) expected value:

$$E(\varepsilon) = 0, \ E(ax) = aE(x)$$
$$E(y_t) = \phi E(y_{t-1})$$
$$= \phi E\big(\phi E(y_{t-2})\big)$$
$$= \ \phi^2 E(y_{t-2})$$
$$E(y_t) = \phi^t y_0$$

Because the expected value $E(y_t)$ is a function of ϕ^t multiplied by the initial value y_0 we can see that when $|\phi| < 1$ the value is pulled back to the mean, and it is stationary. When $|\phi| > 1$ the expected value explodes to infinity and is not stationary. And when $|\phi| = 1$ this suggests that our data is a random walk. The Augmented Dickey-Fuller tests these values of ϕ and is used to validate that our AR(3) model receives only stationary sentiment and PM2.5 data [48].

An AR(1) process can be extended to AR(3) by:

$$y_t = \phi_1 x_t + \phi_2 x_{t-1} + \phi_3 x_{t-2} + \phi_4 x_{t-3} + \varepsilon_t$$

3.4.4 Spurious Correlation in Autoregression

We consider 4 types of random walks:

1. A pure random walk. The variance depends on time and trends to infinity. It cannot be predicted.

$$y_t = y_{t-1} + \varepsilon_t$$

2. A walk with drift.

$$y_t = \alpha + y_{t-1} + \varepsilon_t$$

α: drift

3. A walk with trend, constant variance, and constant growth in its mean.

$$y_t = \alpha + \beta t + \varepsilon_t$$

β: trend

4. A walk with trend and drift.

$$y_t = \alpha + \beta t + y_{t-1} + \varepsilon_t$$

It is not uncommon to discover that attributes in a dataset conform to a random walk. And these walks may appear to have either linear or supra-linear growth which match the growth pattern of your target series. The effect of regressing on attributes which are random walks is a set of coefficients which suppress the true correlation. A random walk which has been transformed to be stationary has also become white nose and it is not possible to spuriously correlate the dependent variable on truly random data. We conclude then that if we transform our sentiment and PM data into a stationary form then we will remove spurious regression as a likely outcome.

3.4.5 Orthogonal Distance Regression

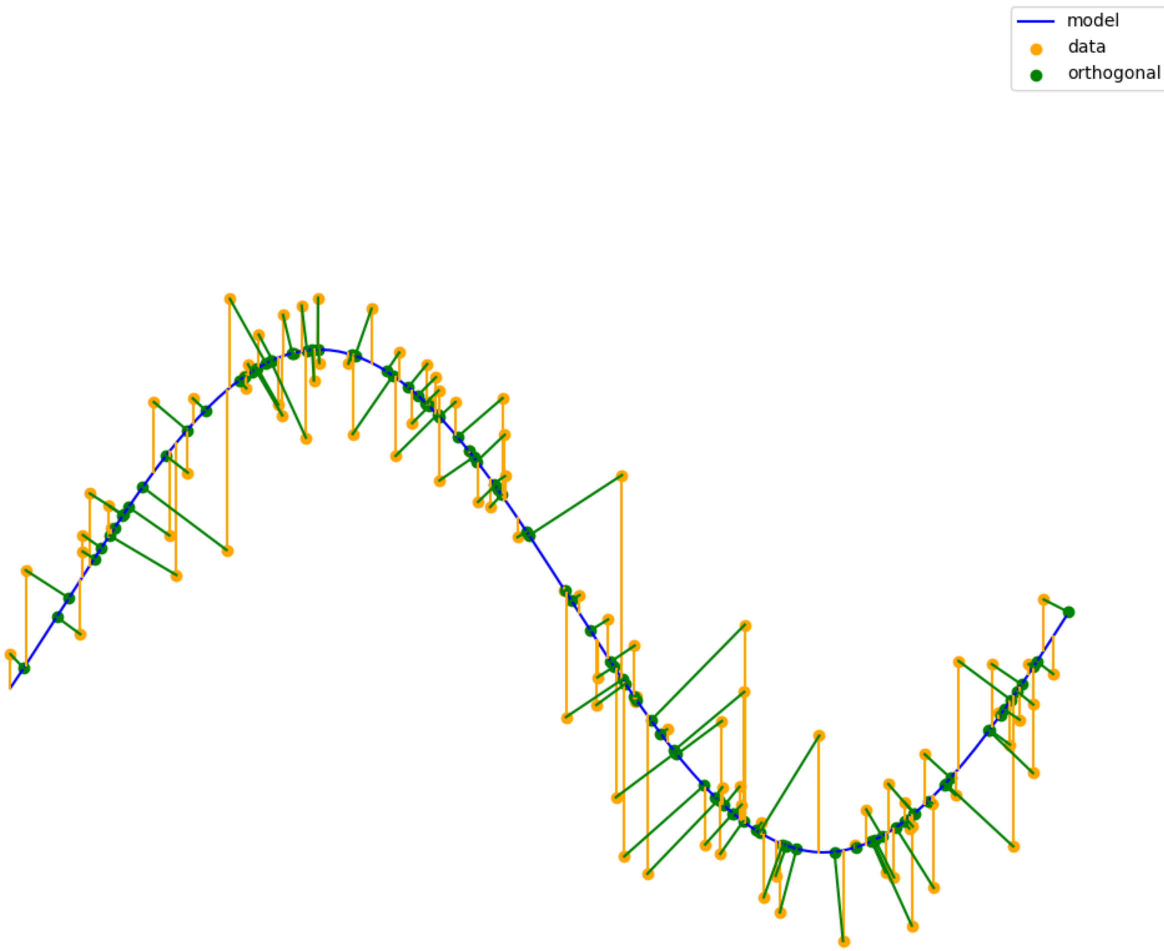
Orthogonal Distance Regression (ODR) was considered in addition to OLS to explore the possibility of significant error in both our independent and dependent variables. We expected that PM2.5 measurements would be accurate, but there may be significant differences between atmospheric readings and a user's sensitivity to air quality. ODR measures errors which are orthogonal to the regression line [figure 1]. Closed form solutions are challenging to derive but we were able to accurately estimate their values through random sampling.

We created TensorFlow models with layers equivalent to polynomials fit with OLS, and used the Adam optimizer. Within the loss function, every batch of input is given an extra so-called candidate dimension with a size of 1000. This means, for every instance in the original input batch there are 1000 new values in the new loss function batch. These 1000 new values per instance are populated by drawing from a random distribution centered on each input instance's value, with a standard deviation equal to its corresponding absolute error. A wider spread increases the chance that an instance with a high absolute error will yield a random sample with an orthogonal value.

Candidate samples are fed back into the model to produce predicted values. We calculate the change in input as well as the change in output for all candidates and find the input-output pair with the shortest Euclidian distance. This is our orthogonal distance, the precision of which is dependent on the size of the candidate dimension. We chose 1000 because it was the highest factor of 10 for which our GPU did not suffer a runtime bottleneck.

A batch of orthogonal distances, now without a candidate dimension, was finally reduced to a single loss value by finding the Euclidian distance of the entire batch of orthogonal points. Our implementation accepted an arbitrary number of input and output attributes to accommodate

AR processes of any lag. For a synthetic AR(3) process results were as expected. When there was no error in the explanatory variables, OLS outperformed ODR. When the explanatory variables had significant error, ODR outperformed OLS except when the response variable was high.



[Figure 1] The blue line labeled model represents a learned function given synthetic data. The yellow points labeled data are the input data and the yellow lines represent the errors calculated by an OLS model. Green dots labeled orthogonal are the orthogonal projection of the actual data on to the model. And the green lines are the approximate orthogonal errors returned in our TensorFlow loss function.

3.4.6 Generating Results

Models for each county were generated with permutations on language model, affective dimension, predictive model type (OLS and ODR), lag order (from 0-14), and polynomial degree (from 1-3) by 10x5-fold cross-validation. Each of the 10 sets of 5 folds were randomly sampled with the goal finding a representative experiment. The model with the median test set r^2 was chosen as our primary metric of success. This extensive randomly sampled cross-validation was primarily used to control for the sensitivity of r^2 on the standard deviation and as a function of time for timeseries.

F-test were used to compare between lag orders, polynomial degrees, and intercept-only models. The restricted intercept-only model always returns the mean response value. An AR(3) linear polynomial OLS model was chosen by comparing F-test p-values between hyperparameters having correlations with the greatest significance shown in section 5. Reported results in 5.1-5.2 were chosen by accepting only those models which were both significantly different from their intercept-only counterpart and had positive test set r^2 values. Such results indicate that the model is different from the mean while also explaining more variance than the mean. Additional metrics include a Dickey-Fuller test for stationarity (addressing assumption 2 section 4.4.2) and a Durbin-Watson test for autocorrelation (addressing assumption 3).

Days with missing sentiment were populated by averaging the next earliest and next latest datum. If a gap in the data occurred at the beginning of a series, then the first observed value was replicated across that gap. Similarly, if a gap occurred at the end of a series, then the last observed value was replicated. All training sets spanned 360 days and all test sets spanned 90 days, so only counties with at least 450 tweets could have no missing data. As the number of tweets in a county grew in general, gap length shrunk dramatically.

4 RESULTS

Analyses begin in 4.1 for Missoula County then follow for all counties in 4.2. We chose to focus on Missoula because it has a moderate number of tweets and is frequently and severely impacted by wildfire. Critical values for Durbin-Watson in [table 2.2] apply to all counties because tweets for all counties were binned first by user, then by day, and all span the same range of dates.

A linear polynomial AR(3) OLS model type was the most successful in exploring the effect of PM2.5 on Twitter sentiment and therefore applies to 4.1-4.2. Our metrics of success are F-test p-value (H_0: parameters for unrestricted and restricted models are the same), Dickey-Fuller p-value (H_0: indicates heteroscedasticity), Durbin-Watson test statistic (H_0 indicates AR(1) autocorrelation), and r^2.

$$y_t = \phi_1 x_t + \phi_2 x_{t-1} + \phi_3 x_{t-2} + \phi_4 x_{t-3} + \varepsilon_t$$

y: sentiment, x: PM2.5, t: day

4.1 Missoula County, MT

Results for the affective dimension of anticipation show a slight correlation between sentiment and PM2.5 as seen in the test set r^2 [table 2.1], using an AR(3) linear polynomial OLS regression with our own language model. Significant homoskedasticity in the errors was found in both test and training sets. The Durbin-Watson test shows no significant AR(1) auto-correlation in the residuals of the test and training sets. The total number of tweets prior to binning daily was 14,324.

The p-values for the F-tests indicate that, for the training sets, the unrestricted linear polynomial models were significantly different from the restricted intercept-only models. Further hypothesis testing, keeping lag order constant, found that 2^{nd} and 3^{rd} degree polynomials were either statistically similar to lower degree polynomials, except 0^{th}, or had less significant individual metrics of success. When the p-value for F-tests was significant and the restricted model was intercept-only, the unrestricted model always had a greater test set r^2. VADER rarely, and ODR more rarely, yielded unrestricted models which significantly differed from their restricted intercept-only counterparts for any county, and were both insignificant for Missoula County.

Comparing F-test p-values and test set r^2 for lags from 0-14 to their alternatives, differing by up to 4, a lag of 3 was chosen to maximize predictive power while minimizing complexity. Lags differing by 1 were never significant except in the case where the restricted model was intercept-only. Rarely were lags differing by 2-3 significant for any county. Our choice in model complexity and lag was limited in scope due to the number of possible permutations. Discovering our set of hyperparameters to significantly implicate PM2.5 as a predictor of tweet sentiment is therefore a foundation for future research in the building of performant public health models.

Dimension	Test DF	Test DW	Test MSE	Test r^2
surprise	1.43E-14	1.914795	0.023985	-0.00478
anticipation	**4.22E-06**	**2.219225**	**0.02808**	**0.068564**

Dimension	Train DF	Train DW	Train MSE	Train r^2
surprise	7.11E-13	1.964262	0.029271	0.029191
anticipation	**6.54E-13**	**2.116742**	**0.028593**	**0.095831**

Dimension	F-test	Intercept	x_t	x_{t-1}	x_{t-2}	x_{t-3}
surprise	0.032161	0.382103	-0.06779	0.138777	0.120325	0.016974
anticipation	**3.09E-07**	**0.818244**	**-0.23335**	**-0.10445**	**0.047185**	**-0.2403**

[Table 2.1] Results for Missoula County from an AR(3) linear polynomial OLS regression using our own language model. F-tests show surprise and anticipation had training sets which varied significantly from their intercept-only models. However, only anticipation had a positive test set r^2. Example words for anticipation can be seen in [tables 3.1-3.2]. All values derive from the median experiment of a 10x5-fold cross-validation sorted on the test set r^2. The Dickey-Fuller and F-tests are reported as p-values. Durbin-Watson is reported as its test statistic with [table 5.2] showing critical values. All significance levels were 95%.

Set	N	DW Low Crit.	DW Low	DW High	DW High Crit.
Test	90	1.566	**1.751**	**2.249**	2.434
Train	360	1.802	**1.848**	**2.152**	2.198

[Table 2.2] Sample sizes and Durbin-Watson test statistics apply for all counties. From 0 to DW low crit. indicates significant positive autocorrelation. From DW low crit. to DW low is inconclusive. Bolded values from DW low to DW high indicate a significant lack of autocorrelation. From DW high to DW high crit. is also inconclusive. And from DW high crit. to 4 indicates significant negative autocorrelation.

abeyance	attendance	competition	distribute	foresightful	impatient
accelerate	auction	completing	divination	foretell	impending
acquiring	audience	completion	dodderer	forethought	importance
addresses	auspicate	conjecture	doomsday	foreword	inaugural
adventure	auspices	consequent	draft	forming	inception
aeronautic	await	contiguous	during	frequence	incidental
aeronautical	beg	contingent	eagerness	genealogic	incipience
aeronautics	begun	continuation	edition	genealogical	incipiency
airport	biennial	continuing	emplace	genealogist	incipient
alchemic	bivouac	continuity	encampment	genealogy	industrialist
alchemical	board	continuousness	endeavor	genesis	industrious
alchemist	boat	continuum	engulf	germinal	industry
alchemize	boater	convergence	essayist	germination	infinity
alchemy	boating	convertibility	evening	gig	inflexibility
allocution	box	copyright	eventual	gradual	inhabited
allurement	brim	correspondence	evergreen	graduality	inhabiting
ancestral	broadside	countdown	exchangeability	gradualness	inquiry
angling	bruise	courtship	expectance	grasping	install
announcement	bugle	cramp	expectancy	gravitate	intended
answering	burrow	craps	expectant	habitual	interim
anticipation	bye	craving	expectation	handcraft	intermission
anticipatory	calculation	creeping	expected	handicraft	interpenetrate
apparent	camp	cue	expecting	handiwork	intuitively
appeal	camper	curiosity	expedition	hankering	invasive
append	camping	curricular	explore	happen	investigation
applicant	campy	daily	extricate	haste	invitation
approaching	candlelight	daybreak	farm	headlight	invocation
arbitration	canton	debenture	farsighted	here	invoke
archaeology	caption	delivery	fate	hereafter	labyrinth
arise	captious	denying	fathom	hereness	lands
arouse	card	destined	ferment	horizon	launch
arrival	career	develop	fermentation	horoscope	lessen
arrive	chemic	developer	fin	hungry	letter
assay	chemical	developing	flipper	hurried	liability
assayer	chemist	dietary	forecast	hurry	liable
astrologer	clock	diffuse	forerunner	hype	linger
astrology	cloth	diffuseness	foresee	immature	lips
astronomer	clue	digress	foreseen	immediately	local
attainable	coming	diligent	foresight	imminent	localize
attempt	commemorative	discreet	foresighted	immortality	locater

[Table 3.1] Words with sentiment over 50% in the anticipation dimension.

locator	notification	prediction	reconstruction	speculative	tributary
long	occupant	predilection	rectify	start	tunnel
lottery	occupier	predispose	recurrent	store	twenty
lull	occupying	prefatorial	refining	straighten	ubiquitous
lust	offset	preliminary	regatta	straightener	ubiquity
mail	olfactory	premeditate	rehabilitation	strive	ultimately
matchmaker	omen	preparation	renovate	submit	uncompromising
maternal	omnipresent	preparatory	reposition	subscribe	undertaking
merge	ongoing	prepare	repositioning	subtitle	undisclosed
midnight	onset	preparedness	representing	suffuse	unfold
mill	opportunity	prerequisite	request	sundown	uninterrupted
millenary	outdo	prescient	restlessness	sunset	university
millennial	overture	prevention	result	surround	unresolved
millennian	packer	previse	resultant	symmetricalness	until
millennium	paddle	primitive	revive	tabulate	untold
mobile	paddler	probability	ripen	tent	unverified
modernisation	pale	proceeding	roulette	tenting	urgent
modernization	paleness	production	rudimentary	theology	vicinal
modulate	pallidness	prognostic	rudiments	there	vicinity
momentum	pallor	programing	sailing	thermocouple	vigil
monetary	parole	programming	sailor	thermometer	virginity
morn	passenger	progress	saliva	thermometric	vision
morrow	patient	prologue	sassy	thirteen	voyage
motion	permeate	prophecy	scrutinize	thought	voyager
mountain	perpetuate	prophet	secular	thousand	waddle
mutable	perspective	prophetic	seductive	till	waddler
mystery	pervade	prophetical	seek	tillage	wade
nascent	petitionary	prospectively	sentiment	tilling	wader
nativity	placenta	prospector	sentimental	time	wading
naturalise	placental	prospicient	sequel	toddler	wait
naturalize	plan	public	serial	tomorrow	waiting
nautical	plump	punt	ship	totterer	waitress
navigable	poke	quest	shortly	tout	wanness
navigation	posited	quicken	shuttle	touter	while
navigational	possibility	readiness	shuttlecock	track	whilst
neighborhood	possibleness	ready	signify	transit	wishful
network	practise	recipient	simmering	transition	wizard
nil	precursor	recognizable	six	transitional	wont
noncompletion	predict	recombination	sonar	transmutability	yacht
nonessential	predicting	reconstruct	source	treadmill	yachting

[Table 3.2] Words with sentiment over 50% in the anticipation dimension.

4.2 All Counties

Significance levels for test sets [tables 4.1-4.2] and the training sets [tables 4.3-4.4] show strikingly that VADER was less responsive to PM2.5 than our own language model. It may be the case that having 9 affective dimensions compared to VADER's 1 best explains this. But a comparison between the two models would require training data to characterize them in a general context.

All counties had at least 1 model with an F-test showing that its prediction of sentiment from PM2.5 was significantly different from the intercept-only alternative. Affective dimensions which failed an F-test or had a test set $r^2 \leq 0$ were not reported for brevity. From the results shown, no county failed the Dickey-Fuller test for homoscedastic errors explained in assumption 2 of section 3.4.2. However, the Durbin-Watson test for AR(1) autocorrelation, explained in assumptions 3 and 6, indicates that some models may be better explained. Increasing model complexity would likely decrease complexity in the residuals and satisfy independence in the errors.

The complex nature of sentiment and air quality mean it's unlikely to be possible to account for all exogenous variables. We conclude in section 5 that our experiments only implicate a correlation between Tweet sentiment and PM2.5, and not correlation strength. This is especially true because of the dependence of r^2 on its standard deviation which would be expected to vary between counties. The presence of machine and transaction data within our tweets, discussed in section 2, is an example of how different types of users can impact overall sentiment. We grouped tweets by user to control for those who tweet more, but it is likely that larger communities have a more diverse userbase.

ODR only outperformed OLS in synthetic benchmarks with error in the explanatory variables and at most moderate error in the response variable, but always lead to accepting H_0 for the F-test for this study. It is likely error in Twitter sentiment was in excess to detect a correlation with ODR. Having provided a framework for scraping tweets, extracting affect, and correlating sentiment with environmental stimuli, further research can further optimize public health forecasting.

County	Tweets	Affect	Dimension	F-test	Test DF	Test DW	Test MSE	Test r^2
Ada, ID	149716	Custom	anticipation	0.012277	5.73E-10	1.8795	0.0432	0.0032
Ada, ID	149716	Custom	trust	0.023285	3.83E-10	1.8672	0.0496	0.0111
Ada, ID	149716	Custom	disgust	0.003992	2.10E-06	2.0317	0.0408	0.0229
Bannock, ID	38241	Custom	sadness	0.011264	1.31E-13	2.1827	0.0392	0.0007
Bannock, ID	38241	VADER	polarity	0.016761	2.38E-05	2.1115	0.0375	0.0035
Bingham, ID	1296	Custom	trust	0.002869	1.45E-08	**2.3974**	0.0342	0.0138
Bingham, ID	1296	Custom	surprise	3.62E-05	2.93E-06	**2.4098**	0.0212	0.0571
Bingham, ID	1296	VADER	polarity	2.59E-05	6.04E-07	2.0479	0.0363	0.0248
Bonneville, ID	6373	Custom	anger	0.036748	3.13E-07	2.0293	0.0310	0.0072
Bonneville, ID	6373	Custom	sadness	0.00349	1.26E-07	1.9912	0.0359	0.0134
Bonneville, ID	6373	Custom	fear	0.042798	8.29E-10	2.0058	0.0546	0.0149
Bonneville, ID	6373	Custom	anticipation	0.008701	1.66E-07	1.8288	0.0315	0.0163
Canyon, ID	24863	Custom	joy	0.010722	1.39E-10	**1.6900**	0.0315	0.0132
Canyon, ID	24863	Custom	anticipation	0.001274	2.31E-09	2.1878	0.0229	0.0278
Elmore, ID	2409	Custom	surprise	3.10E-05	1.77E-08	1.8819	0.0219	0.0461
Kootenai, ID	21363	Custom	polarity	0.022074	2.93E-06	1.8858	0.0354	0.0073
Kootenai, ID	21363	Custom	sadness	0.004539	2.39E-11	1.9017	0.0367	0.0160
Kootenai, ID	21363	Custom	trust	0.000399	3.91E-06	1.9702	0.0443	0.0225
Kootenai, ID	21363	Custom	fear	0.000779	1.57E-12	2.0337	0.0345	0.0336
Latah, ID	7193	Custom	fear	0.000647	1.72E-07	2.0670	0.0225	0.0247
Madison, ID	10608	Custom	polarity	0.004558	8.29E-09	1.9272	0.0424	0.0088
Madison, ID	10608	Custom	anger	0.03281	1.63E-05	2.0341	0.0221	0.0094
Madison, ID	10608	Custom	fear	0.001208	2.43E-09	1.9371	0.0351	0.0125
Madison, ID	10608	Custom	joy	0.001514	2.06E-12	**2.3175**	0.0230	0.0201
Madison, ID	10608	Custom	disgust	5.39E-05	1.54E-07	**2.3043**	0.0319	0.0345
Madison, ID	10608	Custom	sadness	1.11E-06	0.000152	2.1473	0.0273	0.0677
Nez Perce, ID	4000	Custom	anger	0.012702	8.59E-08	2.1098	0.0238	0.0039
Nez Perce, ID	4000	Custom	fear	5.56E-07	7.78E-24	1.9207	0.0300	0.0546
Nez Perce, ID	4000	Custom	sadness	9.86E-07	1.44E-07	**1.7475**	0.0173	0.0618
Twin Falls, ID	4934	Custom	surprise	0.042209	0.002722	1.9114	0.0267	0.0014
Twin Falls, ID	4934	Custom	fear	0.000991	1.77E-07	**2.2602**	0.0260	0.0282
Cascade, MT	21967	Custom	polarity	0.000913	6.18E-16	2.0410	0.0350	0.0158
Custer, MT	214	Custom	sadness	0.00071	1.53E-09	1.8323	0.0225	0.0176
Custer, MT	214	Custom	disgust	1.04E-05	4.12E-07	1.7559	0.0335	0.0414
Custer, MT	214	Custom	anger	4.77E-08	9.21E-08	2.0151	0.0240	0.0655
Deer Lodge, MT	213	Custom	anticipation	0.002167	5.86E-12	2.0227	0.0321	0.0099
Deer Lodge, MT	213	Custom	anger	0.000878	2.61E-13	**2.3643**	0.0349	0.0342
Deer Lodge, MT	213	Custom	trust	8.19E-07	3.48E-14	1.6220	0.0348	0.0798
Deer Lodge, MT	213	Custom	disgust	2.67E-10	1.96E-07	**1.7215**	0.0252	0.1138
Deer Lodge, MT	213	Custom	surprise	1.11E-16	4.40E-11	**1.5463**	0.0210	0.2708

[Table 4.1] Test set results for all counties. P-values ≥ 0.05 bolded.

County	Tweets	Affect	Dimension	F-test	Test DF	Test DW	Test MSE	Test r^2
Fergus, MT	272	Custom	anger	2.03E-07	1.60E-06	**1.7465**	0.0447	0.0597
Fergus, MT	272	Custom	anticipation	2.26E-08	2.62E-24	1.8025	0.0175	0.0668
Fergus, MT	272	Custom	trust	2.15E-09	1.08E-06	1.8692	0.0296	0.0921
Fergus, MT	272	Custom	fear	1.29E-12	2.67E-07	1.8136	0.0205	0.1423
Fergus, MT	272	Custom	polarity	2.91E-14	1.71E-05	2.1276	0.0256	0.1744
Fergus, MT	272	Custom	sadness	1.11E-16	1.80E-08	2.1484	0.0292	0.1839
Fergus, MT	272	Custom	disgust	1.11E-16	2.73E-14	**2.3588**	0.0184	0.3272
Fergus, MT	272	VADER	polarity	6.07E-06	0.000222	1.9833	0.0390	0.0683
Flathead, MT	10593	Custom	joy	0.006435	8.30E-11	**2.6049**	0.0359	0.0121
Gallatin, MT	22328	Custom	disgust	0.035114	5.70E-06	2.2409	0.0420	0.0002
Gallatin, MT	22328	Custom	anger	0.01211	1.32E-05	**1.7291**	0.0312	0.0061
Gallatin, MT	22328	Custom	trust	0.001372	1.65E-05	**2.2664**	0.0372	0.0208
Gallatin, MT	22328	Custom	fear	3.80E-11	1.08E-05	1.8086	0.0431	0.1190
Gallatin, MT	22328	VADER	polarity	0.001059	3.21E-14	2.0395	0.0418	0.0230
Hill, MT	841	Custom	disgust	0.01283	3.69E-08	2.1545	0.0397	0.0027
Hill, MT	841	Custom	surprise	0.000467	4.52E-11	1.9518	0.0201	0.0352
Hill, MT	841	Custom	sadness	1.11E-16	6.51E-09	2.0754	0.0365	0.1776
Hill, MT	841	Custom	trust	1.11E-16	4.20E-08	2.0425	0.0237	0.2577
Hill, MT	841	VADER	polarity	0.022305	2.66E-12	2.2301	0.0187	0.0138
Lewis and Clark, MT	4751	Custom	joy	0.001293	2.15E-06	2.1761	0.0326	0.0155
Lewis and Clark, MT	4751	Custom	fear	0.002329	9.17E-12	1.8896	0.0305	0.0321
Lewis and Clark, MT	4751	Custom	surprise	1.27E-07	2.89E-12	2.2090	0.0248	0.0647
Missoula, MT	14324	Custom	anticipation	3.09E-07	4.22E-06	2.2192	0.0280	0.0686
Park, MT	1641	Custom	fear	9.08E-05	2.21E-08	1.9364	0.0316	0.0473
Park, MT	1641	Custom	anticipation	6.25E-05	1.36E-06	2.2029	0.0271	0.0579
Park, MT	1641	Custom	disgust	3.29E-14	1.83E-09	1.9484	0.0320	0.1491
Richland, MT	871	Custom	disgust	0.03633	1.00E-06	2.0212	0.0124	0.0032
Richland, MT	871	Custom	anger	0.002837	2.85E-06	1.9747	0.0292	0.0266
Richland, MT	871	Custom	anticipation	4.74E-07	4.38E-06	2.0791	0.0316	0.0738
Silver Bow, MT	4784	Custom	disgust	1.08E-06	1.09E-12	1.9700	0.0191	0.0648
Silver Bow, MT	4784	Custom	sadness	1.45E-07	1.17E-07	2.1498	0.0282	0.0699
Yellowstone, MT	21551	Custom	trust	0.005437	0.000466	2.1109	0.0420	0.0146
Yellowstone, MT	21551	Custom	anger	0.001337	2.83E-13	1.8574	0.0419	0.0153
Yellowstone, MT	21551	Custom	sadness	0.000524	1.72E-05	**2.4549**	0.0243	0.0257
Yellowstone, MT	21551	Custom	anticipation	0.000526	7.24E-06	1.9093	0.0307	0.0315
Yellowstone, MT	21551	Custom	disgust	1.22E-09	8.75E-08	2.0533	0.0501	0.1141
Yellowstone, MT	21551	VADER	polarity	0.009171	1.02E-09	1.9550	0.0388	0.0143

[Table 4.2] Test set results for all counties. P-values ≥ 0.05 bolded.

County	Tweets	Affect	Dimension	F-test	Train DF	Train DW	Train MSE	Train r^2
Ada, ID	149716	Custom	anticipation	0.012277	6.42E-20	2.0016	0.0402	0.0353
Ada, ID	149716	Custom	trust	0.023285	3.18E-19	1.9478	0.0459	0.0313
Ada, ID	149716	Custom	disgust	0.003992	1.18E-14	**1.7714**	0.0399	0.0423
Bannock, ID	38241	Custom	sadness	0.011264	8.73E-13	1.9238	0.0277	0.0359
Bannock, ID	38241	VADER	polarity	0.016761	3.88E-15	1.9727	0.0410	0.0334
Bingham, ID	1296	Custom	trust	0.002869	1.63E-18	2.0924	0.0274	0.0443
Bingham, ID	1296	Custom	surprise	3.62E-05	1.19E-13	1.9777	0.0199	0.0697
Bingham, ID	1296	VADER	polarity	2.59E-05	3.52E-14	2.1000	0.0285	0.0716
Bonneville, ID	6373	Custom	anger	0.036748	4.92E-14	2.0691	0.0318	0.0283
Bonneville, ID	6373	Custom	sadness	0.00349	7.90E-18	2.0527	0.0314	0.0431
Bonneville, ID	6373	Custom	fear	0.042798	3.21E-15	**1.8395**	0.0371	0.0273
Bonneville, ID	6373	Custom	anticipation	0.008701	8.13E-19	1.9021	0.0346	0.0375
Canyon, ID	24863	Custom	joy	0.010722	2.04E-13	1.9616	0.0348	0.0362
Canyon, ID	24863	Custom	anticipation	0.001274	3.22E-21	2.0430	0.0355	0.0491
Elmore, ID	2409	Custom	surprise	3.10E-05	2.10E-16	2.0922	0.0261	0.0705
Kootenai, ID	21363	Custom	polarity	0.022074	1.11E-12	**1.8358**	0.0364	0.0316
Kootenai, ID	21363	Custom	sadness	0.004539	4.47E-15	**2.2212**	0.0370	0.0415
Kootenai, ID	21363	Custom	trust	0.000399	3.67E-14	1.9479	0.0414	0.0559
Kootenai, ID	21363	Custom	fear	0.000779	1.98E-19	1.9830	0.0396	0.0520
Latah, ID	7193	Custom	fear	0.000647	1.54E-10	1.9043	0.0309	0.0531
Madison, ID	10608	Custom	polarity	0.004558	9.42E-13	1.9122	0.0356	0.0414
Madison, ID	10608	Custom	anger	0.03281	2.89E-17	2.0306	0.0263	0.0291
Madison, ID	10608	Custom	fear	0.001208	3.07E-14	1.9359	0.0303	0.0494
Madison, ID	10608	Custom	joy	0.001514	8.93E-13	2.0496	0.0292	0.0481
Madison, ID	10608	Custom	disgust	5.39E-05	1.30E-14	**2.1834**	0.0246	0.0674
Madison, ID	10608	Custom	sadness	1.11E-06	1.01E-13	2.0042	0.0305	0.0889
Nez Perce, ID	4000	Custom	anger	0.012702	3.78E-11	2.1201	0.0221	0.0351
Nez Perce, ID	4000	Custom	fear	5.56E-07	8.68E-13	2.1197	0.0269	0.0927
Nez Perce, ID	4000	Custom	sadness	9.86E-07	1.31E-12	1.9396	0.0239	0.0896
Twin Falls, ID	4934	Custom	surprise	0.042209	6.40E-19	**1.7902**	0.0366	0.0274
Twin Falls, ID	4934	Custom	fear	0.000991	1.27E-13	2.1255	0.0253	0.0506
Cascade, MT	21967	Custom	polarity	0.000913	1.32E-14	2.1226	0.0394	0.0511
Custer, MT	214	Custom	sadness	0.00071	8.23E-18	**1.7772**	0.0193	0.0526
Custer, MT	214	Custom	disgust	1.04E-05	1.78E-14	1.8652	0.0195	0.0767
Custer, MT	214	Custom	anger	4.77E-08	3.86E-17	1.9434	0.0209	0.1058
Deer Lodge, MT	213	Custom	anticipation	0.002167	2.62E-18	1.9930	0.0271	0.0459
Deer Lodge, MT	213	Custom	anger	0.000878	3.58E-18	1.9681	0.0316	0.0513
Deer Lodge, MT	213	Custom	trust	8.19E-07	1.37E-12	**1.8007**	0.0255	0.0906
Deer Lodge, MT	213	Custom	disgust	2.67E-10	1.10E-14	**2.2046**	0.0200	0.1326
Deer Lodge, MT	213	Custom	surprise	1.11E-16	3.47E-12	1.9433	0.0264	0.2794

[Table 4.3] Training set results for all counties. P-values ≥ 0.05 bolded.

County	Tweets	Affect	Dimension	F-test	Train DF	Train DW	Train MSE	Train r^2
Fergus, MT	272	Custom	anger	2.03E-07	1.38E-12	2.0973	0.0383	0.0981
Fergus, MT	272	Custom	anticipation	2.26E-08	2.32E-14	2.0602	0.0174	0.1097
Fergus, MT	272	Custom	trust	2.15E-09	2.84E-17	2.1187	0.0340	0.1219
Fergus, MT	272	Custom	fear	1.29E-12	7.45E-10	2.0942	0.0219	0.1591
Fergus, MT	272	Custom	polarity	2.91E-14	1.90E-16	**2.1535**	0.0259	0.1773
Fergus, MT	272	Custom	sadness	1.11E-16	1.18E-12	2.0908	0.0298	0.2035
Fergus, MT	272	Custom	disgust	1.11E-16	4.84E-16	1.9157	0.0212	0.3227
Fergus, MT	272	VADER	polarity	6.07E-06	1.29E-13	2.0544	0.0298	0.0796
Flathead, MT	10593	Custom	joy	0.006435	2.31E-14	1.9585	0.0397	0.0393
Gallatin, MT	22328	Custom	disgust	0.035114	1.50E-12	2.1351	0.0335	0.0286
Gallatin, MT	22328	Custom	anger	0.01211	3.25E-13	1.9132	0.0367	0.0354
Gallatin, MT	22328	Custom	trust	0.001372	2.02E-20	1.9462	0.0348	0.0487
Gallatin, MT	22328	Custom	fear	3.80E-11	1.37E-23	1.8998	0.0300	0.1424
Gallatin, MT	22328	VADER	polarity	0.001059	1.39E-14	2.0774	0.0371	0.0502
Hill, MT	841	Custom	disgust	0.01283	6.44E-16	1.9646	0.0353	0.0350
Hill, MT	841	Custom	surprise	0.000467	9.75E-17	2.0069	0.0289	0.0550
Hill, MT	841	Custom	sadness	1.11E-16	1.26E-12	2.2775	0.0259	0.2080
Hill, MT	841	Custom	trust	1.11E-16	6.18E-12	2.1429	0.0172	0.2796
Hill, MT	841	VADER	polarity	0.022305	2.41E-15	**2.2415**	0.0257	0.0315
Lewis and Clark, MT	4751	Custom	joy	0.001293	6.04E-11	1.9034	0.0312	0.0490
Lewis and Clark, MT	4751	Custom	fear	0.002329	2.17E-12	**2.1735**	0.0290	0.0455
Lewis and Clark, MT	4751	Custom	surprise	1.27E-07	5.61E-14	2.0804	0.0250	0.1006
Missoula, MT	14324	Custom	anticipation	3.09E-07	6.54E-13	2.117	0.0286	0.0958
Park, MT	1641	Custom	fear	9.08E-05	6.05E-12	1.8632	0.0358	0.0645
Park, MT	1641	Custom	anticipation	6.25E-05	4.22E-13	1.9526	0.0273	0.0666
Park, MT	1641	Custom	disgust	3.29E-14	4.40E-18	1.8515	0.0253	0.1768
Richland, MT	871	Custom	disgust	0.03633	1.88E-11	2.0473	0.0192	0.0284
Richland, MT	871	Custom	anger	0.002837	1.30E-14	2.0807	0.0219	0.0443
Richland, MT	871	Custom	anticipation	4.74E-07	6.81E-19	1.9806	0.0279	0.0935
Silver Bow, MT	4784	Custom	disgust	1.08E-06	5.03E-15	1.8585	0.0230	0.0891
Silver Bow, MT	4784	Custom	sadness	1.45E-07	1.35E-13	1.9372	0.0279	0.0999
Yellowstone, MT	21551	Custom	trust	0.005437	7.03E-17	1.9890	0.0355	0.0404
Yellowstone, MT	21551	Custom	anger	0.001337	4.65E-10	1.7523	0.0356	0.0488
Yellowstone, MT	21551	Custom	sadness	0.000524	4.62E-13	2.0336	0.0286	0.0543
Yellowstone, MT	21551	Custom	anticipation	0.000526	1.58E-14	2.1069	0.0350	0.0543
Yellowstone, MT	21551	Custom	disgust	1.22E-09	2.09E-15	2.1128	0.0358	0.1248
Yellowstone, MT	21551	VADER	polarity	0.009171	2.72E-12	2.0677	0.0360	0.0371

[Table 4.4] Training set results for all counties. P-values ≥ 0.05 bolded.

5 CONCLUSIONS

(1) We created a custom scraper because first-party data at commercial rates was infeasible, and third-party sources did not allow us to parameterize our tweet search to the degree we wanted. Containerizing our scraper with Docker on the DigitalOcean cloud decreased our scraping time by orders of magnitude. Even when additional time is permitted, or the scope of data scraping is limited, a single instance of a scraper is better on a cloud platform than in the lab. This is because exceeding rate limits can lead to an IP-based cap lasting multiple days. Twitter's academic research products encourage the discovery of insight but require bandwidth moderation.

(2) Our custom language model could forecast sentiment with a degree of significance as measured through F-tests and test set r^2 for a linear polynomials of PM2.5 using OLS. VADER was effective as a predictor of tweet sentiment in only 6/78 total models where a significant correlation was found. This suggests that VADER is less responsive to PM2.5 than our own sentiment analysis tool. We cannot compare performance of VADER and our affective model in other contexts, but regressing on 9 dimensions of sentiment may have been a key advantage.

(3) The fit between models of different affective dimensions, language models, and counties [tables 4.1-4.4] can be assessed by identifying test and training sets with similar errors. Differences in the fit between individual rows of results are likely imparted by differences in sentiment error. The number of tweets in each varied from 213 in Fergus, MT to 149,716 in Ada, ID indicating different degrees of representativeness. Fewer users in Fergus translates to each having a much greater influence than users in Ada.

(4) Assertions 3 and 6 in 3.4.2 were likely violated resulting in models with variance as a function of time. Future research should perform attribute selection to identify the effect of potential exogenous variables like temperature. Any bag of words model is inflexible and therefore cannot be context aware. It is reasonable to assume this accounts for significant error.

(5) An attention-based language model has the potential to model sentiment with reduced error. This would however rely on the not insignificant task of acquiring training data including multiple affective dimensions. It is also warranted to explore a broader range of hyperparameters and sentiment forecasting model types, especially non-linear.

Although we did not detect AR(1) autocorrelation in most of our results [tables 4.1-4.4], and found that a linear polynomial often outperformed higher order polynomials, but this does not mean a linear model was optimal. The process of making PM2.5 and sentiment data stationary (through first-differencing) prior to modeling has the effect of controlling autocorrelation in the errors. This means our error terms may be more interdependent than they appear. Patterns in the errors suggest insufficient model complexity or excluded variables. The reduction of sentiment error may also have the side-effect of revealing more complex dynamics.

(6) A weak yet significant correlation was found between PM2.5 and Twitter sentiment reliably using our own language model, and occasionally using VADER. Future research may benchmark our affective model with existing NLP tools for multi-domain sentiment analysis to validate the meaningfulness of the dimensions of our model. High bias is not indicated but variance as a function of time is likely due to the violation of error independence and exogeneity. We therefore conclude that PM2.5 is a reasonable predictor of Twitter sentiment.

6 REFERENCES

[1] World Health Organization. (2006). Air quality guidelines: global update 2005: particulate matter, ozone, nitrogen dioxide, and sulfur dioxide. World Health Organization.

[2] Lin, Y.-R., Margolin, D., & Wen, X. (2017). Tracking and analyzing individual distress following terrorist attacks using social media streams: Distress following terrorist attacks. Risk Analysis: An Official Publication of the Society for Risk Analysis, 37(8), 1580–1605.

[3] Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. Journal of Language and Social Psychology, 29(1), 24–54.

[4] Joshi, A., Sparks, R., McHugh, J., Karimi, S., Paris, C., & MacIntyre, C. R. (2020). Harnessing tweets for early detection of an acute disease event. Epidemiology (Cambridge, Mass.), 31(1), 90–97.

[5] Big Data Analytics. Retrieved November 1, 2020, from Ibm.com website: https://www.ibm.com/analytics/hadoop/big-data-analytics

[6] Data Economy: Radical transformation or dystopia? (1/2019). Retrieved December 4, 2020, from www.un.org website: https://www.un.org/development/desa/dpad/wp-content/uploads/sites/45/publication/FTQ_1_Jan_2019.pdf

[7] Data Policy. Retrieved December 4, 2020, from Instagram.com website: https://help.instagram.com/519522125107875

[8] Does Facebook sell my information?. Retrieved December 4, 2020, from Facebook.com website: https://www.facebook.com/help/152637448140583

[9] April, I. Update on the twitter archive at the library of congress. Retrieved December 4, 2020, from Loc.gov website: https://blogs.loc.gov/loc/files/2017/12/2017dec_twitter_white-paper.pdf

[10] Kamperman Sanders, A., Falcão, T., Haider, A., Jambeck, J., LaPointe, C., Vickers, C., & Ziebarth, N. (2018). World Economic and Social Survey 2018: Frontier Technologies for Sustainable Development.

[11] Landguth, E. L., Holden, Z. A., Graham, J., Stark, B., Mokhtari, E. B., Kaleczyc, E., … Noonan, C. (2020). The delayed effect of wildfire season particulate matter on subsequent influenza season in a mountain west region of the USA. Environment International, 139(105668), 105668.

[12] Privacy Policy. Retrieved November 1, 2020, from Twitter.com website: https://twitter.com/en/privacy

[13] Twitter Help Center. (2020, October 29). How to tweet. Retrieved November 1, 2020, from Twitter.com website: https://help.twitter.com/en/using-twitter/how-to-tweet

[14] Twitter Help Center. (2020, August 4). Debunking twitter myths. Retrieved November 1, 2020, from Twitter.com website: https://help.twitter.com/en/using-twitter/twitter-myths

[15] Twitter Help Center. (2020, March 3). How to add your location to a Tweet. Retrieved November 1, 2020, from Twitter.com website: https://help.twitter.com/en/using-twitter/tweet-location

[16] Twitter Help Center. (2020, May 19). Following FAQs. Retrieved November 1, 2020, from Twitter.com website: https://help.twitter.com/en/using-twitter/following-faqs

[17] Twitter Help Center. (2020, May 19). How to access your Twitter data. Retrieved November 1, 2020, from Twitter.com website: https://help.twitter.com/en/managing-your-account/accessing-your-twitter-data

[18] Twitter Help Center. (2020, March 3). About your activity dashboard. Retrieved November 1, 2020, from Twitter.com website: https://help.twitter.com/en/managing-your-account/using-the-tweet-activity-dashboard

[19] Analytics. Retrieved November 1, 2020, from Twitter.com website: https://business.twitter.com/en/advertising/analytics.html

[20] Twitter Analytics. Retrieved November 1, 2020, from Twitter.com website: https://analytics.twitter.com/about

[21] Twitter API. Retrieved December 4, 2020, from Twitter.com website: https://developer.twitter.com/en/products/twitter-api

[22] Limits for academic research. (2020, August 12). Retrieved December 4, 2020, from Twittercommunity.com website: https://twittercommunity.com/t/limits-for-academic-research/140894

[23] Higher levels of access for academic researchers (v2 Twitter API). Retrieved December 4, 2020, from Trello.com website: https://trello.com/c/qGjOCWlB/65-higher-levels-of-access-for-academic-researchers-v2-twitter-api

[24] US Census Bureau. Cartographic boundary files - shapefile. Retrieved from https://www.census.gov/geographies/mapping-files/time-series/geo/carto-boundary-file.html

[25] Shapely. Retrieved December 4, 2020 from GitHub.com website: https://github.com/Toblerity/Shapely

[26] Rasterio: access to geospatial raster data — rasterio documentation. Retrieved December 4, 2020, from Readthedocs.io website: https://rasterio.readthedocs.io/en/latest/

[27] Premium and enterprise operators. Retrieved December 4, 2020, from Twitter.com website: https://developer.twitter.com/en/docs/twitter-api/v1/rules-and-filtering/overview/premium-operators

[28] Twitter Help Center. (2020, November 26). About the notifications timeline. Retrieved December 4, 2020, from Twitter.com website: https://help.twitter.com/en/managing-your-account/understanding-the-notifications-timeline

[29] Twitter Help Center. (2020, March 3). How to add your location to a Tweet. Retrieved December 4, 2020, from Twitter.com website: https://help.twitter.com/en/using-twitter/tweet-location

[30] Twitter Help Center. (2020, March 3). Tweet location FAQs. Retrieved December 4, 2020, from Twitter.com website: https://help.twitter.com/en/safety-and-security/tweet-location-settings

[31] Skyfield. Retrieved December 4, 2020, from Rhodesmill.org website: https://rhodesmill.org/skyfield/

[32] Richardson, L. (2007). Beautiful soup documentation. April.

[33] The selenium browser automation project. Retrieved December 4, 2020, from Selenium.dev website: https://www.selenium.dev/documentation/en/

[34] XML Path Language (XPath). Retrieved December 4, 2020, from Www.w3.org website: https://www.w3.org/TR/1999/REC-xpath-19991116/

[35] How to provision and manage remote Docker hosts with Docker Machine on Ubuntu 16.04. (2016, May 20). Retrieved December 4, 2020, from Digitalocean.com website: https://www.digitalocean.com/community/tutorials/how-to-provision-and-manage-remote-docker-hosts-with-docker-machine-on-ubuntu-16-04

[36] George A. Miller (1995). WordNet: A Lexical Database for English. Communications of the ACM Vol. 38, No. 11: 39-41.

[37] Wolf Garbe. 2019. Symspell. https://github.com/wolfgarbe/SymSpell

[38] Derek Anderson. wordninja, Probabilistically split concatenated words using NLP based on English Wikipedia uni-gram frequencies., 2017. https://github.com/keredson/wordninja.

[39] Saumya Banthia, & Anantha Sharma. (2020). Classification of descriptions and summary using multiple passes of statistical and natural language toolkits.

[40] Ramya Tekumalla, & Juan M. Banda. (2020). Characterizing drug mentions in COVID-19 Twitter Chatter.

[41] M. Munezero, C. S. Montero, E. Sutinen and J. Pajunen, "Are They Different? Affect, Feeling, Emotion, Sentiment, and Opinion Detection in Text," in IEEE Transactions on Affective Computing, vol. 5, no. 2, pp. 101-111, 1 April-June 2014, doi: 10.1109/TAFFC.2014.2317187.

[42] Hutto, C.J. & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.

[43] Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word–emotion association lexicon. Computational Intelligence, 29(3), 436-465.

[44] Google Ngram Exports. Retrieved December 4, 2020, from Googleapis.com website: https://storage.googleapis.com/books/ngrams/books/datasetsv3.html