IPRPD
International Journal of Arts, Humanities and Social Sciences
ISSN 2693-2547 (Print), 2693-2555 (Online)
Volume 03; Issue no 10: October, 2022
DOI: 10.56734/ijahss.v3n10a2

A TUTORIAL ON COLLECTING AND PROCESSING LONGITUDINAL SOCIAL MEDIA DATA

Grace M. Leffler[1], Xin Tong[2]

[1] *Department of Psychology, University of Virginia, USA*

[2] *Ph.D., Associate Professor, Department of Psychology, University of Virginia, USA*

Abstract

Longitudinal research using social media data has been under-explored in social and behavioral sciences. Despite its great potential, longitudinal analysis using social media data faces unique challenges. Researchers must consider many influential factors and incorporate them when designing their studies and conducting analyses. Over the past decade, best practices have originated from both studies focusing on social media data in general and those applying longitudinal designs. This tutorial aims to educate those unfamiliar with such a growing field, outlining the different steps that may exist within data collection, data processing, and data analysis of longitudinal social media data. To illustrate these techniques, we apply our basic steps to a Twitter dataset about the 2020 U.S. wildfires, examining sentiment throughout the wildfire period.

Keywords

Social Media, Longitudinal Text Data, Data Mining, Wildfires

Longitudinal studies help researchers understand changes. There has been a considerable rise in attention paid to longitudinal study design, methodology, and application in many disciplines including but not limited to psychology, education, sociology, economics, political science, and medicine. Over the past few decades, researchers have collected many longitudinal datasets and built various programs with a focus on longitudinal human development (e.g. the National Longitudinal Surveys (NLS; Bureau of Labor Statistics, 1999), the National Longitudinal Study of Adolescent to Adult Health (Add Health; Harris & Udry, 2016), etc.).

Longitudinal research provides valuable insight into causal relationships associated with a variety of topics (e.g. relationships between adverse childhood experiences and mental illness, political violence and war, and climate change and mental health). However, collecting longitudinal data is often difficult and expensive, especially during the coronavirus pandemic with mandated social distancing rules and changes in human behaviors. While mandated social distancing poses additional challenges for in-person data collections, these same mandates and post-pandemic tendencies have promoted the use of social media as a necessary daily activity to express views, connect with and communicate to others. As a result, there are more opportunities for social media data to be used to understand human behaviors and sentiments (e.g. Hswen et al., 2020).We aim to harness the power of the emerging research frontier of combining longitudinal analytic techniques with social media data.

Although social media data contain a large amount of information, there are unique challenges with collecting and analyzing them, especially if longitudinal. While the internet enables researchers to explore relatively new phenomena rapidly and widely, many false equivalencies may occur. For example, those wishing to research mental health must consider that shame and other stigmas related to mental illness vary based on social and economic demographics, meaning that some demographics may be very underrepresented in studies that examine posts or user bios for mentions of a mental illness (Parcesepe & Cabassa, 2013). As far as we are aware, social media data are widely adopted in computer science but less frequently in other fields like social and behavioral sciences, partially because researchers in these fields are less familiar with the procedures of collecting and processing social media text data. Thus, this paper will fill the gap and guide social scientists in using data mining techniques to acquire longitudinal datasets from social media for adequate sampling, research design, and analytical techniques.

A general procedure for implementing longitudinal analysis using social media text data typically encompasses three steps: data collection, data processing, and data analysis. Because there is already rich literature about longitudinal data analytical techniques, we mainly focus on the data collection and data processing steps in this paper. In the following sections, we provide a tutorial for collecting and processing longitudinal social media text data and illustrate these steps using an example with Twitter data drawn from the Los Angeles and Portland wildfires of 2020.

A GENERAL GUIDELINE

1. *Data collection*

Methods behind data collection for longitudinal analysis using social media may vary. A common and straightforward way is to apply for Application Programming Interface (API) credentials from the social media platform itself. Certain platforms do not have a formal API application process, meaning that researchers may need to contact the platform directly to see if they will be supported. Some platforms offer various types of API permissions with different levels of querying capabilities within a certain time period. For example, Twitter offers a standard Sandbox API that allows users to request tweets posted within the previous week after providing minimal documentation of the requestors' identity and the purpose behind their research. The Sandbox Twitter API may be enough for real-time longitudinal studies where researchers follow randomly selected subjects and retrieve their data every week or every few weeks. When researchers are also interested in historical data that are older than a week, an Academic Research Twitter API approval is needed. Unlike Sandbox API, the Academic Research Twitter API option offers more extensive access to tweets at no cost. Additionally, it provides limited access to any tweet stored on Twitter's servers; removed or deleted tweets are inaccessible. To apply for an Academic Research API track, researchers must first have a personal Twitter account. Then, they must apply for a developer account by providing their names, links to webpages that establish their identity, details about the intended types of projects (e.g., the general research area, the purpose and benefit of such research), the methodology behind the projects, and the project distribution methods. After receiving developer permissions, the same type of information will be used to apply to an academic track within the developer portal.

Once obtaining permission to retrieve data, researchers must collect data across an intentional time period. Filters may be applied when data are collected with API. Depending on the platform and API permission type, there may be restrictions on the number of posts and comments that may be retrieved within a certain time period. Certain filters may not be available unless the researchers request to use a research-specific API and/or pay for a premium API type. Filters may be used individually or in conjunction with each other, including geography, username, user profile content, and so on. Additionally, filters can specify information about the tweet source (e.g., a verified Twitter user), the type of online interaction (e.g., a post, comment, or retweet), and/or the content itself (e.g., only tweets containing links, hashtags, or media). Some filters can go so far as to query tweets in response to specified users or on particular tweet threads. The possibilities in using these filters to study particular online conversations or the role of particular online actors cannot be overstated.

Besides using filters, we may also estimate or find users' demographics from social media. There are multiple tools at researchers' disposal. Mislove et al. (2011) estimated users' race using their name and the colloquial meanings. Wang et al. (2019) used the M3-inference model to approximate user's gender and age through multimodal analysis on user's profile image, username, and description. Hswen et al. (2020) chose to include users if they used self-described terms (e.g. “lesbian”, “gay”,”cisgender”, etc.) in their Twitter bios. We may also obtain users' data from other resources. For example, in the real data example in the next section, we retrieved each user's geographic location, based on which the corresponding Air Quality Index (AQI)data was obtained using United States Environmental Protection Agency (EPA) online directory (Environmental Protection Agency, 2020). Some studies lack any of these information at all; however, omitting this step leads to low external validity of the studies.

While longitudinal analysis is feasible with social media data as retrieving a specific user's history is possible with most API platforms, there may be fewer time-bound restrictions on retrieving information in this manner (e.g., API platforms may tend toward providing a user's most recent history). Retrieving social media participants through external platforms may come with added ethical complexities compared to simply data mining within a certain geographic area or using certain keywords, as it allows for informed consent as well as involves retrieving potentially sensitive information along with identifiable information (e.g. a username). Yet as Zhang et al. (2020) stated, although deep learning methods perform better on larger training sets, collection methods that lack rigor may well taint the dataset.

Two potential issues need to be considered in planning the data collection procedure: 1) meaningful time points (MTPs) and 2) potential non-random differences in user type. Creating MTPs depends on the research topic and the variables being analyzed. For example, understanding collective and/or personal trauma may require a longer sampling period than understanding the impact of online affirmations. Examining political-related troll account activity may involve the same number of time points clustered before and after an election. In general, the goal of the study will guide the length of the time period and the number of time points. In addition, the practicality of choosing a time span must also be considered; certain users may only post once over a short time period while others may post hundreds of times. There can be many reasons for their usage level, making it non-random. Individuals may post more often because of a number of reasons such as increased loneliness, increased leisurely time, or higher socioeconomic status. These factors, especially how they apply to the social media platform, must be carefully evaluated in designing the study and determining MTPs. As a general rule of thumb, data mining longitudinal samples will likely require more observations than longitudinal studies conducted through surveys or other traditional measurements to reach a similar statistical power due to the number of different factors influencing participant behaviors and sentiment.

It is important to note that the population selection bias may arise in the data collection procedure and researchers should take this issue into consideration. For example, a Pew Research Center survey indicated that while Twitter users aligned with the broader U.S. population in some political views, these Twitter users differed from the average American on some key social issues, such as whether immigrants strengthen or weaken the country (Wojcik & Hughes, 2019). People may also express themselves differently on social media than what they truly feel. When there is potential population selection bias, alternative methods that have been developed to choose users and/or collect user information outside of the social media platform may assist to acquire more representative sample sets. For example, Wojcik & Hughes (2019) used MTurk, a platform commonly used in social sciences research, to recruit, screen, and measure the self identified demographics of Twitter users before collecting participants' Twitter history. Similarly, Reece et al. (2017) developed a technique of recruiting through MTurk's platform, screening participants, mining their history, and comparing self-reported data to the users' mined Twitter history.

2. *Data processing*

Although there are numerical data, the most important data we collect from social media are text data, for which we need to use text mining techniques to understand content like communication of human emotion. Data processing differs on the type of analysis. In this paper, we highlight data processing with sentiment analysis due to its widespread use with social media data (e.g., Giachanou & Crestani 2016; Zhang et al. 2020; Kim et al. 2021). Sentiment analysis is the interpretation and classification of emotions (positive, negative and neutral) with text data using text analysis techniques. We first clean the text data by removing commas and URL addresses. Bag-of-words sentiment analyses and network analyses will require filler words such as conjunctions (e.g. "and") and articles (e.g. "the") to be removed. Other models may require their inclusion to better understand the sentence structure when analyzing meaning. Stemming is appropriate for many analyses except when the text preprocessing pipeline is used, which requires tokenizing words for deep learning (Baziotis et al., 2017).

The type of sentiment analysis should be selected based on the goal of the study. In general, sentiment analysis techniques may be split into machine learning based, lexicon-based, or hybrid techniques (Giachanou & Crestani, 2016). Machine learning based analysis utilizes training and testing sets, attempting to use either supervised or unsupervised techniques to improve its model. Sometimes, this technique creates very sophisticated understandings of sentiment using linguistic and semantic clues. Lexicon-based methods, on the other hand, compare participant text to a dictionary, typically with hundreds of entries. There are many different lexicons that are specific to the medium (e.g. government documents, blogs) or topic (e.g. stocks, depression) that are sometimes combined with machine learning models. Newer techniques in sentiment analysis on social media data have been applied to account for the emphatic lengthening, sparseness, polarity, commonalities of stop words, and multilinguality in the data, among others. In our example in the next section, we use a bag-of-words model, a simplistic lexicon-based technique using the occurrences of words by category or thematic impression to understand sentiment, regardless of grammar or word order (Giachanou & Crestani, 2016). Alternatives to bag-of-words approaches may be more applicable within certain contexts, especially in understanding thematic meanings or in capturing nuanced messages. Each text type and text source may be better suited by specific types of sentiment analysis techniques; for example, using a technique that accounts for slang or online colloquial phrases may better break down meanings used in social media data, providing a more accurate measurement of sentiment.

3. *Data analysis*

After the data processing, the data structure is longitudinal: sentiment analysis scores for each subject at multiple time points. Note that datasets obtained from social media platforms are unbalanced with individually varying time metrics and contain missing values. Longitudinal data analytical techniques that are used to handle such data should be applied. In addition, a cross validation or an accuracy, recall, and precision approach is often included before correlating sentiment analysis scores with other factors (Kim et al., 2021).

In longitudinal research, growth curve modeling is widely used as it can directly investigate within-subject change over time and between-subject differences in within-subject change. Growth curve modeling can be conducted in the multilevel modeling framework and the structural equation modeling (SEM) framework. Because data collected from social media are unbalanced: participants may have a lack of time point overlap, different numbers of measurements, and unique measurements (McNeish & Matta, 2018). The multilevel modeling framework has several advantages, including that they naturally accommodate time-unstructured or unbalanced data and may investigate nonlinear change patterns (McNeish & Matta, 2018).

Handling of individually varying time metrics in the SEM framework is not straightforward. Definition variables can be used (Mehta & Neale, 2005) so that the factor loading matrix in SEM will be converted based on the definition variables instead of the varying time points. It is also important to note that SEM software can accommodate nonlinearity in terms of the variables but not the parameters, unless structured latent-curve models are used (Blozis & Harring, 2016).

Another approach to consider is the application of deep learning models (Zhang et al., 2020). Zhang et al. (2020) trained data on the textual information, linguistic profiles, demographics, engagement, and big five personality characteristics using a support vector machine, chunking data from different months or time periods to apply to its longitudinal analysis. Some researchers, like Zhang et al. (2020) are of the opinion that insights to true emotion cannot be measured by applying sentiment analysis to one lone social media post. It is important to note that the method in Zhang et al. (2020) was applied to a dataset differentiating between depressed and non-depressed users based on depressed users stating that they had depression on Twitter. Reece et al. (2017) emphasizes that studies using these approaches may simply be tracking users' sentiments as they further identify with their mental health disorder rather than a progression or depressive markers generalizable to all depressed populations; however, this novel approach may prove viable for studies with more rigorous sampling.

AN ILLUSTRATIVE EXAMPLE USING TWITTER DATA

As one of the most popular social media platforms, Twitter has 76.9 million active users in the U.S. as of January 2022 and 52% of Twitter users report using it daily (Statista Research Department, 2022). Because of the popularity of Twitter, in this section, we demonstrate the application of the general guideline described in the previous section with an applied Twitter dataset drawn from the California and Oregon wildfires of 2020. Due to the climate change and drought effects, there has been an increase in wildfires globally. We investigate the longitudinal impact of wildfires near Los Angeles and Portland on human sentiment and hypothesize that people's negative sentiments are associated with wildfire imagery and dramatically reduced air quality over time. The analysis was implemented in R, a free software environment for statistical computing and graphics (R Core Team, 2020). We provide important annotated R code below to illustrate key precedures. All our R programming code is available on our Github site or upon request.

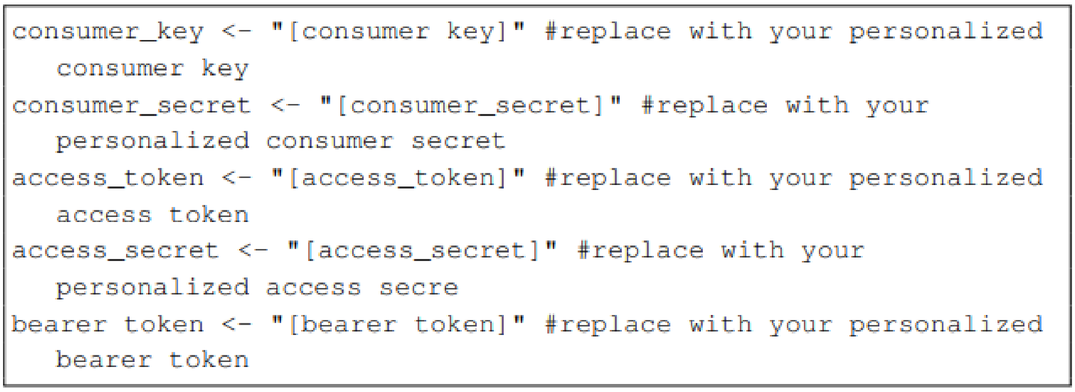
Method

We demonstrate how to collect longitudinal Twitter data, via both the Sandbox API and Academic Research API tracks. In this example, we collected tweets from people in Los Angeles and Portland to study their emotions. If our study was real-time, the Sandbox API could be used. Otherwise, if we intended to study previous changes of sentiment, we could use the Academic Research API. We additionally collected wildfire imagery, AQI, and Twitter bio sentiment data using a personally developed wildfire lexicon, the Liu and Hu lexicon, and the Environmental Protection Agency's dataset (Liu et al., 2005; Environmental Protection Agency, 2020).

Sandbox API. The standard Sandbox accounts use the twitteR packages (Gentry, 2016), which can be installed and loaded as shown below.

```
install.packages("twitteR")
library(twitteR)
```

To gain authentication for this API type, four credentials listed in the developer platform site must be saved to run the Open Authorization (OAuth).



```
consumer_key <- "[consumer key]" #replace with your personalized
   consumer key
consumer_secret <- "[consumer_secret]" #replace with your
   personalized consumer secret
access_token <- "[access_token]" #replace with your personalized
   access token
access_secret <- "[access_secret]" #replace with your
   personalized access secre
bearer token <- "[bearer token]" #replace with your personalized
   bearer token
```

Additionally, the http package must be loaded in order to use these four credentials for final authentication.

```
install.packages("httr")

library(httr)
setup_twitter_oauth(consumer_key, consumer_secret,
access_token, access_secret)
```

Below demonstrates what we searched on September 27 of 2020 to collect tweets created within the past week. An expected count of tweets and a query, such as ’wildfire’, are necessary for the code to run.

```
caliSep <- searchTwitter(’wildfire’, n = 100, since=’2020-09-20’,

until=’2020-09-27’, geocode = "34,-118,60mi", lang="en")
```

This query will generate a dataframe with columns referencing aspects about the tweet (e.g. date and time, screenname, like count) and rows representing each tweet. This data frame can easily be manipulated and saved as a comma-separated values (CSV) file.

Some other useful functions in the twitteR package include ’favorites’, allowing the researcher to choose the most recently favorited tweets of a Twitter user or users, and ’getUser’, allowing the researcher to view the basic information (e.g., followers, tweet count, Twitter bio) of a user or users. The twitteR package functions can easily find relevant tweets as well as information about the Twitter users and their followers. But there are querying and functionality limitations built into the package because it is not meant for researchers. A more in-depth description is provided below for the academictwitteR package, as longitudinal researchers will be more likely to use it for the practicality of having no time-bound restrictions on data.

Academic Research API. A major difference of the Academic Research API from the standard Sandbox API is that additional information is needed to authenticate the researcher. Under the academic track, users must first install and load the academictwitteR package (Barrie & Ho, 2021).

```
install.packages("academictwitteR")
library(academictwitteR)
```

Unlike the standard Sandbox API, the Academic Research API only requires a bearer key to authenticate one’s identity to connect to the API. Users of academictwitteR have the choice of specifying their bearer token only once or specifying it in each query. It is generally recommended to save it only once, as this option stores the token outside of the R console to preserve privacy of information. To only save it once, first run *set_bearer()*. A separate tab with the .Renviron should automatically open. Type *TWITTER_BEARER=YOURBEARERCODE* in the console tab, replacing ’YOURBEARERCODE’ with your unique bearer code. Next, your bearer code should appear in .Renviron’s terminal output. A quick way to check if the token was saved is through using *get_bearer()*, which should return your bearer token as output.

In our example, we randomly sampled participants who tweeted in early September 2020 using the terms “wildfires”, “wildfire”, and “arsonist” in Los Angeles, California and Portland, Oregon. The *get_all_tweets()* function can be used to search for tweets from specific users, with specific key words in the tweet text, or both. Each query requires a tweet count, bearer token, period start date, and period end date to run properly. It greatly helps to specify a data path and a file name in order to generate back-up JavaScript Object Notation (JSON) files in a specific folder and the R *rjson* package can be installed to help manipulate JSON files. If these terms are specified, the academictwitteR query generates two JSON files within the data path argument’s newly created folder, one related to the Twitter users of that query and one related to the tweets themselves. These files cannot be easily converted to CSV files due to their storage of three dimensional data to better compact more data surrounding the tweet entry itself. Also contained within this same folder are the exact query arguments to generate this data. If there are enough tweets that meet these specifications, then there should be two hundred tweets generated as specified by the last line of the query.

Additional specifications, such as geography, tweet content, user, or language can be added to modify the search. Below is an example for two hundred tweets in Portland in the English language that have the terms wildfire or arsonist. Note that the bearer token is simply listed as ’get-bearer()’. The data path is the general folder where the data will be stored. The file name is the specific folder that will be generated, storing the search query, user data, and tweet data. If researchers do not wish to set their bearer token, the ’get_bearer()’ may be replaced with the actual bearer token within this search query.

```
get_all_tweets("wildfire OR arsonist", start_tweets = "2020-09-06 T00:00:00Z",
          end tweets = "2020-09-20T00:00:00Z",
          place="portland", lang="en",
          file = "Wildfiredata/",
          data_path = "initial_portland",
          bearer_token= get_bearer(),
          n = 200)
```

For these participants, we can set our working directory to the folder with the file name we specified above before saving the usersnames of the Twitter users who created the retrieved posts as a new variable.

```
setwd("/FOLDER")
portland.user.data <- fromJSON(file=" data_users_1381238752035663874.json")
users.portland <- as.data.frame(NULL)
for(i    in    1:n){users.portland[i,1]    <-    portland.user.data[1]$users
   [[i]]$username}
for(i    in    1:n){users.portland[i,2]    <-    portland.user.data[1]$users
   [[i]]$description}
```

It is easier in some cases to save specific data from the JSON files and then request more data from these users at different time points than our starting time points, as done below.

```
get_all_tweets(start_tweets = "2020-09-21T00:00:00Z",users=users. portland[,1]
          end tweets = "2020-10-20T00:00:00Z",
          place="portland", lang="en",
          file = "Wildfiredata/",
          data_path = "final_portland",
          bearer_token= get_bearer(),
          n = 10000)
```

After the data are collected, for simplicity, we subset the dataset to exclude press users and any tweets that were just links or news headlines. The final dataset include a total of 588 participants, selected by geo-locating the wildfire tweets to create two study groups: those who had a geo-location within sixty miles of Los Angeles, California (502 participants) or Portland, Oregon (86 participants). 1,403 tweets were collected. Los Angeles represented the bulk of tweets (1,219), followed by Portland (184).

While many user characteristics were included, few of them could be recorded for all or even most participants. The final dataset included user geography, Twitter bio sentiment score, day of tweet, daily Air Quality Index (AQI) of the user's respective city, and the amount of wildfire imagery within the tweet. The AQI data were drawn from the EPA's website for that particular day. The terms used for wildfire imagery included 'fire', 'arson', 'forest', 'burn', 'smoke', 'smoking', 'tree', 'gender reveal', 'reveal party', flam', 'blaz,' and 'contain'. Twitter bio sentiment scores were an especially important factor, lending insight to how positively they view themselves and/or their world view. The dependent values, emotion, were obtained using Emoxicon analysis, yielding sentiment scores for sadness (SAD) (Golino, 2018). A subset of the dataset is provided in Table 1. The full dataset is available upon request.

Participant ID	SAD Score	Day	AQI Value	Bio.Score	Wildfire Imagery	Geography	Word count
140	0	28	166	0	0	Los Angeles	4
541	0	11	201	0	0	Los Angeles	12
492	1	6	107	0	1	Portland	19
433	0	11	179	-1	1	Los Angeles	6

Table 1: A subset of the example dataset

Because longitudinal data analytical techniques are not the focus of this tutorial, for the demonstration purpose, we simply applied two growth curve models in the multilevel modeling framework to investigate the underlying trend of SAD scores. Model A was a linear unconditional growth curve model. Model B, on the other hand, was a quadratic growth curve model. In practice, the change pattern may be different from those in the two models we applied and needs to be carefully examined. After the growth pattern is determined, we added bio scores, wildfire imagery, AQI scores, and geography as covariates into the model to explain the between-subject differences in the SAD scores.

Results

Data analysis was conducted in R. As the more complex Model B did not fit the data substantially better than Model A (the AIC for Models A and B are 1924.53 and 1920.39, respectively), Model A was selected for the parsimonious purpose. However, as we pointed out previously, this is just for demonstration and the true change pattern may be different from those in the two models we applied. In fact, the trajectory plot of the SAD scores (Figure 1) indicates that some factors may be creating both positive or negative relationships and potentially indicating a mixture of different patterns instead of a single linear pattern. Given the results for Model A, the average intercept was 0.395 and the average slope was -0.009. There was a general and significant decrease in SAD scores for participants over this period (*p* = .00).

Due to the large between-subject variations in the latent slopes, we added four covariates to explain the between-subject differences in the change of SAD scores. Specifically, Air Quality Indexes (AQI) for that day, wildfire imagery, participant bio sentiment scores, and geography were added to Model A. The correlations among them were tested for multicollinearity. Using a threshold of 4 in analyzing variance inflation factors, none of the variables were dropped. This model demonstrated significant effects of geography and wildfire imagery on the latent intercept (β = 0.33, *p* = .020; β = 0.35, *p* = .012), and significant effect of geography on the latent slope (β = -0.02, *p* = .010). Bio scores and AQI values did not significantly affect both the initial value and the rate of change for the SAD scores.

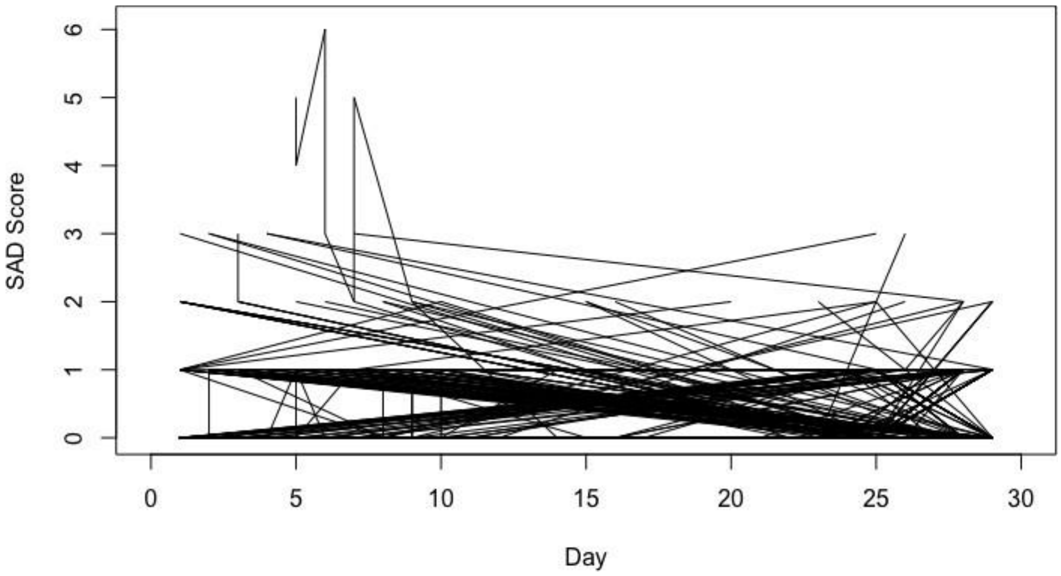


Figure 1. Trajectory plot for the SAD scores

CONCLUSION

This example demonstrated the applicability of longitudinal analysis to Twitter data. Although the general decrease of SAD sentiment in Tweets seemed small (average slope=-0.009), the variance of the latent slopes could not be ignored. There was a combined effect of time with geography on participants' SAD scores. As time progressed, Portlandians were more likely to have reductions in the amount of SAD sentiment conveyed through their tweets compared to Angelenos. There are many explanations for this phenomenon. One main explanation is that California had more extensive and longer lasting wildfires, as indicated by Los Angeles and Portland having mean AQI values of 162.8 and 86.29 over this time period, respectively. As their respective fires began and progressed at different rates, other models and data collection should further investigate how to track wildfires' effects.

Although AQI values were not found to be significant within the final model, they may be significantly related to sentiment in models with more expansive time periods, especially days when the wildfires were totally or almost extinguished. One reason is that there may be no significant relationship between Twitter sentiment and AQI values when AQI values are already elevated or abnormal. At AQI values of 100 or above, people's daily lives may be impeded enough that further fluctuations in air quality produce no further negative sentiments. Thus, there may be a certain threshold in air quality at which there is no noticeable change in sentiment through social media posts.

Additionally, this analysis demonstrated that tweets containing wildfire imagery were more likely to be sad. Tweets that focus more on the realities of the fire—both in its devastation and impact on daily life—are more likely to convey sadness. While other variables were not significant, this finding indicates that Tweet subjects may very much contribute to the measurement of participants' sentiment. Interestingly, there was no significant effect of wildfire imagery on the change of SAD sentiment. Further studies on natural disasters may explore the role of news consumption in perhaps elevating or maintaining negative affect related to natural disasters as well as the trajectory of sentiment on the disaster over a prolonged period of time.

Note that as demonstrated in Figure 1, there may be different change patterns for SAD scores among the participants. Most participants had 2-3 tweets over the data collection period. Collecting more samples per participant may improve the precision of the estimates of the trajectories as well as increase the likelihood of understanding whether the covariates explain the sentiment scores.

DISCUSSION

With social media data, longitudinal studies can be conducted virtually, which may substantially benefit researchers in social and behavioral sciences, especially during and following the mandated social distancing periods. More importantly, using social media data may allow unique access to overlooked groups and participants who may be more distrustful of institutions and authorities. Although virtual longitudinal survey samples may have higher discontinuation among higher stress populations, high online usage is also positively related to stress (Rübsamen et al., 2017). However, there are also certain limitations to obtain longitudinal data using social media. Social media platforms may grant access to individuals who are not representative of their demographics; for example, those indicating a marginalized identity on social media may be more susceptible to adverse mental health due to mediating factors (Stanton et al., 2017). Specifically, Twitter demographics highly skew toward individuals who are more liberal, younger, higher socioeconomic status, more highly educated, and less attached to their community than the average American (Wojcik & Hughes, 2019). People may also express themselves differently on social media than in their normal lives. Although sampling bias correction methods for longitudinal data has been developed (e.g., Mazen & Tong, 2020), the application of these methods to correct for selection bias in social media data needs to be further investigated and examined.

Other practical limitations include the methodology behind and ethical concerns related to the data. There are missing variables and coding difficulties including the potential for human error and a skewed training set model. Data drawn from these accounts should be taken as a mixture of identities, persons, and experiences. There are also many ethical concerns with data that is technically a part of the public record due to it containing potentially sensitive material, not involving active consent on the behalf of participants, and being a part of a changing public record. In addition, text mining methods are typically used for social media text data analysis. However, the accuracy of these methods is not always ideal, so the measurement reliability is not guaranteed. With the development of new reliable text mining methods in the future, the longitudinal analysis of social media data can also be improved.

Future research is tasked with making these steps more rigorous and better understanding the connections between longitudinal techniques and social media sentiment. Firstly, the field could benefit from studies that examine the practicality of these findings, relating to temporal emotions and overall moods to online sentiment. Finding the best types of sentiment analysis to apply to each social media platform or in investigating different topics will prove essential. A realistic understanding of how geographical vocabulary, the strengths of lexicons in different subjects, and vocabulary may impact results by social media platform may be necessary before results from these studies may truly be understood. Lastly, and most importantly, more research must be done into which demographics use each platform, how each demographic uses the platform, and how to collect representative demographics. With time, the underlying methodology for these studies will and should change.

Works Citation

Barrie, C. & Ho, J. C. (2021). academictwitter: An r package to access the twitter academic research product track v2 api endpoint. https://doi.org/10.5281/zenodo.4714637.

Baziotis, C., Pelekis, N., & Doulkeridis, C. (2017). Proceedings of the 11th International Workshop on Semantic Evaluation. In *Association for Computational Linguistics* (pp. 747–754). http://dx.doi.org/10.18653/v1/S17-2126.

Blozis, S. A. & Harring, J. R. (2016). On the estimation of nonlinear mixed-effects models and latent curve models for longitudinal data. *Structural Equation Modeling: A Multidisciplinary Journal*, 23(6), 904–920. https://doi.org/10.1080/10705511.2016.1190932.

Bureau of Labor Statistics (1999). National Longitudinal Survey of Older Men, 1967-1990 (rounds 1-13). https://www.nlsinfo.org/investigator/pages/search?s=NLSM.

Environmental Protection Agency (2020). Air quality index report. https://www.epa.gov/outdoor-air-quality-data/air-quality-index-report.

Gentry, J. (2016). https://cran.r-project.org/web/packages/twitteR/twitteR.pdf.

Giachanou, A. & Crestani, F. (2016). Like It or Not: A Survey of Twitter Sentiment Analysis Methods. *ACM Computing Surveys*, 49(2), 1–41. http://dx.doi.org/10.1145/2938640.

Golino, H. (2018). Emoxicon: Combining data mining and Rasch modeling to identify emotions in texts. In *Pacific-Rim Objective Measurement Symposium*. https://proms.promsociety.org/2018/sessions/emoxicon-combining-data-mining-and-rasch- modeling-to-identify-emotions-in-texts/.

Harris, K. M. & Udry, J. R. (2016). National Longitudinal Study of Adolescent to Adult Health (Add Health. https://doi.org/10.3886/ICPSR21600.v21.

Hswen, Y., Zhang, A., Sewalk, K. C., Tuli, G., Brownstein, J. S., & Hawkins, J. B. (2020). Investigation of Geographic and Macrolevel Variations in LGBTQ Patient Experiences

Longitudinal Social Media Analysis. *Journal of Medical Internet Research*, 22(7). https://doi.org/10.2196/1708.

Kim, J., Lee, D., & Park, E. (2021). Machine Learning for Mental Health in Social Media: Bibliometric Study. *Journal of Medical Internet Research*, 23(3). https://doi.org/10.2196/24870.

Liu, B., Hu, M., & Cheng, J. (2005). Opinion observer: Analyzing and comparing opinions no the web. In *Proceedings of the 14th international World Wide Web conference (WWW-2005)*: International World Wide Web Conference Committee.https://www.cs.uic.edu/ liub/publications/www05-p536.pdf.

Mazen, J. A. M. & Tong, X. (2020). Bias Correction for Replacement Samples in Longitudinal Research. *Multivariate Behavior Research*, (pp. 1–23). https://doi.org/10.1080/00273171.2020.1794774.

McNeish, D. & Matta, T. (2018). Differentiating between mixed-effects and latent-curve approaches to growth modeling. *Behavior Research Methods*, 50, 1398–1414. https://doi.org/10.3758/s13428-017-0976-5.

Mehta, P. D. & Neale, M. C. (2005). People are variables too: Multilevel structural equations modeling. *Psychological Methods*, 10(3), 259–284. https://psycnet.apa.org/doi/10.1037/1082-989X.10.3.259.

Mislove, A., Lehmann, S., Ahn, Y.-Y., Onnela, J.-P., & Rosenquist, J. N. (2011). Understanding the demographics of twitter users. In *Fifth international AAAI conference on weblogs and social media*, volume 20 (pp. 554–557).: Association for the Advancement of Artificial Intelligence. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/viewFile/2816/3234.

Parcesepe, A. M. & Cabassa, L. J. (2013). Public stigma of mental illness in the United States: a systematic literature review. *Administration and policy in mental health*, 40(5), 384–399. https://doi.org/10.1007/s10488-012-0430-z.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Reece, A. G., Reagan, A. J., Lix, K. L. M., Dodds, P. S., Danforth, C. M., & Langer, E. J. (2017). Forecasting the onset and course of mental illness with Twitter data. *Nature Scientific Reports*, 7(13006). https://doi.org/10.1038/s41598-017-12961-9.

Rübsamen, N., Akmatov, M. K., Castell, S., Karch, A., & Mikolajczyk, R. T. (2017). Factors associated with attrition in a longitudinal online study: results from the HaBIDS panel. *BMC Medical Research Methodology*, 17(132). https://doi.org/10.1186/s12874-017-0408-3.

Stanton, A. G., Jerald, M. C., Ward, L. M., & Avery, L. R. (2017). Social media contributions to strong Black woman ideal endorsement and Black women’s mental health. *Psychology of Women Quarterly*, 41(4), 465–478. https://doi.org/10.1177

Statista Research Department (2022a). Frequency of Twitter use in the United States as of 3rd quarter 2020. https://www.statista.com/statistics/234245/twitter-usage-frequency-in-the-united- states

Statista Research Department (2022b). Leading countries based on number of Twitter users as of January 2022. https://www.statista.com/statistics/242606/number-of-active-twitter-users-in- selected-countries/.

Wang, W., Hale, S., Adelani, D. I., Grabowicz, P., Hartman, T., Flöck, F., & Jurgens, D. (2019).

Demographic inference and representative population estimates from multilingual social media data. In *The World Wide Web Conference (WWW ’19)*. https://doi.org/10.1145/3308558.3313684..

Wojcik, S. & Hughes, A. (2019). Sizing Up Twitter Users. *Pew Research Center*. https://www.pewresearch.org/internet/2019/04/24/sizing-up-twitter-users/.

Zhang, Y., Lyu, H., Liu, Y., Zhang, X., Wang, Y., & Luo, J. (2020). Monitoring Depression Trend on Twitter during the COVID-19 Pandemic: Observational Study. *Social and Information Networks*. https://arxiv.org/abs/2007.00228.