Analysis of the Impact of COVID-19 on Education Based on Geotagged Twitter

Zhu Wang
ADVIS Lab
Department of Computer Science
University of Illinois at Chicago
zwang260@uic.edu

Isabel F. Cruz
ADVIS Lab
Department of Computer Science
University of Illinois at Chicago
ifcruz@uic.edu

ABSTRACT

More than 150 colleges have reported hundreds of COVID-19 confirmed cases over all the states as the campuses have reopened and the schools have resumed in-person classes, after switching overnight to online teaching in the spring. We conduct a large scale study on education by using a geotagged Twitter dataset, which spans the whole U.S. during parts of the spring, summer, and fall terms of 2020. We analyze the temporal and spatial patterns of COVID-19 cases. Then, we conduct content and sentiment analysis to discover which topics and which thoughts people located at U.S. colleges and universities are communicating.

CCS CONCEPTS

• **Information systems → Spatial-temporal systems**; **Online analytical processing**; *Social tagging systems*;

KEYWORDS

Geospatial Data integration, Semantics, Uncertainty, Range Queries

ACM Reference Format:
Zhu Wang and Isabel F. Cruz. 2020. Analysis of the Impact of COVID-19 on Education Based on Geotagged Twitter. In *1st ACM SIGSPATIAL International Workshop on Modeling and Understanding the Spread of COVID-19 (COVID-19), November 3, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3423459.3430756

1 INTRODUCTION

The COVID-19 pandemic has had a widespread impact on all aspects of people's life in the last months. In particular, the education domain has been affected in the United States due to school and university closures and a subsequent move to online platforms. This move, which in many circumstances occurred practically overnight, has been a challenge to students, parents, and faculty. However, new challenges have been directly connected with the pandemic itself. University campuses became the new hot spots for COVID-19 in the fall semester [6], because many schools are providing in-person classes since August.

In a study, "203 'college town' counties where students comprise at least 10% of the population found that about half had experienced their worst weeks of the pandemic as students returned in August, and about half of those were experiencing peak infections" in September [6, 22]. What are the attitudes displayed and emotions felt by people in those towns, counties, and regions as they witness directly new infection peaks?

Because of the social distance practice in the era of COVID-19, social media platforms such as Twitter and Facebook can potentially be a conduit of choice to express and share opinions, emotions, and other life aspects related to the pandemic for those involved in academic life and for citizens in general. In this paper our focus is on Twitter. In fact, there is a wealth of millions of tweets with different times, locations, and users, in what constitutes a stream of snapshots throughout the pandemic, which we analyze in this paper, following the lead of so many others who have analyzed tweets in a variety of domains [3, 4, 7, 9, 10, 12, 13, 15, 16, 21, 23].

The emergence of Twitter brought important changes to how health and urban information are discovered and transmitted. When Twitter started in 2006, community health centers and healthcare providers in Chicago were communicating flu cases by *fax* and New York City had started their 3-1-1 call service for citizens to communicate non-emergency situations to local government, just three years earlier. Nowadays, researchers and healthcare entities rely on Twitter to detect flu trends and conduct disease surveillance [4, 9]. As for the communication of emerging urban situations, the 311 service can now be accessed via Twitter in many cities across the U.S. and is recognized as an important citizen participation tool [5]. Because tweets have an associated time and often spatial information, they lend themselves well to exploring the sentiments of people in different regions as the disease evolves in time. In the case of education, time and space are critical because different schools may have different semester or quarter periods, henceforth named *terms*, and distinct policy and regulations in each region or state.

Because tweets have an associated time and often spatial information, they lend themselves well to exploring the sentiments of people in different regions as disease evolves in time. In the case of education, time and space are critical because different schools may have different semester or quarter periods and distinct policy and regulations, in each region or state. The IEEE data port [8] provides a dataset of geotagged tweets including over 6 billions of COVID-19 related tweets up to now, 5% being education related worldwide. We extracted two time periods of tweets in states across the U.S. as follows. The former is from March to July including the spring and first summer terms, and the latter from August 24 to September 5 covering either the first week or two of the fall term.

COVID-19, November 3, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-8168-0/20/11…$15.00
https://doi.org/10.1145/3423459.3430756

In this paper, we collect and analyze at a large scale dataset of geotagged tweets to explore the reactions of people to the COVID-19 pandemic in the education domain, containing 673,601 tweets in the above two time periods. We study the temporal and spatial patterns and compare the density of tweets with the confirmed COVID-19 cases at U.S. colleges and universities. Then, we conduct content and sentiment analysis to discover which topics and which thoughts people are communicating.

This paper is organized as follows. Section 2 introduces data collection and pre-processing for our dataset. In Section 3, we demonstrate the methods applied in this study, including spatio-temporal patterns analysis, sentiment analysis, and content analysis. In Section 4, we present briefly some of the existing techniques that analyze social media (such as Yelp or Twitter). Then, we conclude the paper and discuss future research in Section 5.

2 DATASET

In this section we describe our dataset and how we process the dataset prior to the analytic steps.

2.1 Data collection

We use the Twitter API to hydrate tweets from the Coronavirus dataset that are written in English and occur in the U.S. from March to July and from August 24 to September 5. Also, we will be using tweets that are either geotagged or have otherwise associated with them the users' location, as provided by their profile, either in location or description. Then, we select tweets related to the education using a pre-defined keyword set: `school`, `course`, `class`, `student`, `teach`, `exam`, `educate`, `education`, `campus`, `university`, `college`, `tuition`, `learn`, `study`, `quiz`, `midterm`, `homework`, and `assignment`. That is, if a tweet contains one or more words from this set, they will be included in our study. Finally, we collected 560,780 tweets from March to July and 112,821 tweets from August 24 to September 5 from 265,654 users.

2.2 Pre-processing

We start by cleaning the dataset by organizing the geo-information, for example, we add to the users' displayed city or state information regarding latitude and longitude or vice versa. We use the Google Map API to get both the information on the coordinates and on the cities.

Raw data from the text of the tweets are noisy, thus leading to a sparse vector space and an increase in run time and storage. To address this problem, the text pre-processing tasks involve several steps: (1) standardizing the corpus to change all upper case letters to lower cases, (2) removing non-English words, URLs, special characters and stopwords, including the stopwords both from the dictionary and from a personalized set, in our case common words related to COVID-19, such as `covid` and `virus`, thus reducing the sparsity of the vector space, (3) tokenizing sentences or less structured text into a word level corpus. The NLTK toolkit[1] is used to perform the pre-processing of the text.

3 METHODS

In this section, we introduce the explanatory data analysis we will be performing in our study. First, we summarize the characteristics of the Twitter users, which users state their roles as parents, students, teachers or professors, and official accounts of the schools in their profiles and tweets, while others are not specified (see Table 1). Then, we study the spatio-temporal patterns of the tweets in our dataset, such as their hourly distribution in different days and their density and daily distributions in different states. Last, we study deeply the texts in the tweets to classify users' attitudes as positive or negative in different regions, and to extract the topics that were discussed on Twitter.

3.1 Spatio-temporal patterns

We explored the daily distribution of tweets in the various periods of the semesters from the different states, because we noticed that different schools have different academic calendars. Next, we display the hour distribution in each day of the week. Then, we compare the density of tweets on the map with the COVID-19 college and university cases tracker using the subset of college and university data.

Figures 1 and 2 show the different peaks of tweet counts in various states. California, Texas, New York and Florida generated the largest number of tweets in the two time periods. Each state has a similar increasing/decreasing pattern over time. The first peak in Figure 1 was around the start of the outbreak when many schools began closing their campus and moving to online classes. Other peaks in both periods correspond to the beginning and end of each term.

Hourly tweet distributions are shown in Figures 3 and 4. Figure 3 is about the spring and summer terms from March to July and Figure 4 is about the fall term. We converted the UTC time zone to local time zone for each tweet according to the spatial information given by the tweet. The interesting finding from the slight difference between the two terms is that the percentage of tweets from 1am to 3am in the spring term is larger than in the fall term, which we explain based on the intense workload of the final weeks of the term in comparison with the first couple of weeks in the fall term.

We extract the city and coordinate information of the tweets corresponding to the locations of colleges and universities, and display the volume of tweets on the U.S. map in Figure 5. According to the New York Times, a large number of new confirmed cases continues to emerge on college campuses [6, 22]. The Times provides an updated tracker map with the confirmed COVID-19 cases in colleges and universities in Figure 6. We found that the location distributions and density of tweets display a similar pattern to the confirmed cases in our college map, which attest to the relationship

Table 1: Summary of users' characteristics.

Roles	Counts	Percentage(%)
Parents	436	0.16
Student	141,568	53.29
Teacher/Professor	262	0.09
Official account	36	0.01

[1] http://www.nltk.org/

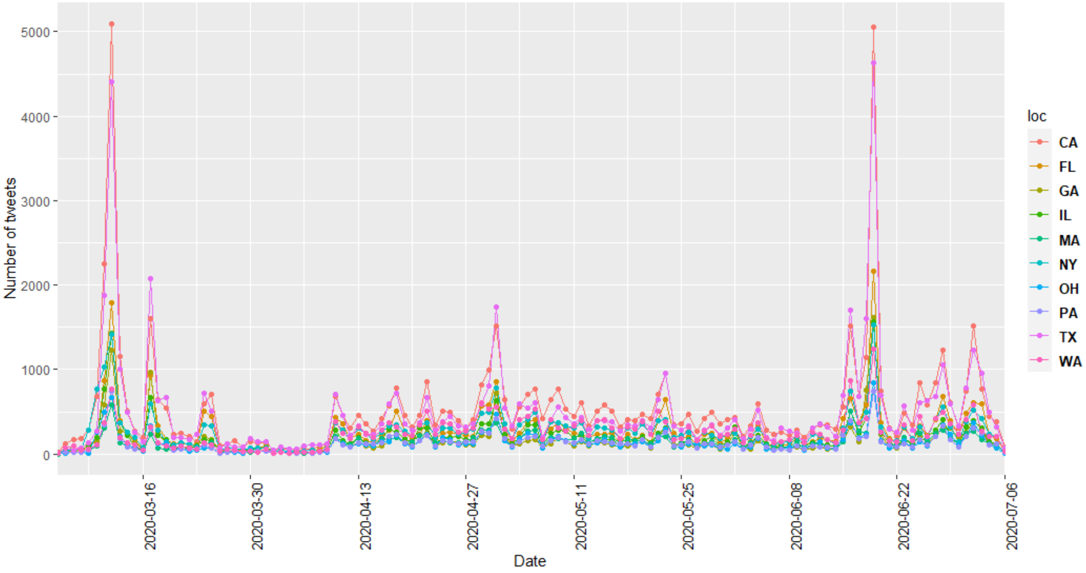


Figure 1: State-level tweet counts from March to July in the 10 states that generate the largest counts.

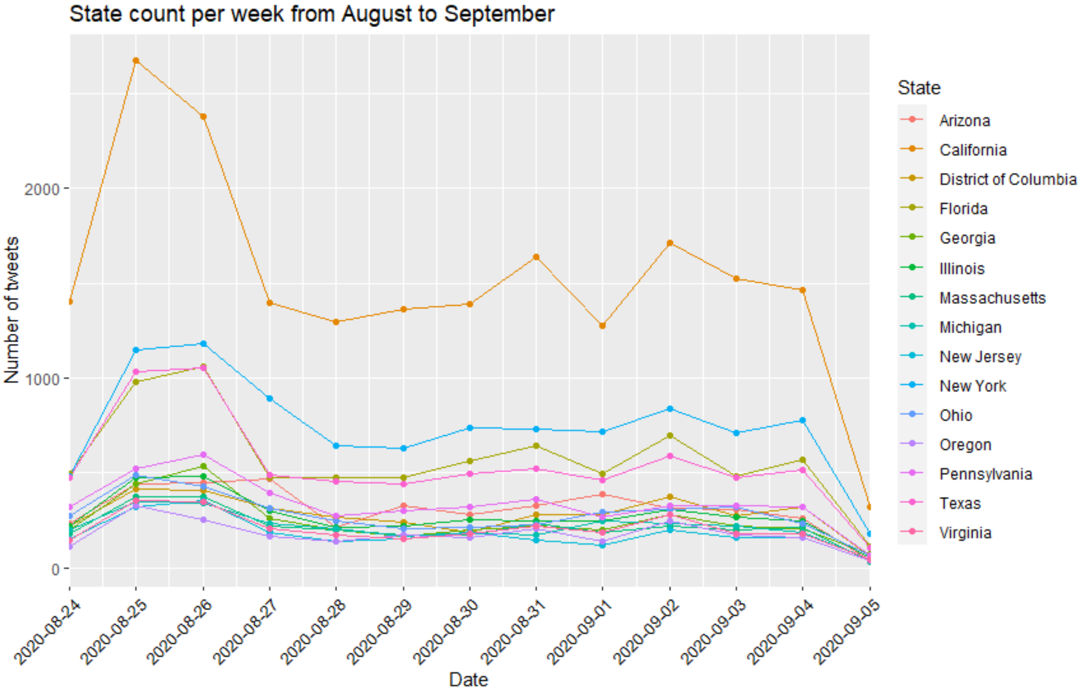


Figure 2: State-level tweet counts from August 24 to September 5 in the 15 states generating the most tweets.

between intense reactions and confirmed cases. A better comparison would be between the average of the number of cases per person compared with the average of the number of tweets per person. Otherwise one may wonder if the number of tweets is higher because the density of the population is also higher. However, the latter average is difficult to determine as a person may have different Twitter accounts and each person can post any number of tweets even if only using one account.

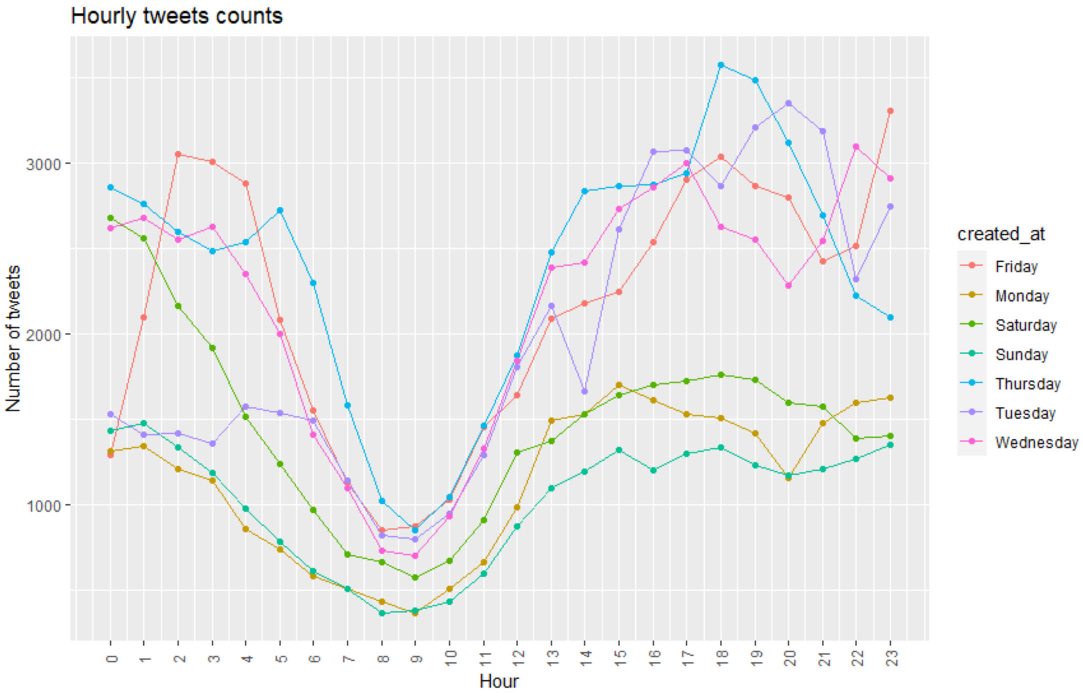


Figure 3: Hourly tweet counts in spring.

3.2 Sentiment analysis

People express their opinions about their daily life and events on micro-blogs such as Twitter, therefore, leveraging sentiment analysis of micro-blogs is beneficial to understand the users' attitudes towards the COVID-19 pandemic, and how it affects different domains, in our case the education domain. In this study, we deploy a dictionary based on sentiment method without having a ground truth corpus available. The sentiment dictionary we use is the *Language Assessment by Mechanical Turk, labMT* [3].

The polarity of a tweet, that is the emotion it expresses, can be determined using the labMT dictionary, which provides a happiness score for each word from 1 (sad) to 9 (happy). By computing the numeric average of the happiness score of all the words in a tweet, we determine its sentiment category as positive, neutral, or negative. We examine the convergent validity of negative tweets against the COVID-19 confirmed cases from colleges and universities. As a result, the scores obtained using labMT correlate positively with the confirmed positive cases with value 0.45.

Figure 7 illustrates the number of positive, neutral, and negative tweets in the 15 states that have the largest number of confirmed COVID-19 cases in colleges. Unexpectedly the number of positive tweets exceeds, for all states, the number of neutral and the number of negative tweets. In some states, the number of positive tweets even exceeds the sum of the neutral and negative tweets. Since we applied dictionaries to detect sentiment, a word like positive is a strong indicator of a positive sentiment. It also happens to be one of the words that contributes the most to the number of positives tweets in Figure 7, even if the word positive appears in positive cases or test result is positive, which are far from expressing a positive sentiment.

Moreover, we observe the daily sentiment—positive, neutral, or negative—using the percentage of the number of tweets for each sentiment with respect to the total number of tweets in the first time period—which is depicted in Figure 8. There are outliers in the different periods. The first two peaks (high values) of negative sentiment appear at the beginning of the pandemic when many schools start cancelling classes. From July 18 to July 20, there is an even higher negative peak. A possible explanation is that many states went into reopening phase around these dates, but students and faculty still did not feel safe to return to the campus.

To better understand the attitudes and opinions of users, we list the 10 most frequent words in positive and negative tweets in California and Illinois in Figures 9(a) and 9(b). There are many common words in these two states. The word debt appears frequently, which seems to indicate that students are concerned with their loans, since they need to pay the same tuition for online classes. Indeed, there are many discussions about tuition and expenses.

3.3 Content analysis

In this section, to further understand the social and psychological meaning associated with tweets, we use the *Linguistic Inquiry and Word Count (LIWC2015)* dictionary [2] preceded by *Latent Dirichlet Allocation (LDA)* [1] to infer coherent topics.

Since ground truth data are not available, we apply topic modeling as an unsupervised probabilistic method to discover latent topics. We treat the text of each tweet as a document to create

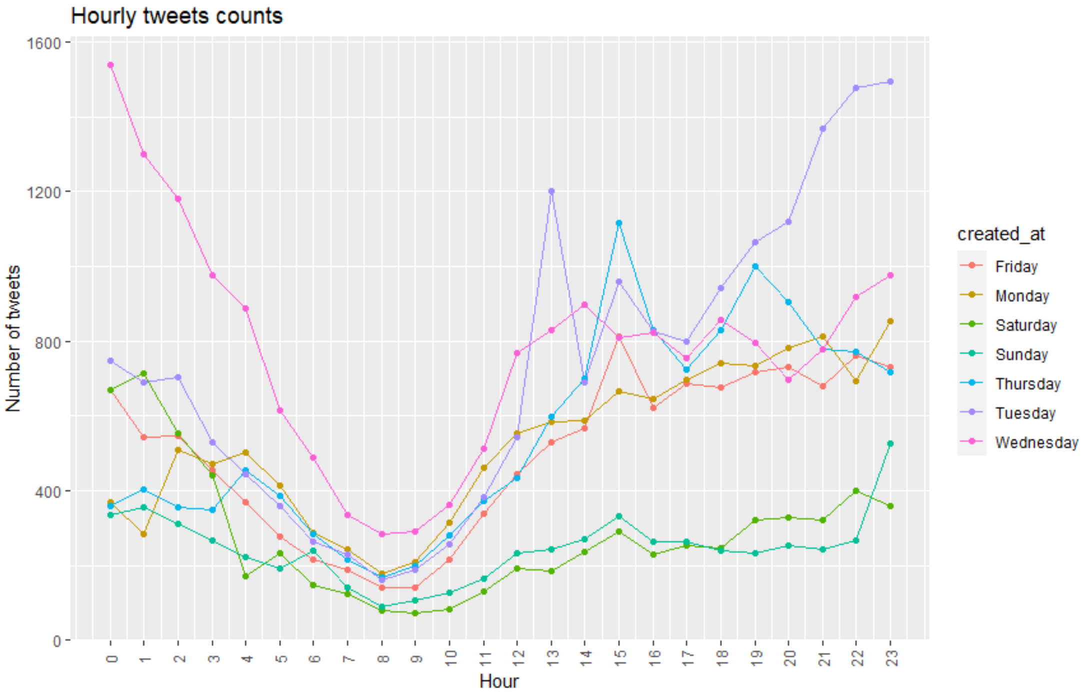


Figure 4: Hourly tweet counts in fall.

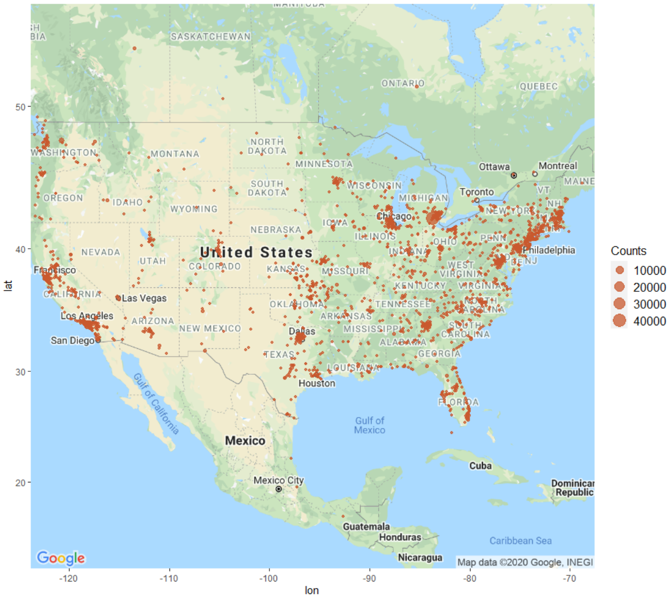


Figure 5: Density of city-level tweets.

the corpus, and we apply lemmatization to improve the model performance by keeping nouns, verbs, adjectives and adverbs. In this study, we mainly focus on the education domain, so we have fewer latent topics than a Twitter dataset that includes many other topics. We determine the optimal number of topics, n, topic coherence, an

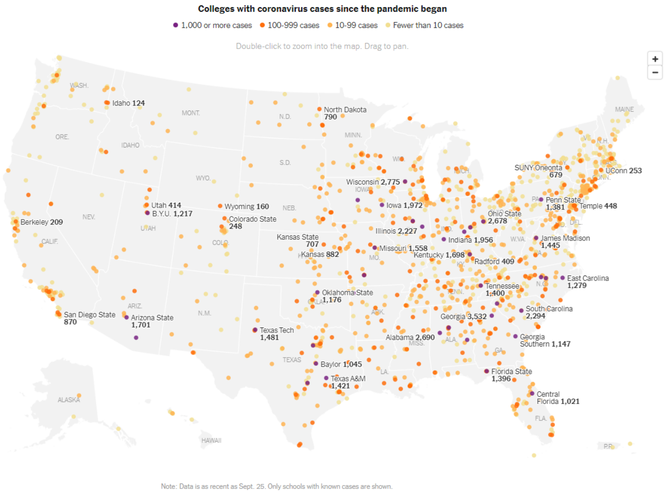


Figure 6: Colleges with confirmed COVID-19 cases [20].

important metric in LDA [11], with $n = 4$ yielding the highest coherence of 0.53. We examine the produced topics and the associated keywords using the pyLDAvis package [18], which displays LDA results in an interactive chart, to verify that the discovered topics are reasonable. In Figure 10, each circle represents a topic in the inter-topic distance map. Our discovered topics are indeed reasonable because the topic circles are fairly big and non-overlapping.

Table 2 shows a sample of the topic results containing the top 10 most probable words in each topic. Topics appear coherent even

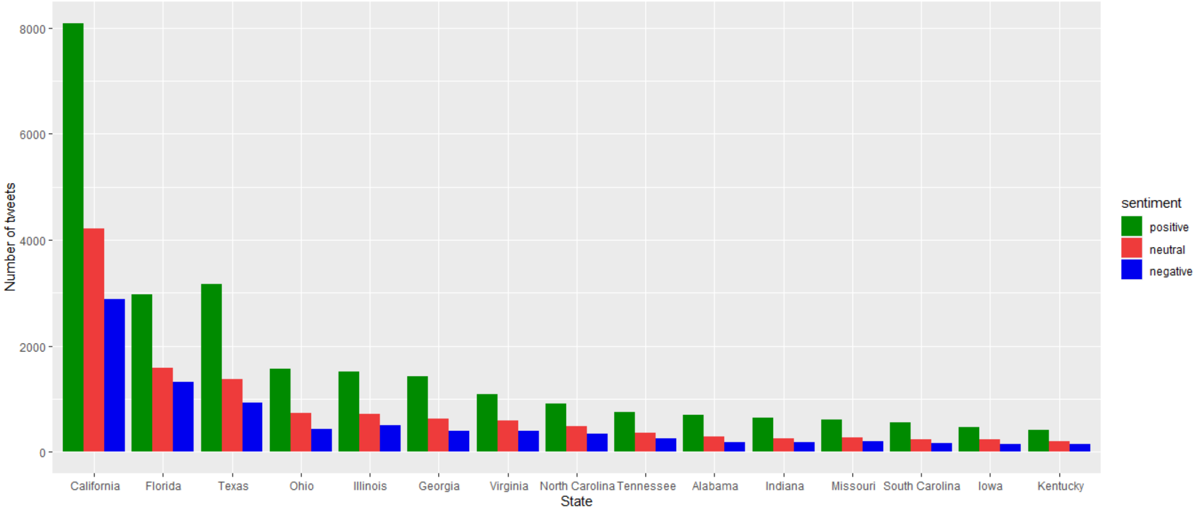


Figure 7: Sentiment analysis by state.

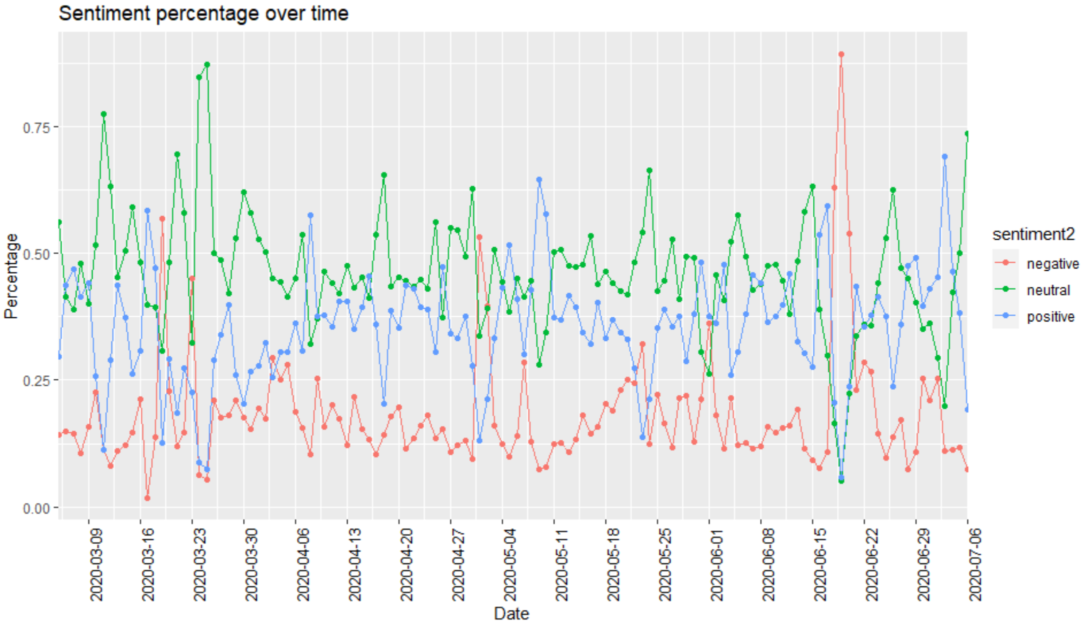


Figure 8: Daily sentiment in all states.

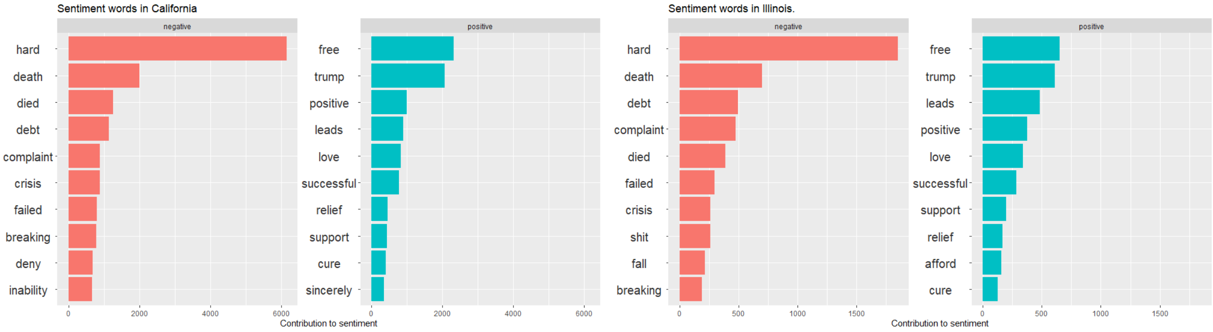
though some of words seem somewhat random, such as `free` and `today`. Topic 1 contains K12 and high school related words, such as `kid, child` and `high`, which appear in the group with high probability scores assigned by LDA.

High-level education is discussed in Topic 2, so we infer that college students are concerned with funding and tuition issues, as already explained in Section 3.2.

To go beyond the classification of tweets as positive, neutral, or negative, we use the 2015 Linguistic Inquiry and Word Count (LIWC15) to link daily word use to psychologically meaningful categories [19]. LIWC is able "to show attention focus, emotion, social relationships, thinking styles, and individual differences" [19]. We call these categories *language features*, and concentrate on `anxiety`, `positive emotion`, `family`, and `friend` (see Figure 11).

For each feature, r, the LIWC dictionary gives us a set of words $w_1, w_2, \ldots, w_j$ associated with that feature. For example, the words `nervous, afraid, tense` are a subset of the words associated with `anxiety` [17]. We can compute the score of r for the corpus of tweets T [2] in each region (for example, state, city), as an average:

$$score(r) = \frac{\sum_{k=1}^{|T|} \frac{s_{t_k}}{|t_k|}}{|T|} \tag{1}$$



(a) California

(b) Illinois

Figure 9: Top sentiment words in California and Illinois.

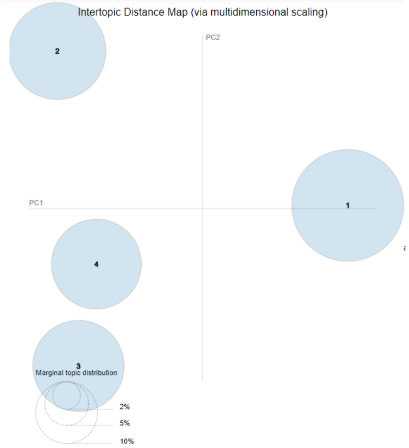


Figure 10: Intertopic distance map of the LDA results as obtained using pyLDAvis [18].

Table 2: Topics resulting from LDA.

Topic 1	Topic 2	Topic 3	Topic 4
school	learn	test	case
go	campus	help	time
get	online	even	day
kid	college	change	start
year	classroom	return	see
take	university	reopen	good
child	job	free	feel
high	tuition	try	share
teacher	fund	grow	home
risk	lab	believe	today

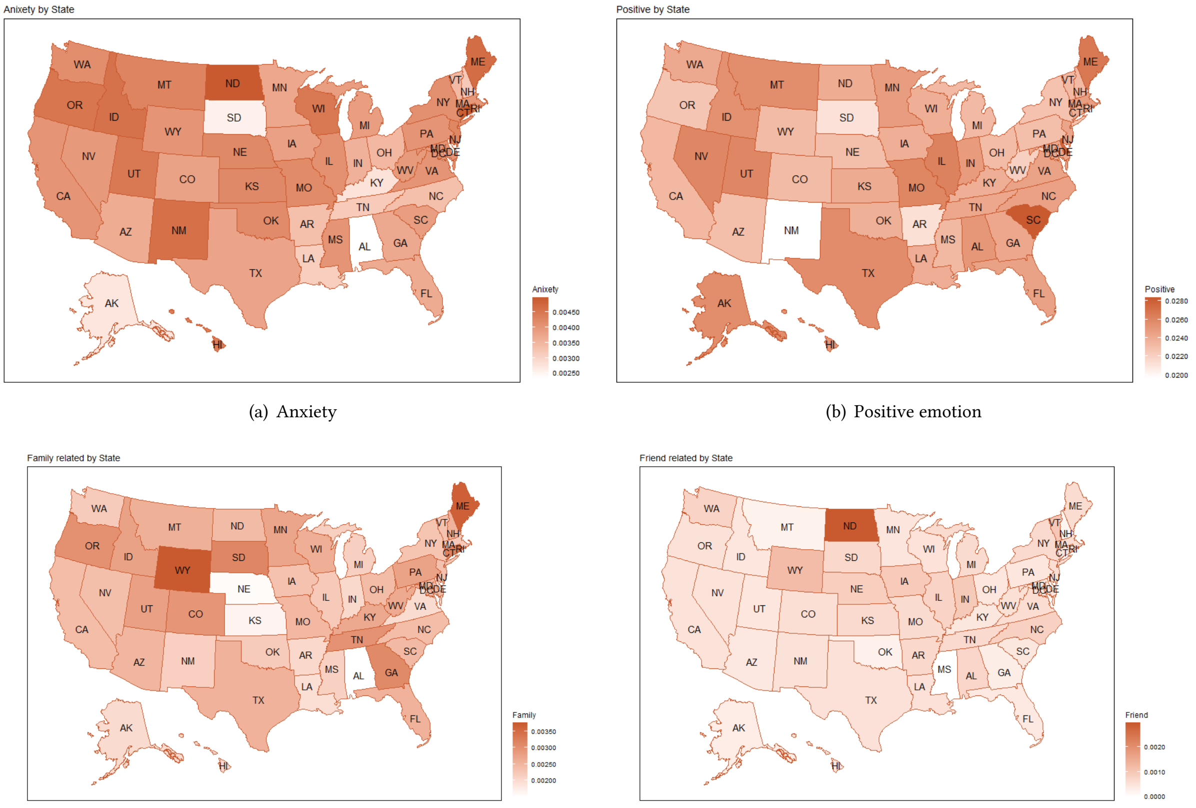
where $|t_k|$ is the number of words in tweet t_k, and s_{t_k} is the number of occurrences of all the words $w_1, w_2, \ldots, w_j$ in tweet t_k.

Figure 7 shows the language features at the state level. From Figure 11(a), we conclude that the levels of anxiety of the people in North Dakota and Maine are the highest in this study. A sample tweet says that `Going #backtoschool can be stressful, especially with the added uncertainty surrounding the #COVID19 pandemic`. However, Maine is also a state with relatively high positive emotions as shown in Figure 11(b). The states with many COVID-19 confirmed cases such as Georgia and North Carolina show less positive emotions. As shown in Figure 11(c), Maine is also one of the states where tweets relate more to family while North Dakota, as shown in Figure 11(d), is the state where tweets relate more to friends. Overall, tweets seem to be more about family than about friends as we compare the intensity of the color shades in Figures 11(c) and 11(d) possibly as the result of stay at home orders. For example, `I stay at home with my family to take online`.

4 RELATED WORK

In recent years, social media platforms have rapidly grown. Twitter is one of the most important platforms and gives away significant spatial and temporal information about the users and their posts. Users express and share details about their life, including health information and opinions on emerging events. For example, Twitter-based approaches can be used to detect late breaking news [16] or local news in spite of data scarcity [23]. The study of public health using Twitter encompass health monitoring and surveillance for early prediction of disease outbreaks. Dredze et al. [4] describe a system, called Carmen, which utilizes geocoding tools for influenza surveillance and considers geo-information from both tweets and users. Lee et al. [9] have collected over 6 million of flu-related tweets, with which they can analyze influenza rates in real time using a novel flu surveillance system.

A system by Paul et al. [13], called Compass, applies neural network methods for sentiment modeling and a dictionary for text classification to capture democratic vs. republican sentiment for the 2016 U.S. presidential election, at the county and state levels. A study of neighborhood happiness, diet, and physical activity has been performed by Nguyen et al. [12]. It applies sentiment analysis to regions, namely neighborhoods, based on geotagged tweets and socio-demographic characteristics, as provided by census data. Martinez et al. examine the use of and perceptions about e-cigarettes



(a) Anxiety

(b) Positive emotion

(c) Family

(d) Friend

Figure 11: Language features.

in the U.S. [10]. They use a sample of tweets that they manually geocoded and interpreted, which has the advantage of creating a detailed categorization of the tweets.

Sadilek et al. [15] and Wang et al. [21] perform studies to detect and predict foodborne illness using respectively Twitter and Yelp. They analyze the tweets using language features and language models. Zhao et al. [24] compared the content of Twitter with a typical traditional news medium, the New York Times, using unsupervised topic modeling. They develop a new Twitter-LDA model that is effective for short tweets. Jaidka et al. [7] want to monitor well-being at large scale. They find that text-based methods that use language dictionaries including labMT and LIWC2015, which we used, can easily work at a large scale. However, they found that supervised data-driven methods are more robust, when compared with a gold standard. Similarly to our paper, Paul and Dredze [14] use an unsupervised model, in their case called Ailment Topic Aspect Model (ATAM), and LDA to learn how users express their illnesses and ailments in tweets.

5 CONCLUSIONS AND FUTURE WORK

While Twitter has been used to analyze and predict other diseases, in the case of COVID-19, everything is new: the disease itself, its extent, how it propagates and affects several sectors, including education, and which emotions it brings out. Our large scale study analyzes 673,601 tweets over two time periods in 2020.

Our study is about several aspects, including psychological categories [19], which capture emotion, social relationships, and individual differences. We analyze individual data, and aggregate that information, as a function of time or of state. We illustrate the impact of COVID-19 on education by means of spatio-temporal patterns, sentiment and content analysis applied to a large geotagged Twitter dataset, and capture the topics of greatest concern such as funding and tuition. We found many similarities such as daily tweet numbers (Figure 1) and common patterns such as sentiment percentage in different states (Figure 7).

All in all, we converted a large dataset of tweets into meaningful geospatial information. There are many interesting facts that can be observed. We also verify the correlation of the results that are obtained from sentiment analysis with the number of confirmed positive cases.

As we mentioned in Section 3.2, there is ambiguity when we encounter the word `positive`, which usually denotes a positive sentiment. However, in the case of `positive result`, that tweet is the opposite of being positive, while a `negative result` is great news, and definitely *positive*. The problem arises because we use bags of words. In future work, we will be looking at capturing better the semantics of such cases. It is to our advantage that tweets are focused on COVID-19 only, thus limiting the number of possible ambiguous word associations. Another possibility is to manually label a random subset of tweets so as to apply machine learning algorithms for sentiment analysis. At this time (October of 2020), it seems that the pandemic will last several more months. If that is the case, then we will have a baseline to compare results for 2021 with those we obtain in this paper.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their insightful comments. This work was partially supported by NSF award III-1618126.

REFERENCES

[1] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* 3 (2003), 993–1022. http://jmlr.org/papers/v3/blei03a.html

[2] C. K. Chung and J. W. Pennebaker. 2018. What Do We Know When We LIWC A Person? Text Analysis As An Assessment Tool for Traits, Personal Concerns and Life Stories. In *The SAGE Handbook of Personality and Individual Differences*, V. Zeigler-Hill and T. K. Shackelford (Eds.) (Eds.). Sage Reference, 341–360. https://doi.org/10.4135/9781526451163.n16

[3] Peter Sheridan Dodds, Kameron Decker Harris, Isabel M. Kloumann, Catherine A. Bliss, and Christopher M. Danforth. 2011. Temporal Patterns of Happiness and Information in a Global Social Network: Hedonometrics and Twitter. *PloS One* 6, 12 (2011), e26752.

[4] Mark Dredze, Michael J. Paul, Shane Bergsma, and Hieu Tran. 2013. Carmen: A Twitter Geolocation System with Applications to Public Health. In *AAAI Workshop on Expanding the Boundaries of Health informatics using AI (HIAI)*, Vol. 23. Citeseer, 45.

[5] Xian Gao. 2017. Networked Co-Production of 311 Services: Investigating the Use of Twitter in Five U.S. Cities. *International Journal of Public Administration* 41 (03 2017), 1–13. https://doi.org/10.1080/01900692.2017.1298126

[6] Shawn Hubler and Anemona Hartocollis. 2020. How Colleges Became the New Covid Hot Spots. (September 11, 2020). Retrieved October 18, 2020 from https://www.nytimes.com/2020/09/11/us/college-campus-outbreak-covid.html (Updated October 2, 2020).

[7] Kokil Jaidka, Salvatore Giorgi, H. Andrew Schwartz, Margaret L. Kern, Lyle H. Ungar, and Johannes C. Eichstaedt. 2020. Estimating Geographic Subjective Well-being from Twitter: A Comparison of Dictionary and Data-driven Language Methods. *Proceedings of the National Academy of Sciences* 117, 19 (2020), 10165–10171.

[8] Rabindra Lamsal. 2020. Coronavirus (COVID-19) Tweets Dataset. (2020). https://doi.org/10.21227/781w-ef42

[9] Kathy Lee, Ankit Agrawal, and Alok Choudhary. 2013. Real-time Digital Flu Surveillance using Twitter data. In *2nd Workshop on Data Mining for Medicine and Healthcare.*

[10] Lourdes S. Martinez, Sharon Hughes, Eric R. Walsh-Buhi, and Ming-Hsiang Tsou. 2018. Okay, We Get It. You Vape: An Analysis of Geocoded Content, Context, and Sentiment Regarding E-cigarettes on Twitter. *Journal of Health Communication* 23, 6 (2018), 550–562.

[11] David Newman, Edwin V. Bonilla, and Wray L. Buntine. 2011. Improving Topic Coherence with Regularized Topic Models. In *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems (NIPS) 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain*, John Shawe-Taylor, Richard S. Zemel, Peter L. Bartlett, Fernando C. N. Pereira, and Kilian Q. Weinberger (Eds.). 496–504. http://papers.nips.cc/paper/4291-improving-topic-coherence-with-regularized-topic-models

[12] Quynh C Nguyen, Suraj Kath, Hsien-Wen Meng, Dapeng Li, Ken R Smith, James A VanDerslice, Ming Wen, and Feifei Li. 2016. Leveraging Geotagged Twitter Data to Examine Neighborhood Happiness, Diet, and Physical Activity. *Applied Geography* 73 (2016), 77–88.

[13] Debjyoti Paul, Feifei Li, Murali Krishna Teja, Xin Yu, and Richie Frost. 2017. Compass: Spatio Temporal Sentiment Analysis of US Election: What Twitter Says!. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, August 13 - 17, 2017.* ACM, 1585–1594. https://doi.org/10.1145/3097983.3098053

[14] Michael J. Paul and Mark Dredze. 2011. You are What You Tweet: Analyzing Twitter for Public Health. In *Fifth International AAAI Conference on Weblogs and Social Media.* Citeseer.

[15] Adam Sadilek, Henry A. Kautz, Lauren DiPrete, Brian Labus, Eric Portman, Jack Teitel, and Vincent Silenzio. 2017. Deploying nEmesis: Preventing Foodborne Illness by Data Mining Social Media. *AI Mag.* 38, 1 (2017), 37–48. https://doi.org/10.1609/aimag.v38i1.2711

[16] Jagan Sankaranarayanan, Hanan Samet, Benjamin E. Teitler, Michael D. Lieberman, and Jon Sperling. 2009. TwitterStand: News in Tweets. In *17th ACM SIGSPATIAL International Symposium on Advances in Geographic Information Systems, ACM-GIS 2009, November 4-6, 2009, Seattle, Washington, USA, Proceedings.* ACM, 42–51. https://doi.org/10.1145/1653771.1653781

[17] "Kovach Computing Services". 2007. LIWC Dictionary (Linguistic Inquiry and Word Count). (2007). https://www.kovcomp.co.uk/wordstat/LIWC.html Accessed September 26, 2020.

[18] Carson Sievert and Anemona Kenny Shirley. 2015. pyLDAvis. (April 5, 2015). Retrieved October 21, 2020 from https://github.com/bmabey/pyLDAvis

[19] Yla R. Tausczik and James W. Pennebaker. 2010. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology* 29, 1 (2010), 24–54. https://doi.org/10.1177/0261927X09351676

[20] New York Times. 2020. Tracking Covid at U.S. Colleges and Universities. (2020). https://www.nytimes.com/interactive/2020/us/covid-college-cases-tracker.html Accessed September 25, 2020.

[21] Zhu Wang, Booma Sowkarthiga Balasubramani, and Isabel F. Cruz. 2017. Predictive Analytics Using Text Classification for Restaurant Inspections. In *ACM SIGSPATIAL Workshop on Smart Cities and Urban Analytics*, Huy T. Vo and Bill Howe (Eds.). ACM, 14:1–14:4. https://doi.org/10.1145/3152178.3152192

[22] Sarah Watson, Shawn Hubler, Danielle Ivory, and Robert Gebeloff. 2020. A New Front in America's Pandemic: College Towns. (September 6, 2020). Retrieved October 18, 2005 from https://www.nytimes.com/2020/09/06/us/colleges-coronavirus-students.html

[23] Hong Wei, Jagan Sankaranarayanan, and Hanan Samet. 2020. Enhancing Local Live Tweet Stream to Detect News. *GeoInformatica* 24, 2 (2020), 411–441. https://doi.org/10.1007/s10707-019-00392-9

[24] Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan, and Xiaoming Li. 2011. Comparing Twitter and Traditional Media Using Topic Models. In *European Conference on Information Retrieval.* Springer, 338–349.