Investigating Human Behaviour During COVID-19 by Examining Twitter Data

Syema Ailia
Computer Science Department
Northeastern Illinois University
Chicago, USA
sailia@neiu.edu

Dr. Kelly P. Gaither
Department of Women's Health
Dell Medical School
University of Texas at Austin
Computational Engineering Dept.
Mississippi State University
Computer Science Dept.
Austin, USA
kelly@tacc.utexas.edu

Dr. Dave Semeraro
Department of Women's Health
Dell Medical School
University of Texas at Austin
Austin, USA
semeraro@tacc.utexas.edu

Dr. Justin A. Drake
Department of Women's Health
Dell Medical School
University of Texas at Austin
Philosophy Dept.
University of Texas at Austin
Biomedical Engineering Dept.
Austin, USA
jdrake@tacc.utexas.edu

***Abstract*— We examine sentiment analysis on Twitterdata. The contributions of this paper are: (1) We introduce sentiment polarity scores on Twitter text. (2) We explore the mapping of sentiment value per Illinois county. (3) Deduce the results through data visualisation.**

I. INTRODUCTION

Twitter is a widely used platform for users to discuss topics important to them. One of the trending topics being discussed is the rapid spread of the coronavirus disease 2019 (COVID-19) in the United States. These discussions contain valuable sentiments which may directly/indirectly correlate to the way users react toward the pandemic. The aim of this study is to determine whether there is a correlation between the sentiment values of COVID-19 related "Tweets" and their locations. The locations in question are the rural and urban counties in the state of Illinois. Sentiment Analysis will be conducted by using Python to parse and extract the Twitter data and utilise its libraries; the Natural Language Processing (NLP) from the Natural Language Toolkit (NLTK) and Pandas for tweets that were posted in the months of March, April, May and Jun 2020.

II. DATA DESCRIPTION

A. *Twitter API*

Twitter is an American microblogging and social networking service on which users post and interact with messages known as "Tweets". Registered users can post, like, and retweet tweets, but unregistered users can only read them. Users access Twitter through its website interface, through Short Message Service (SMS) or its mobile-device application software. Twitter provides users access to their data though their application programming interfaces (API)[1].

B. *JSON*

The Twitter API returns Tweets and their related data encoded in JavaScript Object Notation (JSON)[2]. JSON is based on key-value pairs, with named attributes and associated values. These attributes, and their state are used to describe objects. These objects all encapsulate core attributes that describe the object. Each Tweet has an author, a message, a unique ID, a timestamp of when it was posted, and sometimes geo metadata shared by the user. In this research, we are interested in the "text" key, which contains the message of the user, and their geo metadata, if enabled by the user, named "place" which returns the city and state they had posted the Tweet from. Figure 1. illustrates the structure of the JSON Twitter data and their objects and *some* of their attributes:

Figure 1.


```
{
  "created_at" : "Thu Apr 06 15:24:15 +0000 2017" ,
  "id_str" : "850006245121695744" ,
  "text" : "1\/ Today we\u2019re sharing our vision for the future of the Twi
  "user" : {
    "id" : 2244994945 ,
    "name" : "Twitter Dev" ,
    "screen_name" : "TwitterDev" ,
    "location" : "Internet" ,
    "url" : "https:\/\/dev.twitter.com\/" ,
    "description" : "Your official source for Twitter Platform news, updates
  } ,
  "place" : {
  } ,
  "entities" : {
    "hashtags" : [
    ] ,
    "urls" : [
      {
        "url" : "https:\/\/t.co\/XweGngmxlP" ,
        "unwound" : {
          "url" : "https:\/\/cards.twitter.com\/cards\/18ce53wgo4h\/3xo1c" ,
          "title" : "Building the Future of the Twitter API Platform"
        }
      }
    ] ,
    "user_mentions" : [
    ]
  }
}
```


Funded by NSF Grant #1852538

III. Python Libraries

A. Pandas

Pandas is a software library written for the Python programming language for data manipulation and analysis[3]. In particular, it offers data structures and operations for manipulating numerical tables and time series. Pandas allows us to create 2-dimensional data structures automatically: objects can be explicitly aligned to a set of labels called a Dataframe. For our purposes, we use Pandas to create a table of sentiment values and their corresponding locations.

B. NLTK

The Natural Language Toolkit (NLTK)[4] is a Python package for natural language processing. From this package we download the submodules: util and Vader[5]. From the Vader module, we use their SentimentIntensityAnalyzer package and its polarity_scores function on a given sentence (in this instance, the text from the Twitter data) which returns a sentiment intensity score as a float for sentiment strength. Positive values are positive valence, negative values are negative valence.

C. Plotly

Plotly provides open source graphing libraries for Python[6]. We will be using the Choropleth map function under this package. A Choropleth Map is a map composed of coloured polygons. It is used to represent spatial variations of a quantity.

IV. Pre-processing of Data

The process of extracting only parts of the data for our use requires multiple steps.

A. Location Extraction

To ensure that we only work with Tweets from Illinois, Tweets that met the following conditions were written to a separate text file:

- The Tweet's "place" object is not none.
- The Tweet's "place" objects "full_name" value contains the string ", IL".

B. Text Cleaning

When handling text for sentiment analysis, it is vital to perform a round of text cleaning techniques to take care of nonsensical text, numbers and punctuation. For this process, the following constraints are applied using Regular Expressions (regex)[7] and written to a separate text file:

- Lowercase text, remove text in square brackets, remove punctuation and remove words containing numbers.
- Nonsensical text and additional punctuation such as apostrophes.

C. Removing Stop Words

A stop word is a commonly used word (such as "the", "a", "an", "in"). We do not want to consider these words when applying sentiment analysis on the sentence as they provide little meaning. NLTK in Python has a list of stopwords stored in 16 different languages. The Tweet text is formatted as a string, and therefore must be split into separate words in order to check if each word is a stop word. We do this by using a word_tokenize method from the NLTK package on the sentence, then apply the stopwords method on each word in the sentence. Once the stopwords are removed, we join the tokenised words back together to form a single string.

D. Illinois County FIPS Codes

To map sentiment values per county, we download a CSV file from an external database called United States (US) Cities[8] that contain Illinois county Federal Information Processing Standard (FIPS) codes.

IV. Analysis

The Tweet text is ready to be analysed by the SentimentIntensityAnalyzer. For each Tweet text, we apply the SentimentIntensityAnalyzer, generate a sentiment polarity score and set it as a key in a python dictionary. For each county's sentiment score, we subtract the negative polarity from zero, then add the positive score. We then averaged these calculations for each county.

Finally, we create a Dataframe with a Series of FIPS codes and their corresponding averaged sentiment scores. Figure 2 shows the first 5 lines of the output from this Dataframe.

Figure 2.

	FIPS	Sentiment Average
0	17043	0.011387
1	17031	0.019545
2	17097	0.018622
3	17167	0.029923
4	17033	0.000000

IV. Data Visualisation

To view the mapping of sentiments per county, we created an array of hex colour values which are gradients between the values of 0.0 to 0.1 that are divided into 8 parts. These values are then used to create a Choropleth map using the Plotly package. The following figures are the results for the months of March, April, May and June respectively.

Figure 3. Average sentiment of Tweets per county in March

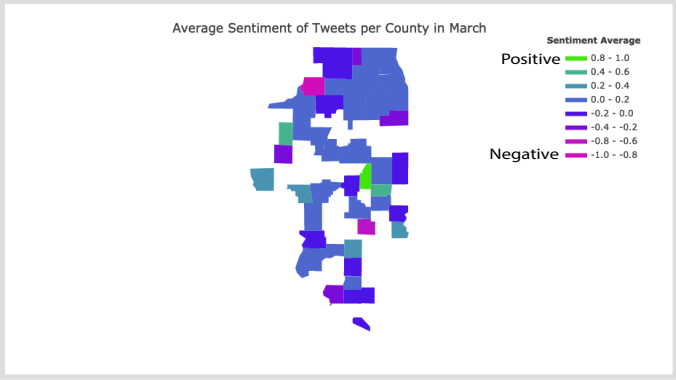


Figure 4. Average sentiment of Tweets per county in April

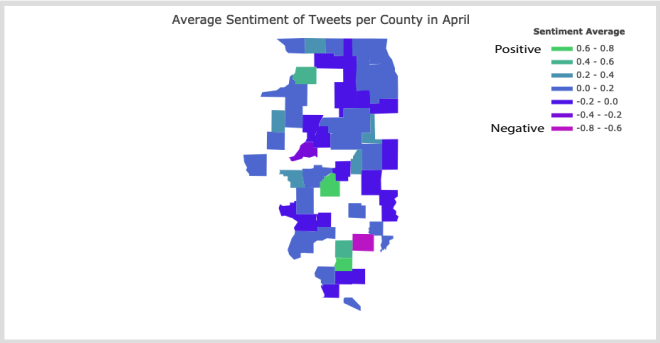


Figure 5. Average sentiment of Tweets per county in May

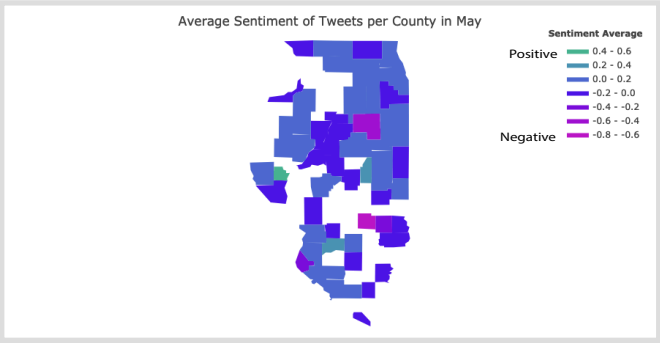
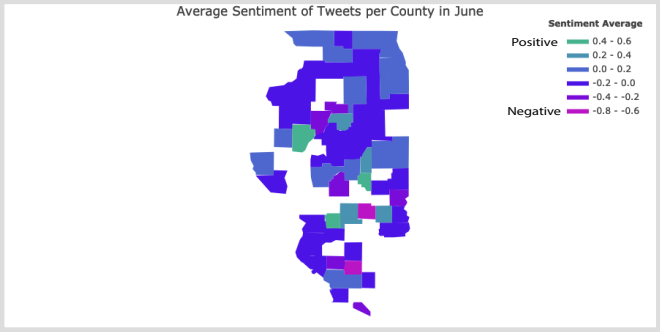


Figure 6. Average sentiment of Tweets per county in June



V. Conclusion

In the spring and early summer, very few people discussed topics regarding COVID-19 in rural and southern counties. As COVID-19 spread throughout the state[9], more users in rural and southern counties began to discuss COVID-19.

Additionally, the majority of users in the northeast of Illinois (including Chicago) had maintained fairly consistent neutral sentiment in their tweets, while the sentiment of rural and southern counties was volatile between the months of March and June.

VI. Future Work

- Integrate machine learning capabilities to increase recognition of concepts such as context, sarcasm, and misapplied words.
- How does the measured sentiment by each region correlate with that region's infection rate via COVID-19 data on census level?
- Aggregate a large set of data by live stream via the Twitter API and showcase an interactive dashboard displaying sentiments over a given time period.

Acknowledgment

I would like to express my very great appreciation to the CyberInfrastructure Research 4 Social Change and to the NSF, Award #1852538, for giving me the opportunity to conduct this research.

Thanks to the Texas Advanced Computing Center staff for their training, computing resources and aid throughout this REU experience.

I would like to offer my special thanks to Dr. Kelly Gaither and her valuable and constructive suggestions during the planning and development of this research.

I wish to thank Dr. Dave Semeraro for their contribution to this project and for their valuable technical support on this project.

I am grateful for the assistance given by Dr. Justin Drake for his useful and constructive recommendations on this project.

Special thanks should be given to Rosalia Gomez, my research project supervisor for her professional guidance and valuable support.

References

1. "Developer." *Twitter*, Twitter, developer.twitter.com/en.
2. "Introducing JSON." *JSON*, www.json.org/json-en.html.
3. *Pandas*, pandas.pydata.org/.
4. "Natural Language Toolkit¶." *Natural Language Toolkit* - NLTK 3.5 Documentation, www.nltk.org/.
5. 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.
6. "Plotly Python Graphing Library." *Plotly*, plotly.com/python/.
7. *Regular Expressions*, pubs.opengroup.org/onlinepubs/9699919799/basedefs/V1_chap09.html.
8. "HTML5 Illinois Map." *Simplemaps*, simplemaps.com/county-il.
9. The New York Times. "Illinois Coronavirus Map and Case Count." *The New York Times*, The New York Times, 1 Apr. 2020, www.nytimes.com/interactive/2020/us/illinois-coronavirus-cases.html.