Examining Collective Consciousness through Machine Learning and Social Media Sentiment Analysis

Arwain Giannini-Karlin
School of Engineering and Technology
University of Washington Tacoma
Tacoma WA 98402 – 3100
Email: arwain11@gmail.com

Abstract- The goal of this research is to examine consciousness as a collective practice and evaluate a correlated psychological effect that major environmental and socio-political events have on populations across the globe.[i] It is my intention to utilize hybrid sentiment analysis techniques, drawing from both knowledge-based and statistical machine learning methods, on social media data to determine; a.) if a correlation exists with respect to the event itself and an exhibited psychological feature. b.) if the exhibited psychological feature is detected significantly in various geographical locations and if there is a correlation between geographical distance from the event and magnitude of the feature. c.) a time series consisting of a comparison of the scale of the event and the magnitude of the exhibited feature also pre-event and post-event duration of the feature if any. d.) can a psychological ripple effect be observed after a major event and what attributes does the effect exhibit over time.

Index Terms- *Machine Learning, Natural Language Processing, Sentiment Analysis, Consciousness, Social Media, Artificial Intelligence, Data Science, Distributed Cognition*

I. Introduction

Consciousness is the state or quality of awareness, or, of being aware of an external object or something within oneself. Thanks to developments in technology over the past few decades, consciousness has become a significant topic of interdisciplinary research in cognitive science, with significant contributions from fields such as psychology, anthropology, neuropsychology and neuroscience. The primary focus is on understanding what it means biologically and psychologically for information to be present in consciousness—that is, on determining the neural and psychological correlates of consciousness.[ii]

Social media data has dramatically refined approaches in computational sociology and continues to provide breakthroughs on how we interact as species. Social research informs politicians and policy makers, educators, planners, legislators, administrators, developers, business magnates, managers, social workers, non-governmental organizations, non-profit organizations, and people interested in resolving social issues in general. There is often a great deal of crossover between social research, market research, and other statistical fields.[iii]

Through machine learning (ML) and natural language processing (NLP) analysis of social media data during periods of major socio-political change and environmental events it is my intention to observe a correlation of psychological features present during these times and analyze their distributions and recurrent effects throughout global populations. Alongside this effort, it is the goal of this project to determine if any data can be observed of significant value with respect to deconstructing consciousness to any degree. The importance of this eloquently stated by Max Planck, "I regard consciousness as fundamental. I regard matter as derivative from consciousness. We cannot get behind consciousness. Everything that we talk about, everything that we regard as existing, postulates consciousness."

II. Motivation

With the popularity of social networks, opinion mining and sentiment analysis become a field of interest for many researchers.[iv] While opinion mining has proved useful/profitable in several realms of e-commerce and marketing, the application of context analysis and text mining to specifically quantify collective consciousness remains to be seen. Potential valuable data from this analysis may include: observable patterns in psychological behavior over time, psychological ripple effect mapping, cross referencing of psychological features and geographical location providing quantitative value of the impact of an event, if a correlative effect is determined this data could be used in reverse to predict location/type of an event occurring.

However, similar research exists examining collective intelligence as it relates to disaster studies through sentiment analysis. One such study, "Collective Intelligence in Disaster: Examination of the Phenomenon in the Aftermath of the 2007 Virginia Tech Shooting", effectively provided evidence for an emerging phenomenon of highly distributed, decentralized problem-solving resulting from a tragic event combined with the utilization of modern forms of communication.[v] That being said, this work does not employ NLP or ML analysis and maintains a geographically isolated focus. Access to global Twitter data will allow this research to observe behavior patterns (through sentiment analysis) over a much larger scale. Similarly, ML techniques utilized over large data sets can reveal correlations that have remained subliminal in other forms of analysis.

III. Problem Statement

The goal of this research is to examine consciousness and evaluate a correlated psychological effect that major environmental and socio-political events have on populations over time. I plan to use Twitter data and open source libraries to obtain global data during specific time periods where major socio-political and environmental events have occurred. From this data, I plan to utilize Python NLP and ML libraries and algorithms to process and analyze the data, using sentiment analysis, and observe the global effects of isolated events and also the ripple effects of global events over time. Potential valuable data from this analysis may include: observable patterns in psychological behavior over time, psychological ripple effect mapping, cross referencing of psychological features and geographical location providing quantitative value of the impact of an event, if a correlative effect is determined this data could be used in reverse to predict location/type of an event occurring. Lastly, it is the goal of this project to determine if any data can be observed of significant value with respect to deconstructing consciousness to any degree. My motivation for this research is to gain hands on experience with NLP and ML libraries, to become familiar with the process of conducting formal academic research and on a personal note it is an opportunity for me to explore consciousness, neuroscience and cognitive science in greater depth.

IV. Technical Approach

- Technologies include: Twitter open source API, Python, Python NLP and ML libraries (scikit-learn, NLTK, Pandas, Pickle)
- Data collection was done through utilization of Twitter's open source API to collect tweets (10,000/day)
- NLP, specifically the NLTK library to parse the tweets and segregate them based on events/psychological features.
- Apply sentiment analysis to obtain positive/negative results (NLTK-Vader Sentiment Intensity Analyzer).
- Results from the sentiment analysis were analyzed via kmeans (scikit-learn).
- Visualization via Python's matplotlib library

V. Educational Statement

It was my intention, to obtain experience cleaning and preparing data, processing textual data (NLP/text analysis), implementing machine learning algorithms and modeling data through this research project. With the assistance and guidance of Professor Sakpal this has become a reality.

Through this project I have learned techniques in data collection that are critical to extracting data in a manner resulting in data that can be used for scientific analysis later. This was one of my struggles with this project, getting the Twitter API to maintain connection through several scheduled queries. I was eventually able to mitigate the APIs restriction by spacing the queries at least 4 hours apart (the API supposedly resets after 15 minutes), however, at times this would result in fewer Tweets being extracted without warning from the API. This resulted in choppy data collection, I was later able to manually extract data that produced results closer to those hypothesized.

Overall, I feel that I learned skills critical to future professional and academic inclinations into the Data Science realm. Through this project I learned how to extract, organize, clean and manipulate large data sets for textual analysis and machine learning processing. Specifically, these tasks included; the use of regular expression search patterns to clean text data in an efficient and thorough manner; preprocessing data for sentiment analysis utilizing NLTK's Vader; data frame manipulations on large sets to handle missing/broken data; date object cataloging and extraction of past time through Python's datetime object library; storage of data/machine learning models through the Pickle library; kmeans, operations, set up and visualization utilizing matplotlib. All of these experiences have expanded upon my knowledge as a computer scientist and developer, and I am grateful to Professor Sakpal for guiding me in this research.

VI. Data Extraction

Three events were chosen for analysis:

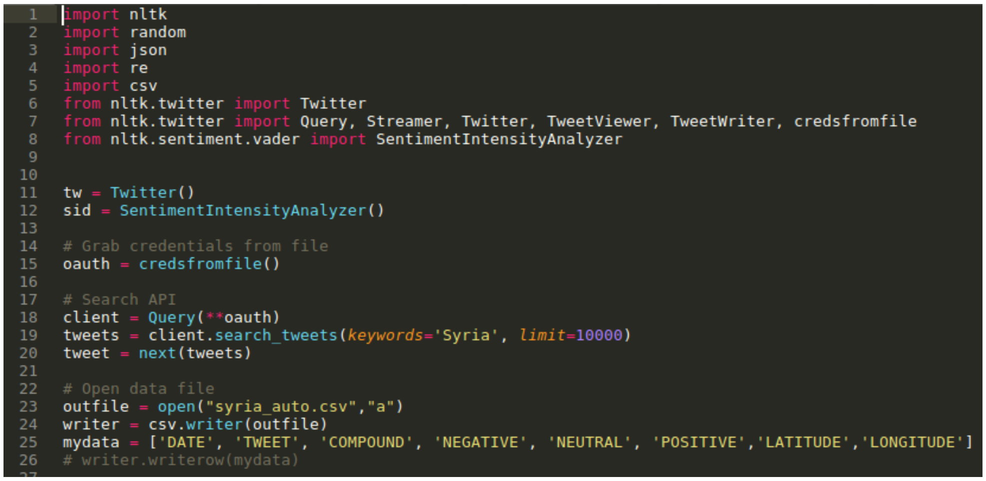
1.) 2018 missile strikes against Syria 14th April 2018 (SY)

2.) 2018 North Korea-United States summit (KS)

3.) 2108 Bitcoin Analysis over time (BTC)

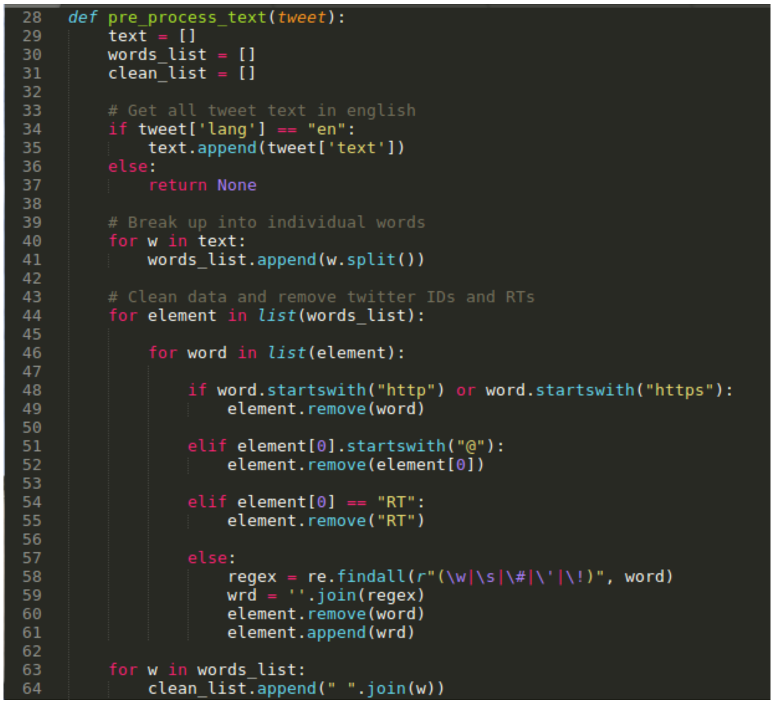
All events were extracted using a similar 5 stage process. The initial stage (Stage 1) connects to the Twitter API and searches events based on keywords. Text data/Tweets are then preprocessed utilizing regular expressions (Stage 2). Next, date and time data are extracted directly from the tweet object (Stage 3). Then sentiment analysis is conducted via Vader Sentiment Intensity Analyzer (Stage 4). Lastly, geodata (if present) is extracted directly from the tweet object (Stage 5).

Stage 1



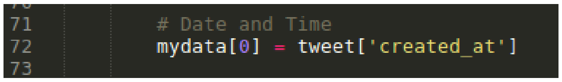
```python
import nltk
import random
import json
import re
import csv
from nltk.twitter import Twitter
from nltk.twitter import Query, Streamer, Twitter, TweetViewer, TweetWriter, credsfromfile
from nltk.sentiment.vader import SentimentIntensityAnalyzer


tw = Twitter()
sid = SentimentIntensityAnalyzer()

# Grab credentials from file
oauth = credsfromfile()

# Search API
client = Query(**oauth)
tweets = client.search_tweets(keywords='Syria', limit=10000)
tweet = next(tweets)

# Open data file
outfile = open("syria_auto.csv","a")
writer = csv.writer(outfile)
mydata = ['DATE', 'TWEET', 'COMPOUND', 'NEGATIVE', 'NEUTRAL', 'POSITIVE','LATITUDE','LONGITUDE']
# writer.writerow(mydata)
```

Stage 2



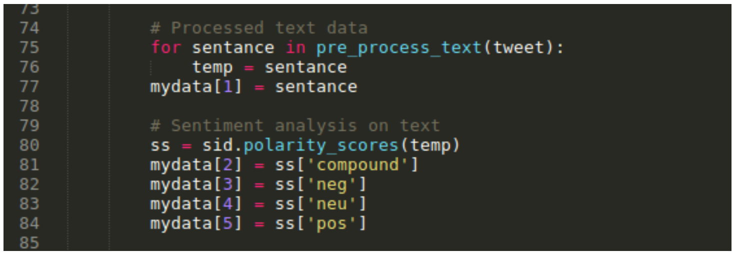
```python
def pre_process_text(tweet):
    text = []
    words_list = []
    clean_list = []

    # Get all tweet text in english
    if tweet['lang'] == "en":
        text.append(tweet['text'])
    else:
        return None

    # Break up into individual words
    for w in text:
        words_list.append(w.split())

    # Clean data and remove twitter IDs and RTs
    for element in list(words_list):

        for word in list(element):

            if word.startswith("http") or word.startswith("https"):
                element.remove(word)

            elif element[0].startswith("@"):
                element.remove(element[0])

            elif element[0] == "RT":
                element.remove("RT")

            else:
                regex = re.findall(r"(\w|\s|\#|\'|\!)", word)
                wrd = ''.join(regex)
                element.remove(word)
                element.append(wrd)

    for w in words_list:
        clean_list.append(" ".join(w))
```

Stage 3



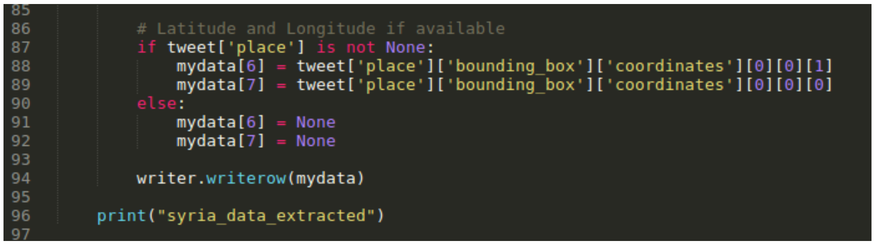
```
        # Date and Time
        mydata[0] = tweet['created_at']
```

Stage 4



```
        # Processed text data
        for sentance in pre_process_text(tweet):
            temp = sentance
        mydata[1] = sentance

        # Sentiment analysis on text
        ss = sid.polarity_scores(temp)
        mydata[2] = ss['compound']
        mydata[3] = ss['neg']
        mydata[4] = ss['neu']
        mydata[5] = ss['pos']
```

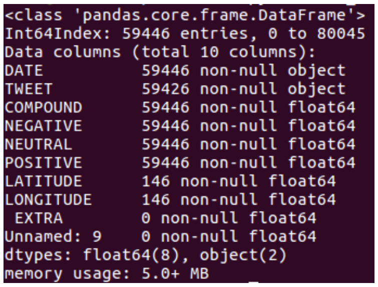
Stage 5



```
        # Latitude and Longitude if available
        if tweet['place'] is not None:
            mydata[6] = tweet['place']['bounding_box']['coordinates'][0][0][1]
            mydata[7] = tweet['place']['bounding_box']['coordinates'][0][0][0]
        else:
            mydata[6] = None
            mydata[7] = None

        writer.writerow(mydata)

    print("syria_data_extracted")
```

VII. Data Analysis and Visualization

Twitter gives its users the option of turning off their geolocation data, while this is a great feature that ensures user privacy, it is a hinderance in obtaining global sentiment. Through the process of collecting this data it was elucidated that 99.7% of users tweeting about bitcoin had blocked geolocation data.



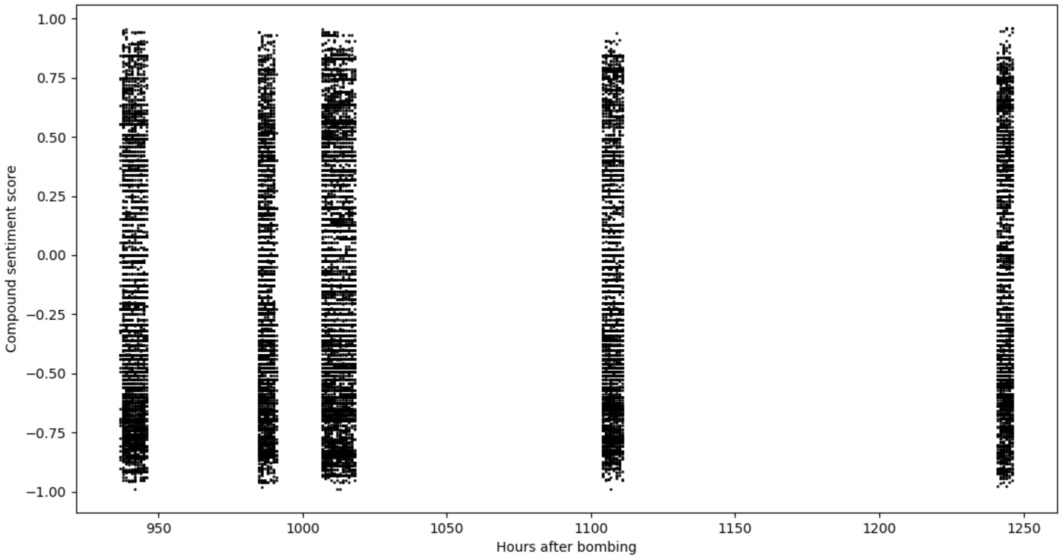
```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 59446 entries, 0 to 80045
Data columns (total 10 columns):
DATE            59446 non-null object
TWEET           59426 non-null object
COMPOUND        59446 non-null float64
NEGATIVE        59446 non-null float64
NEUTRAL         59446 non-null float64
POSITIVE        59446 non-null float64
LATITUDE        146 non-null float64
LONGITUDE       146 non-null float64
 EXTRA          0 non-null float64
Unnamed: 9      0 non-null float64
dtypes: float64(8), object(2)
memory usage: 5.0+ MB
```

This in combination with earlier improper data extraction (earlier techniques only achieved text and time-retrieved data, instead of tweet object) made it essentially impossible to establish an unbiased base to extrapolate geo-specific data with respect to sentiment. It is also worth noting that, future research in this area should maintain method/s of conducting sentiment analysis in multiple languages to achieve an unbiased global sentiment. However, good sentiment data over time was achieved in later runs after initial techniques were refined. Matplotlib was utilized to plot earlier sets directly and also in conjunction with k-means to color code labeled points and mark resulting centroid placement established by the algorithm.

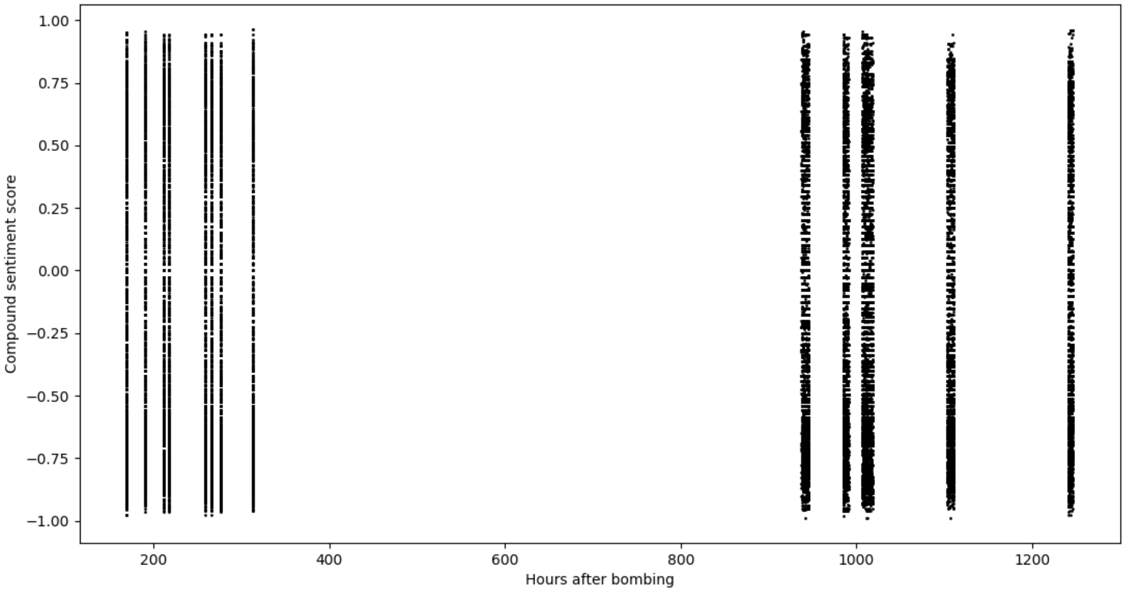
These are as follows:

2018 missile strikes against Syria 14th April 2018 (SY)

Data extracted using later techniques:

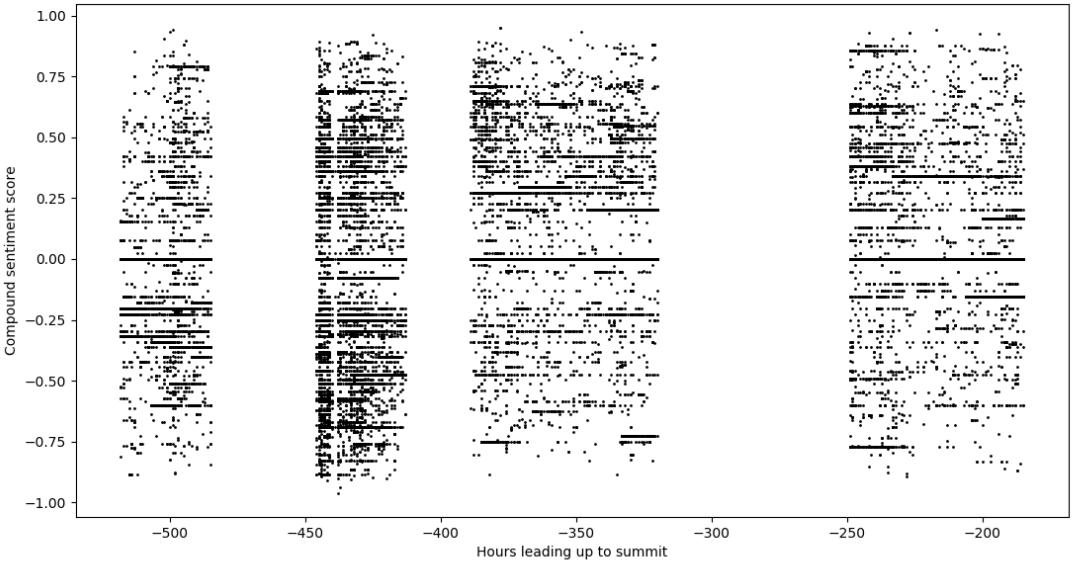


Data combined with earlier extraction:

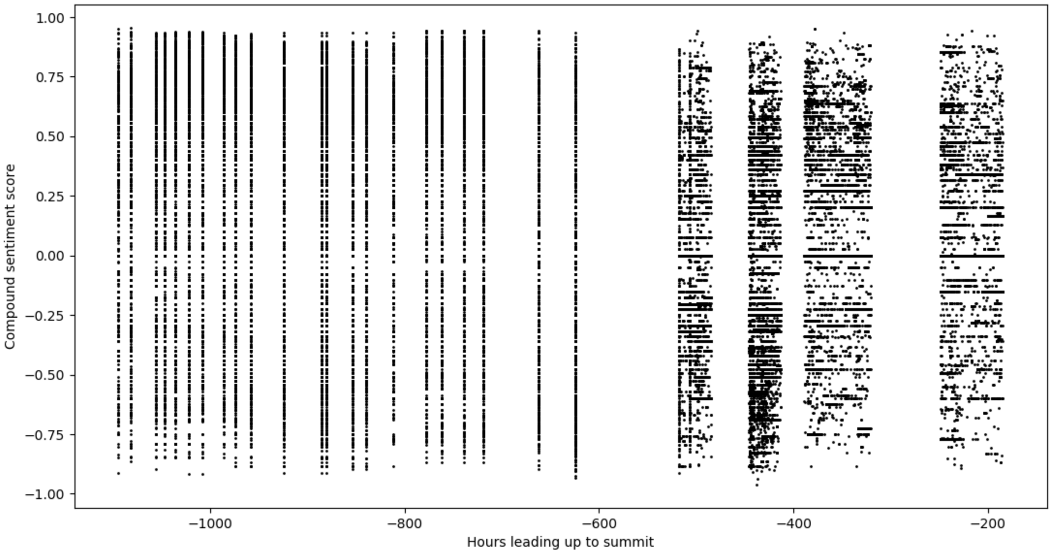


2018 North Korea-United States summit (KS)

Data extracted using later techniques:

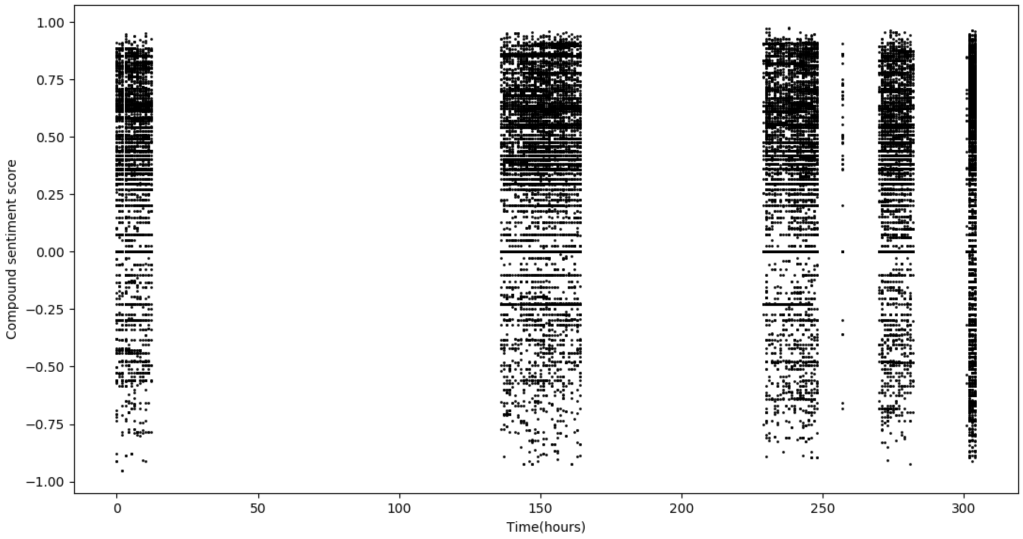


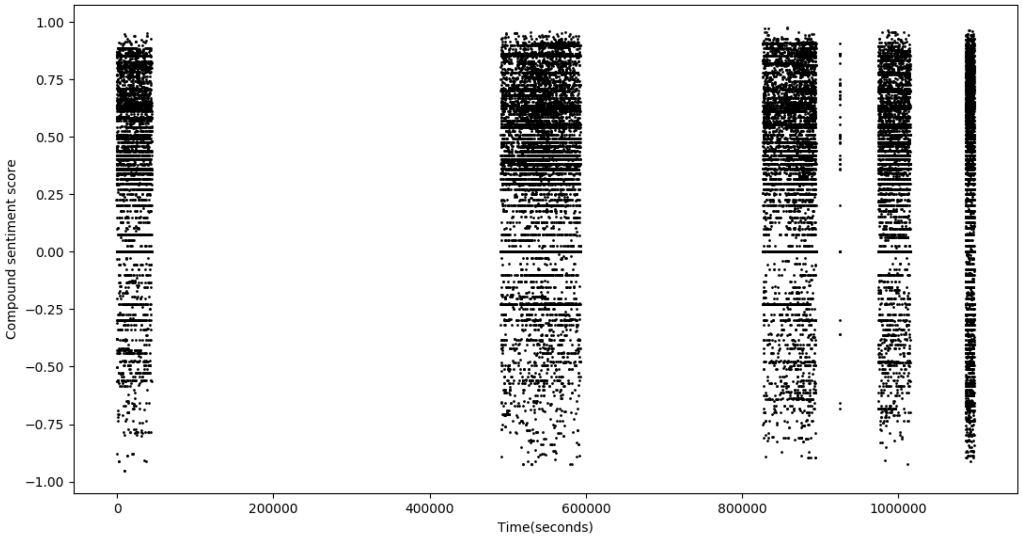
Data combined with earlier extraction:



2108 Bitcoin Analysis over time (BTC)

Data extracted using later techniques:

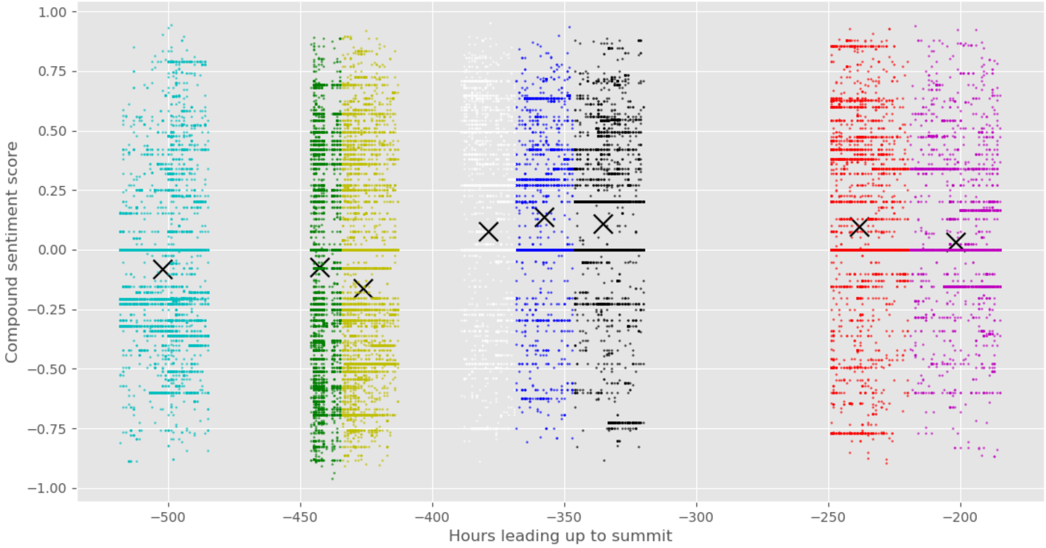




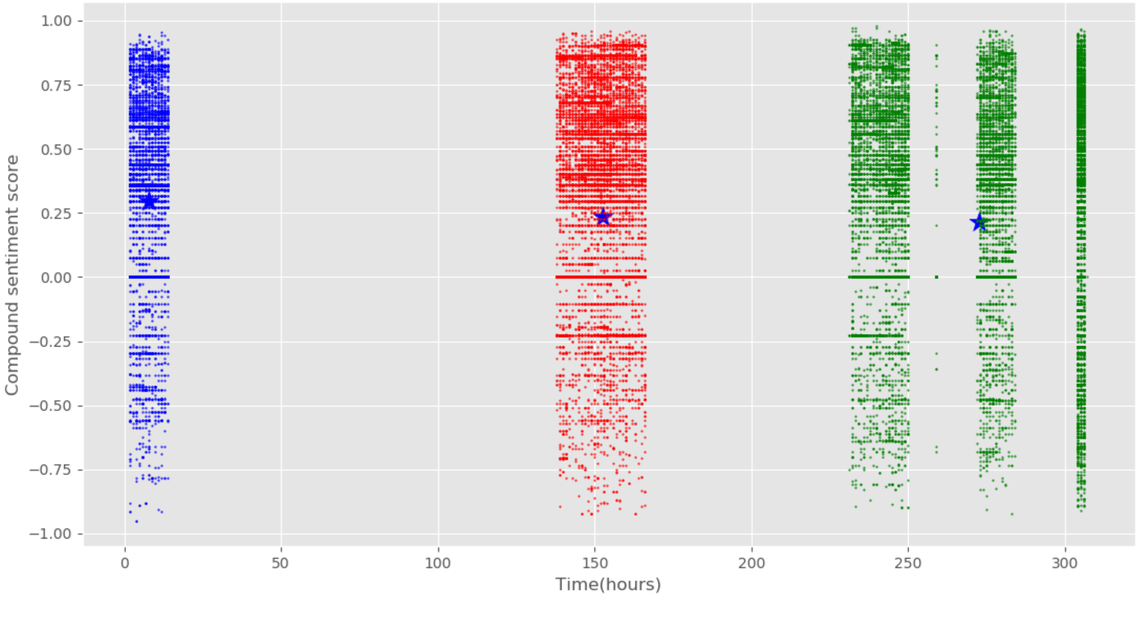
K-Means Analysis

K-Means analysis clustered data with defined x-axis parameters based on the time of extraction/creation, this was not surprising as the API extracts the latest relevant tweets first followed by older tweets and given the time necessary for the free API to reset the data is essentially in waves. However, centroid points on the y-axis are exemplary of the overall sentiment given a time set. So while this is not the traditional methodology, or rather how k-means is generally used we are able to extract a change in sentiment over time. The best example of this was with respect to the KS data set. Here k-means was given a parameter of 8 clusters, the algorithm chose to break these up laterally as predicted, however, the centroid points are varied across the y-axis in correlation with fluctuations in overall sentiment of the time.

KS dataset with k-means (8 clusters):

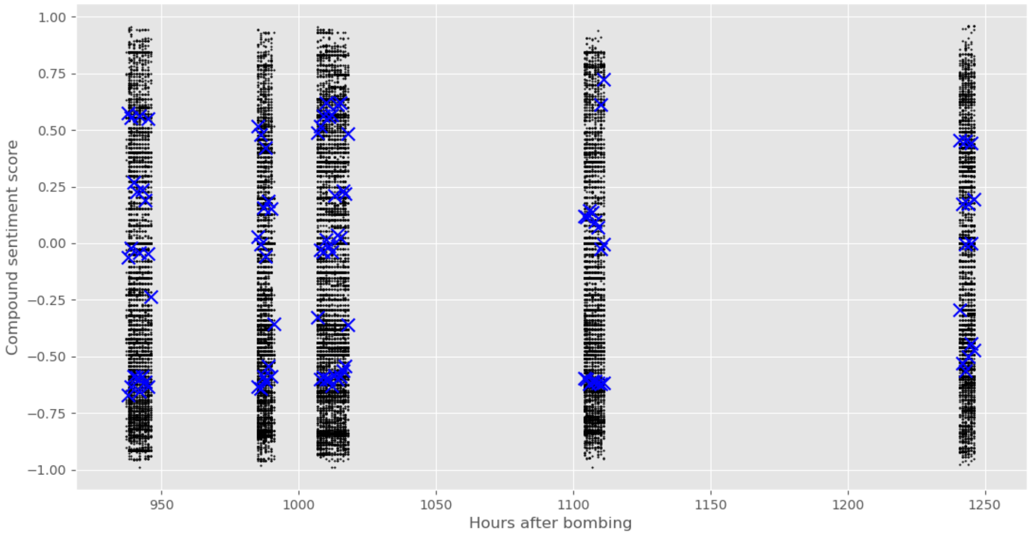


BTC dataset with k-means (3 clusters):

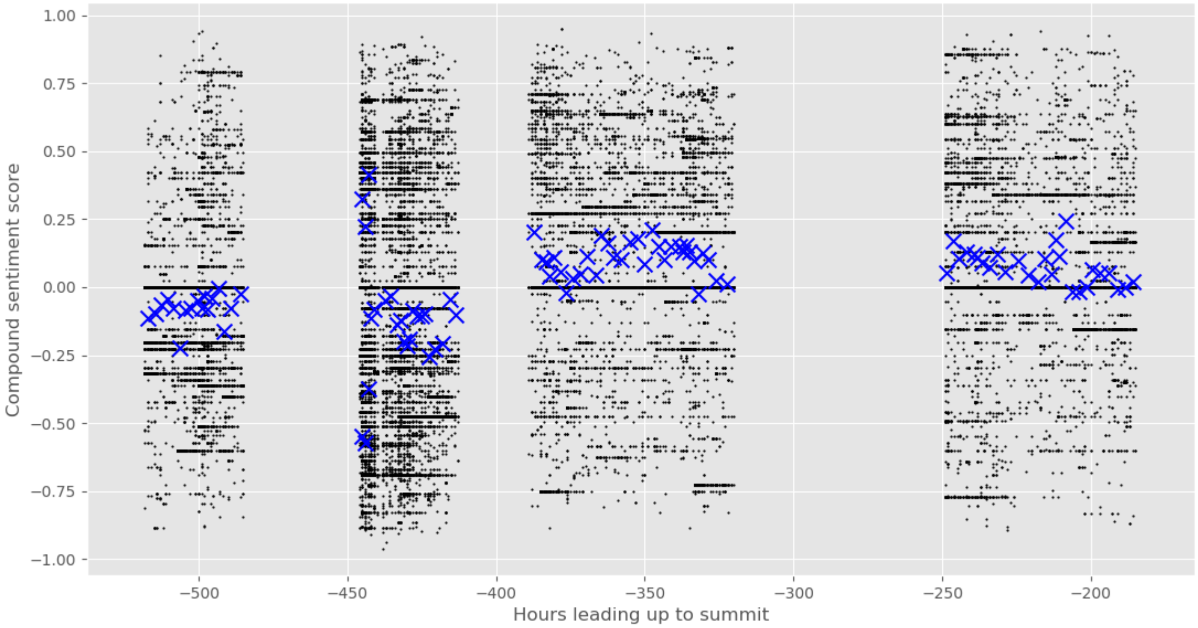


After establishing that clustering would remain laterally defined by extraction time, this projects focus was shifted to the centroid positioning itself for analysis.

SY dataset with k-means (100 clusters):



KS dataset with k-means (100 clusters):



Given more time with this project analyzing these sub-clusters in correlation with political events of the time may prove useful in extrapolating a more refined sentiment. Overall, the KS dataset proved to be most useful in extracting a clear negative to positive shift.

VIII. Conclusion

While global data visualization was the intended conquest of this project and this could not be achieved for lack of available data, sentiment analysis appears to be accurately examined with respect to the events analyzed. Even in the raw SY dataset (without k-means) we can see a heavy concentration of negative sentiment the closer the proximity to the origin (attack). This is again mirrored by k-means analysis establishing a heavy concentration of centroid positioning within this region as well.

Similarly, with the BTC data set, this set was extracted during a time period when bitcoin was enjoying a rapid increase in value after a tumultuous year. Our k-means analysis confirms this with centroid positioning averaging in the +.25 range along the y-axis. Ironically, this event was to be a base to show extreme sentiment shift from negative to positive and vice versa, but during the time period of extraction value was on the rise.

Overall, the KS dataset proved to be most interesting. We see a negative correlation shifting to positive around approximately -400 hours roughly 16 days leading to the summit. This is interesting, as this is the time period that president Trump announced that the summit may be shifted to a later date. Afterwards, the date was reestablished and we see a positive trend for the remaining duration. Also, the KS dataset is not as polar as the other sets, here k-means really made it easier to achieve an overall sentiment analysis where otherwise visually may have proved difficult.

In conclusion, future examination of the sub clusters of any one of the data sets may provide interesting analysis. However, for the purposes of this project Twitter data may not offer the data necessary to examine geo-sentiment. Other sets (Facebook, Google, Yahoo!) may provide more in-depth examination for these purposes.

IX. Acknowledgments

I would like to thank Professor Raghavi Sakpal for guiding me on this journey, this has been a valuable learning experience thank you for giving me the time and resources I needed to complete this project.

I would also like to thank the Twitter API for allowing free access to user data.

.

X. References

[i] - Robert van Gulick (2004). "Consciousness". Stanford Encyclopedia of Philosophy.

[ii] - Trnka R., Lorencova R. (2016). section on consciousness on pp.33-42 in Quantum anthropology: Man, cultures, and groups in a quantum perspective. Prague: Charles University Karolinum Press. ISBN 978-80-246-3470-8.

[iii] - Lynn R. Kahle; Pierre Valette-Florence (2012). Marketplace Lifestyles in an Age of Social Media. New York: M.E. Sharpe, Inc. ISBN 978-0-7656-2561-8.

[iv] - Pak, Alexander, and, Patrick Paroubek. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. Universit´e De Paris-Sud, Laboratoire LIMSI-CNRS, Bˆatiment 508, crowdsourcing-class.org/assignments/downloads/pak-paroubek.pdf.

[v] - S. Vieweg, L. Palen, S. Liu, A. Hughes, J. Sutton (2008). Collective Intelligence in Disaster: An Examination of the Phenomenon in the Aftermath of the 2007 Virginia Tech Shooting. Proceedings of the 5th International ISCRAM Conference, Washington DC, USA, May 2008.