Social Media Analysis in Disaster Management

Sahithi Lakamana, Jennifer Le, Meenakshi Meenakshi, Miao Zheng
George Mason University
Fairfax, Virginia
AIT582-002, *Applications of Metadata in Complex Big Data Problems*
[slakaman, jle29, mmeenaks, mzheng3] @gmu.edu

Abstract

As natural disasters occur, social media is a popular method for real-time updates and awareness/outreach efforts in major weather events. The growing use of social media has made it as an important component for disaster management. Hurricane evacuation is a complex and tedious process; Analyzing the data generated during an event can help government and adjoining agencies effectively take action towards public safety and disaster management. The increased activity of users during natural calamities on social media platforms like Facebook, Twitter along with the geo tagging of the post, provides a method to analyze and monitor activities of users. The data focuses around Hurricane Harvey from 2017, which impacted Houston, Texas and its surrounding areas. In this project, we attempt to classify tweets and extract intelligence using natural language processing of Twitter data in order to establish certain keywords, or commonly used words, which can serve as a measure of awareness in severe weather events.

Problem Statement

Social media (instant) updates has become a popular method to spread news and alerts in real-time (Imran, 2013). With such big data generating at large volumes and increasing velocity, filtering through noise becomes necessary to determine the classification of tweets and how they can be used to increase public safety in future natural disasters or life-threatening events. How can analyzing Twitter users' tweets pertaining to Hurricane Harvey in 2017 benefit future disaster management? A series of data extraction, cleaning/transformation and analysis of tweets will be conducted to determine intelligence, if any, can contribute to disaster management improvements.

Literature

Stowe, et al., 2016
Predicting and modeling the behavior of individuals during a crisis is a challenging task. Most studies that have been completed used face-to face interviews, surveys, data collected post-crisis from residents, etc. But in past recent years, social media has emerged as a powerful source of information (Dennis, 2016). Social media is not only used for expressing opinions but other particular features such as geo-tagged posts, which let the users share the opinion *and* location in real-time. This has been successfully used to identify spatial patterns of users during natural disasters (Stowe, 2016). A number of studies have been performed on the tweets during and after Hurricane Sandy [or other significant US-landfall hurricanes] which led to the insight that most tweets came from the locations where the assistance was needed (Dennis, 2016).

A similar study performed by researchers at the University of Colorado developed and analyzed an approach to identify and specifically classify tweets related to hazardous, disaster-events (Stowe, 2016). In this approach, data mining was performed for Tweets using about a dozen keywords during the time frame between October 2012 and April 2013 for Hurricane Sandy. The collection amounted to a total of over 22 million tweets from 8 million Twitter users. This publication particularly applied strict requirements during the data selection process. After applying selection criteria, just under 100 Twitter users and almost 7,500 tweets to analyze during a 17-day period, beginning one week before the hurricane

made landfall. These tweets however, were required to be predominantly in English.

The primary goal of this University-led research was to capture tweets that relate to a hazard event. In other words, 1 - if filtering through social media noise was possible and 2 - if it can be fine-tuned to classify these relevant tweets. Fine-tuning can be looked upon as more than just a general categorization. For example, if the dataset contained one million tweets, 1- which of these tweets can be identified as hurricane relevant. Of those that were hurricane relevant, 2 - can these be classified into specific labels such as before, during and after actions? For example, is the tweet stating to prepare for a storm→ before? Is the tweet asking for donations → after? Is the tweet asking for rescue → during? Extracting knowledge from the fine-tuned categories can benefit disaster management in specific areas, rather than generalizing. If the specific category can be identified, relief efforts and disaster management can be better applied in future events.

This University-led research combines the variety of linguistics, within tweets and contextual data in reference to other tweets of the same user. For example, especially during the data selection process, a tweet was examined individually and in context of the other tweets the same user posted, to possibly gain metadata. In contrast, other research publications have developed approaches to classify tweets in application to sentiment and general domain. For example, is this tweet: happy/sad, good/bad, fact/opinion, news, etc. What differentiates this research from the University of Colorado from others is that it uses data mining of ***all*** tweets in general and filters from there. This truly applies noise filtering techniques. After the filters and criteria, they attempt to fine-grain the categorization with intention to improve disaster management and response. The importance of this application is that social media is recognized as resource to spread information through a variety of channels (Stowe, 2016) very quickly. In other words, it is applied that social media is used to generate big data in terms of volume, velocity and variety. Overcoming these big data challenges is essential to applying improvements within disaster management to serve public safety matters.

Ultimately, the authors determined it is possible to identify relevant tweets with high precision. However, fine-grain classification is more difficult and needs additional work - specially to define specific categories. Specific categories include but are not limited to: sentiment, action, reporting, information, movement, to name a few. More in detail: action can include rescue, movement can include evacuations, etc.

Data sparsity also proved to be challenging for machine learning, as there was a lack of positive examples to provide as training data. It is possible this is due to meeting the stringent data selection criteria. Nonetheless, the goal of this University of Colorado research was to leverage ***both*** relevance and fine-grain classification. The goal is attainable more to in the 'after' phase of a disaster event with batch processing. However, to improve disaster management, the authors recommended developing an approach to real-time processing, which has yet to be achieved (not verified since 2016). In cases and applications of real-time processing of tweets *during* disaster-related events, it is possible to improve overall management and response in emergency situations.

Lamb, et al. 2013

A research publication referenced by Stowe (University of Colorado) is the study of a flu classification system from Twitter data performed by the Department of Computer Science at Johns Hopkins University in Baltimore, Maryland. This study is similar to the developmental approach by Stowe in that it attempted to fine-grain flu-related tweets. By fine tuning content analysis, in this case the flu, this study attempts to improve the surveillance of influenza using Twitter (Lamb, 2013).

Similarly, this approach uses Twitter generated data to find ways of improving public health and safety. Instead of a hazard event, this case uses the application with influenza, or more commonly referenced as the flu, rather than with disaster/hazard events. Again, what differentiates this research along with the University of Colorado is that they attempt to fine-grain the classification rather than a simple a natural language processing approach of classifying general tweet content. There are a few referenced

research publications that have proven they can generally classify tweets.

This influenza surveillance research attempts to monitor trends, likewise to the Uni. Colorado attempt at before, during and after Hurricane Sandy. Since there are annual reports on flu data, the research with Lamb et al. is able to compare results with traditional surveillance methods and data collected by other resources such as the CDC. These two publications differ significantly in that one is about a seasonal natural disaster, that can strike *specific states during specific seasons*; the other in that it revolves around a seasonal illness that can affect *anyone and anywhere*. In addition, both researches had different goals within the fine-tuning classification but still relate back to public safety.

In our individual research, we attempt to redevelop and distinguish classification approaches with Twitter data in reference to Hurricane Harvey (2017). Since it is a more relevant major disaster event, we may be able to assess the improvements- if any, of disaster management. We attempt to solve the problem statement by investigating how filtering through social media data and determining the classification of tweets can increase public safety, particularly in disaster management. This approach will assess its findings with previous research attempts and how it has differentiated from previous developed methods in improving disaster management using Twitter data.

Background and History

Twitter, being one of the most popular social networks, also functions like a real-time news service (Augustyn, 2009). Many Twitter (tweets) datasets could provide organizations and government agencies with real-time information assistance, and possibly save lives with efforts from certain agencies. Some of the information regarding natural disasters may be issued by local and national news, but studies have shown that nearly 80% of these disaster related tweets are "citizen-reported" tweets of people who were experiencing the disaster themselves (Mims, 2012). There have been various researches performed in terms of identifying disaster related events and generating accurate real-time summaries.

Researchers focusing in this area have concluded that: *"Our data indicate that Twitter activity cannot be defined completely in terms of generative and synthetic information production. Twitter is not simply a platform for broadcasting information, but one of informational interaction. […] navigation of this unwieldy space is difficult. Many of these conventions have evolved to aid this navigation, directing other users to valuable information, placing virtual signposts within a complex information space* (Mims, 2012)*".*

In 2015, a research was completed on summarizing Twitter tweets of a 7.8 magnitude earthquake that happened in Nepal on April 25th, 2015. Based on the research gathered by Gabriel Tseng, between April 25th and May 28th after the disaster happened, there were 33,610 tweets by people in Nepal regarding disaster information (Tseng, 2017). Useful situational tweets were extracted, containing information such as disaster updates. To prepare the data, Tseng manually isolated the characteristics of tweets and used the document analysis tool "td-idf" (Python) to identify significant words (Tseng, 2017). The tool SpaCy was used to tokenize the tweets, however, the downside of SpaCy was that it was difficult to tokenizing the abbreviations and the twitter specific linguistics (Tseng, 2017). Tseng's goal was to generate something useful for rescue teams, using 'pymathproj' to optimize the ILP problem, her results were especially remarkable at providing locational information of the areas impacted by the earthquake.

As Hurricane Sandy hit the East Coast in 2012, millions of tweets were tweeted containing information about the status of the disaster. Researchers from the journal Science Advances gathered 52 million geographically pinpointed tweets and came to the conclusion that the area with most tweets inferred that the area received most assistance from Federal Emergency Management Agency (FEMA) grants (Dennis, 2016). Also, in that journal, Pascal Van Hentenryck, a professor of the University of Michigan also said that the study using Twitter to predict damage during disasters is not ready yet, but may be glimpse in the future (Dennis, 2016). The journals, researches and publications ensure the popularity of using Twitter data and tweets as a

source of gathering intelligence applied in many different natural disaster related studies.

Hurricane Harvey is a more recent natural disaster within the US that has caused significant damage as compared to previous similar events such as Hurricane Katrina and Hurricane Sandy in 2005 and 2012, respectively. In comparison, Harvey caused approximately $125B in damages while Katrina at $161B and Sandy $71B (Amadeo, 2019). All three storms mainly affected three different states. With the increased usage of twitter data for research in disaster management explained earlier, the combination of a more recent, high-profile disaster event and social media data will provide an update to similar publications.

Significance

Raising awareness through social media is an instant approach to educate the general public. The significance of analyzing social media data, particularly tweets, can determine the impact of outreach campaigns, emergency preparedness, donation and relief efforts for impending or passing natural disasters. Profoundly, this can contribute to improvements in behavioral research and decision making within disaster management organizations such as Federal Emergency Management Agency (FEMA). These decisions ultimately affect public safety and funding following a natural disaster; will the public be prepared, safe and able to recover in infrastructure/economically?

In other impacts external to disaster management, this project contributes to and collaborates with natural language processing (NLP) techniques. This technique in machine learning can analyze larger volumes of data with minimal human interaction, in contrast to humans analyzing/reviewing texts manually. Through NLP, texts are reviewed without bias, however, sentiment analysis can be challenging in that it is a machine attempting to extract emotional intent behind the texts. Still, NLP can hold capacity for computers to read text, interpret it (classifications), return sentiment analysis and ultimately, extract intelligence in which decisions can be made and knowledge to be gained.

Other research publications have performed similar approaches of sentiment analysis and the classification and identification of topical (hazard and or disaster-related) tweets. These studies have used various natural disaster events such as tornadoes or hurricanes but all have used Twitter as the chosen social media platform with tweets (Stowe, 2016). In addition, other studies have reviewed the impact of social media and the role of big data- if there are effective solutions for inefficiencies within disaster management. In summary, most attempts at social media analysis in disaster-related events have developed a practical system that can classify tweets, ultimately to extract reliable information to be used in public safety and disaster management.

Motivation

Devastating hurricanes in general make big news headlines to raise awareness before and after the events. Afterwards, it can be common to see donation and charity efforts, especially through social media as well. We can see these efforts by monetary donations, food/supply, clothing, etc. Hurricane Harvey motivates us to explore the NLP techniques because of the growing trend in using social media to raise awareness; emerging technologies allow us to receive instant, real-time updates on what is going on around us. In reference to Hurricane Harvey, it is more of a recent devastating storm similar to Hurricane Katrina. In contrast to Hurricane Katrina, it occurred well over ten years ago and during a time when technologies and social media platforms were still emerging. Compared to ten years ago, many apps and social media platforms have developed and established their presence, along with NLP approaches that have continued to improve. The combination of these developments inspires us to apply those improvements and extract valuable data from this research.

Description of Dataset

Obtained from Kaggle, the final dataset used for this project contains tweets pertaining to Hurricane [and later downgraded to Tropical Storm] Harvey and the related damages in the surrounding areas in Texas in 2017. The 20MB dataset with 6 attributes contained approximately 4 million tweets to be filtered for

noise and required preprocessing steps to be discussed in the next sections. Among tweet texts, a timestamp was also included as an attribute.

A secondary source of data was obtained by the University of North Texas. The 46GB JSON file contained over 7 million tweets and over 130 attributes. This data was extracted from Twitter containing tweets, timestamps, location and many other attributes over a specific time frame. Complications with this dataset will be discussed in the *Challenges* section.

Proposed Approach

As data is explored, Python, R/R-Studio and Tableau will be used in data analytics and visualizations for any trends or preliminary findings. Python will be used for data preprocessing, natural language processing (NLP) (sentiment analysis) of tweets/texts. The models will be utilized with training and testing data, which will ultimately determine classification accuracies along with additional findings.

In order to extract any facts from the dataset during the data preprocessing stages, data mining, cleaning and transformation will be applied first and will consume much of the timeline, as it is a timely process. As those steps complete, visualizations and sentiment analysis can follow. Immediately following visualizations, further analyses may be needed depending on findings. For example, through exploratory data analyses, if a correlation seems to be apparent between variables, it may be worthy to further investigate the correlation with statistical analyses such as regressions, correlation plots, etc.

Questions that we hope to answer or find results:

- What are some keywords that are most frequently used?
- Do hashtags relate back to anything?
    - What are the main topics of concern?
    - Do the hashtags raise awareness (post-event) or preparedness (pre-event)?
- Can we identify any groups?
- Can any predictions be made?
    - Future impacts for disaster management
    - Feedback that can contribute to public safety

Preprocessing

The preprocessing stages comprised of data loading, extraction and cleaning. The dataset can be variety of types including JSON, csv, or others. More on this will be discussed in the challenges section. Data cleaning included removing special characters, tokenization, procedural lemmatization and removal of stop words.

Tokenization

Tokenization divides words into character stream; it splits texts into tokens. These tokens can be a few words, a sentence or even a paragraph, depending on the tweet. Please note: Twitter currently has a character limit of 280 characters. This process was implemented using the *word_tokenize* package from Python within the NLTK package. With good intention, after the previously mentioned preprocessing stages, these tokens have the potential to be grouped together to provide useful semantic for processing and analysis.

Stop Words and Punctuation Removal

The preprocessing steps assist in removing ineffective words (commonly referred to as stop words) such as: the, a, in, an, are, at, for, I, then, etc. Some special characters included in the data file containing tweets were mostly brackets, punctuation marks (which can sometimes be made out as emojis), dashes and spacing. Removal of these words and characters can result in cleaner data for any model. The removal of brackets is especially important because those tweets with brackets can be categorized differently than those tweets without brackets. For example: tweet 1 - ‘the hurricane is making landfall!’ and tweet 2 - ‘[the hurricane is making landfall’ can set inaccuracies within machine learning because of an unexpected symbol, in this case, brackets. The stop words were removed using a predefined set in the Natural Language Toolkit (NLTK) package in Python. Other special characters were removed using the regular expressions and functions in python.

Lemmatization

Lemmatization with natural language processing can be referred to as merging words that are alike such as: like and likes, want, wants and wanted, help, helps, etc. Merging these similar words aid in the categorization in that there will not be different counts when a plural or past-tense word is used. This can be visualized in a following word cloud.

Other preliminary stages of data cleaning involve removing outliers and/or anomalies found within the dataset with exploratory data analysis (EDA). It is not uncommon for these beginning stages to become tedious in order to build up gradually until an analysis and/or algorithm can (see appendix, Fig. 23 a data wisdom pyramid shows the data stage at the bottom level). An example of additional cleaning was splitting certain attributes into multiple attributes. By doing so, these attributes may be useful for plotting or additional analyses later in the analytical process. For example, with a location attribute, a record could be something along the lines of 'Houston, Texas'. By splitting 'Houston' from 'Texas', there are more opportunities for analyses such as isolating states or cities.

Preliminary Findings

Preliminary analyses included attempts to find the keywords that are most significant during crisis of this kind. Two word clouds were generated [through Python] to better aid in visualizing these keywords, which can be much difficult to filter in tabular/plain text format. Initial obstacles were found after generating a word cloud, indicating that our dataset needed additional cleaning (see appendix, Fig. 24). The revised, cleaned word cloud was developed using the dataset containing tweets during the time frame of the hurricane itself, approximately August 25, 2017. The second word cloud was from a separate dataset because the hurricane downgraded to a tropical storm (appendix, Fig. 25) eventually, as it continued moving away from the Houston, Texas area and eastward. The date range for this second word cloud included tweets through August 30th, approximately 5 days after the storm made landfall in Texas. The end product of the additional cleaning displayed the following word cloud: (fig. 1)

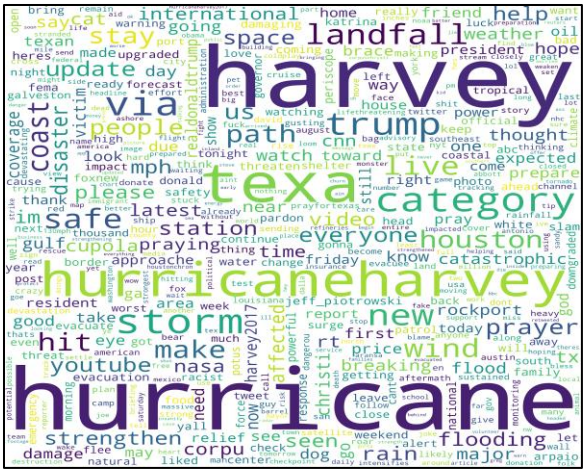

Fig. 1. Word Cloud

It can be inferred that the larger the text/word, the more frequent it appears in the dataset. Clearly, keywords such as 'hurricane', 'harvey', 'texas' and 'safe' were common keywords throughout the tweet dataset. Surprisingly, other keywords such as 'prayers', 'youtube', 'trump' were found in the word cloud, indicating some tweets raising awareness (asking for hopes and prayers or posting attached Youtube videos) or even some tweets aiming at politics. Visualizing these keywords can be used advantageously for future awareness or safety campaigns.

The frequency of keywords that are used can be very helpful in identifying the need of the hour in these disasters. Thus, the overall frequency of the words with the most frequent words used was performed by analyzing hashtags.

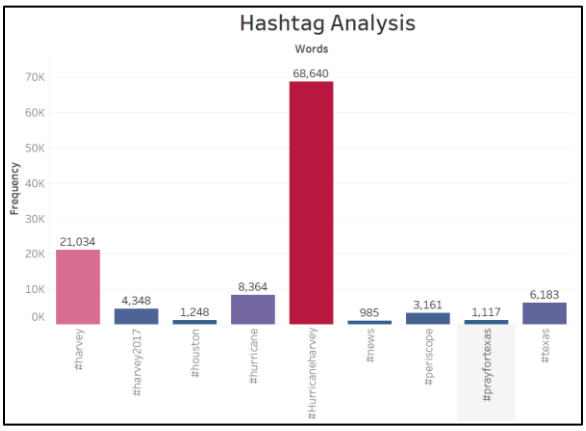

Fig. 2. Frequency of keywords Used

Similar to the word cloud, we can see that still, frequent keywords were along the lines of ‘hurricane’, ‘harvey’, ‘hurricaneharvey’, which can relate back to overall awareness. Note: periscope, third result from the right, is a live video [recording] app, acquired by Twitter in 2015. Based upon the hashtag counts, raising awareness of the hurricane seem to be the focus in this dataset. The top words do not include post-event expected words such as ‘donations’, ‘aftermath’, etc. Ultimately, this could be informative that most of the attention goes to before and during the storm and that many tweets pertaining to the storm after the event decreases significantly.

Exploratory Data Analysis

In addition to preliminary findings, exploratory data analysis was performed to discover any patterns or anomalies and help with any statistical and graphical representations. Essentially in this preliminary analytical stage, we attempt to answer any questions from the proposed approach before delving deeper into advanced approaches.

In the plot below, a time series was generated (with Tableau) to observe trends within the tweets and the timeframe of the approaching hurricane. Here, the number of tweets (count) is plotted with the days of the month in August. Please note, Hurricane Harvey made landfall in Houston, Texas on the evening of August 25, 2017.

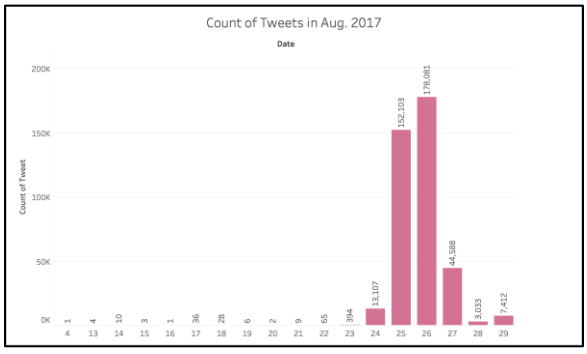

Fig. 3. Number of tweets over time (day)

Unsurprisingly, there were a high number of tweets counted the day of the hurricane. However, the day before the hurricane made landfall, there are significantly fewer tweets—over 90% less [the day before]. It is premature to make any inferences between the discrepancy; citizens could be very busy preparing for the storm or it is also possible that the storm was not taken seriously as a matter of impact that the storm may have.

A high volume of tweet counts surged the day of and even more so the day after. Decreases in tweets appear at least 24 hours post-landfall. Again, it is premature to make any inferences but that can show how much attention is lost from an event once it passes. It is beneficial to note that after the storm, a mandatory curfew was implemented after a series of crime and looting (Amadeo, 2019); this could open a new hashtag analysis but more data is needed.

For a more in-depth approach with the time series, a plot was generated for number of tweets by the hour. This is overall during the month of August as well.

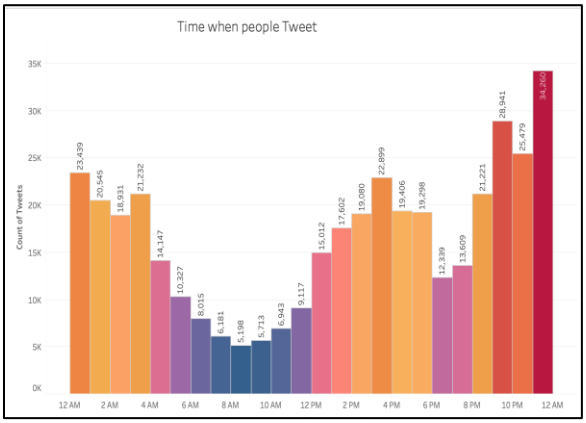

Fig. 4. Number of tweets over time (hourly)

In the plot above, there are very small counts of tweets in the morning hours between 6am and 10am. There seems to be a multimodal distribution, to say the least, with surges of tweets at different time frames of the day. In the later hours after 9pm, the number of tweets are observed to be at its highest. Again, please note that the hurricane made landfall around the Texas Gulf Coast as a Category 4 hurricane at approximately 10pm. It is very possible that there could be a correlation between the hour the hurricane made landfall and the number of tweets surging around that time.

How the locals felt emotionally during the time the storm made landfall maybe something truly immeasurable. Yet, an attempt is made to measure this sentiment though tweets over the time frame of

the storm. In the following Fig. 5, a time series plot of the days leading up to and after the hurricane was measured with the sentiment generated from tweets. In contrast to the previous plots [of number of tweets over time], a varying range of mixed sentiments (feelings) seem to be highest immediately before the storm. The sentiment seems to be more neutralized the day after the storm. It spikes in variation again a couple days after, on the 27th, mostly with negative sentiment. In the days after, the sentiment value lingers off, likewise the number of tweets. Again, in the few days after the storm, a curfew was set in the surrounding areas due to increased crime and looting. It would be interesting to compare these findings with crime statistics for the affected areas at that time of the hurricane. Nonetheless, there was a sense of discontent or dissatisfaction after the storm.

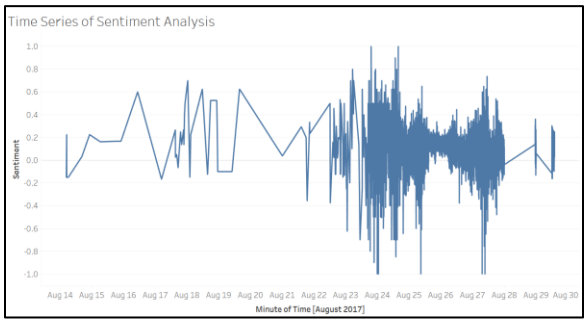

Fig. 5. A measurement of sentiment over time

Further into more exploratory steps, we take another approach at the sentiment over time; this can help in visualizing the number of tweets as the sentiment dies down. If the sentiment diminishes, are there a lesser number of tweets? In other words, if the sentiment fades away over time, does that mean there are less people tweeting about the hurricane?

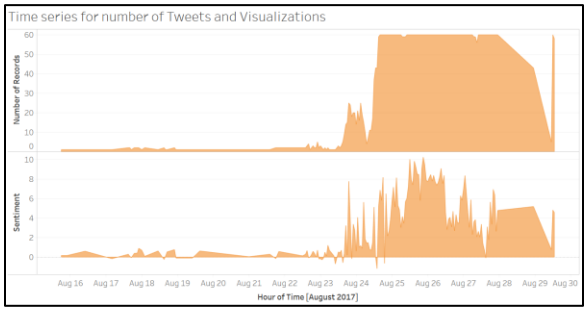

Fig. 6. Sentiment over time with comparison of number of tweets

As observed in figure 6, there are still a high number of tweets during the hurricane and the leading days after. As the days pass, the number of tweets are on the decline until it spikes again between August 29 and 30. Again, does the announcement of a mandatory curfew affect twitter activity? More research would be needed to determine causation.

In the next sections, we take a more advanced and technical approach after gaining insight from EDA. These next steps, descriptive data mining tasks are used for clustering. It is at these advanced analytical stages is where we hope to gain in-depth values to be uncovered by the dataset.

Analytical Approaches

Sentiment Analysis

During a natural disaster, the sentiment of people changes over the course of the disaster. Thus, the overall sentiment of the tweets were analyzed by using the 'TextBlob' package (from Python) and plotted each the overall sentiment with polarity and subjectivity. In addition to polarity, a high positive score would suggest that the hurricane [management] efforts are running on par with a scale where people are satisfied, whereas a high negative score on the other hand, would suggest people were discontent in the way the operations are handled. Both the polarity and subjectivity values cannot exceed 1; polarity can be a -1 whereas subjectivity is always a positive number.

Looking at the analysis, it was observed that most tweets have a neutral sentiment. Alongside, the overall positive sentiments outnumber the negative sentiments.

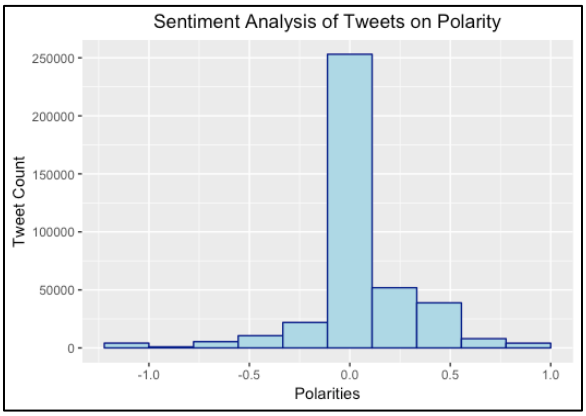

Fig. 7. Sentiment Analysis of Tweets on Polarity

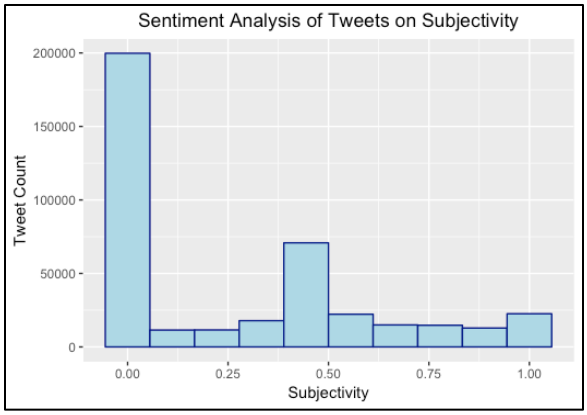

Fig. 8. Sentiment Analysis of Tweets on Subjectivity

Polarity and Subjectivity

Polarity can be described as the emotion expressed in a sentence or statement; it measures the strength of the sentiment. Subjectivity can be described as a varying opinion. It can also be viewed as the opposite of objectivity, or facts. These varying opinions can describe sentiment or feelings towards a topic. Again, polarity and subjectivity values cannot exceed 1 and polarity can reach a value of -1 (implying some kind of discontent) and subjectivity is always a positive value up to 1. Essentially, polarity is the intensity of the emotion expressed and subjectivity is the varying opinion, both of which are towards a statement, topic or sentiment.

In the following Fig. 9, a scatterplot is generated to visualize the correlation between polarity and subjectivity. It can be observed that there are many instances of neutral polarity and subjectivity, meaning the tweets pertaining to the hurricane were without emotions. To better phrase it, there were many tweets that neither expressed negativity or positivity.

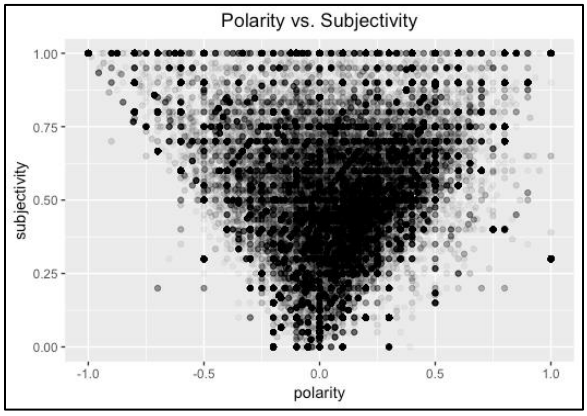

Fig. 9. Scatterplot of Polarity vs. Subjectivity

In an additional attempt to better visualize polarity and subjectivity, an alternative plot was generated (‘kmeans’ package in R) to visualize the clustering. In Fig. 10, a legend was used to differentiate the cluster groups within polarity and subjectivity. Additional analyses with clustering was performed in the next sections.

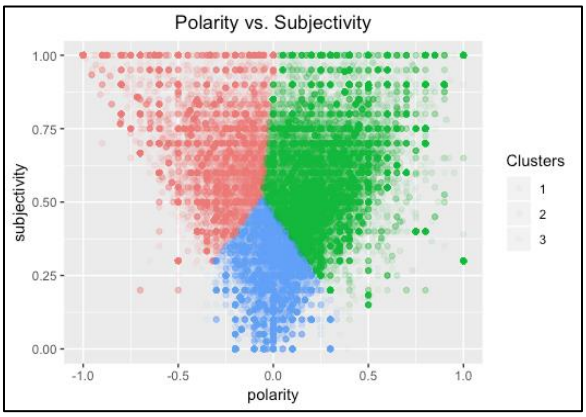

Fig. 10. Improved - Scatterplot of Polarity vs. Subjectivity

Clustering

Clustering the tweets into groups to get brief insight to the underlying patterns is an advanced technique. K-Means clustering was performed to determine the type of tweets that were generated. Here, the words in the tweets are converted to a word vector using ‘TfidfVectorizer’ (in Python); which is the term frequency inverse document frequency (tfidf-). Once every word is converted to a vector, the k-means (with k=3) is applied and the top terms in each cluster are printed to have an idea of what that cluster represent. Below in fig. 11 is an example of one of the cluster contents.

```
Top terms per cluster:
Cluster 0:
 hurricane
 harvey
 texas
 hurricaneharvey
 trump
 storm
 live
 space
 coast
 news
```

Fig. 11. Clustering Results

Another cluster content included many terms relating to some kind of sentiment, such as: thoughts, prayers,

[stay] safe, etc. This group was expected to display a range outside of neutral in polarity. Contents of the remaining clusters of the same analysis can be found in the appendix, figure 19. These cluster groups give us insight on the similar topics people tweet about but with different context.

Text Classification

Also, in this technical analysis, attempts were made to analyze if the tweets displayed situational awareness, as in are actionable tweets found? For example, there is a difference between those who want to help during the hurricane and those who need to be helped during a hurricane. A tweet can express the fact that locals in expected areas of impact need to evacuate. A separate tweet can express that they need help [physically] evacuating, or even rescue if it was too late to evacuate. Possibly due to the limited dataset, it was found that there were not many (almost none) instances of tweets requesting for such help.

Supervised learning with binary classification (Naïve Bayes text classification) was performed to determine if it was possible to differentiate relevant and irrelevant tweets. In this case, tweets were manually labeled for relevancy, whether or not a tweet provided shelter information. During hurricane, aid and shelters are provided to the public for those seeking safety from the storm. If a tweet contained information pertaining shelters, e.g., web links or addresses, it was labeled as relevant; this includes volunteers and aid in general as well. All other tweets would be labeled as irrelevant. The classification model provided 90% accuracy in determining shelter information within a tweet. However, this was in a small sample set of 600 tweets due to the limited dataset.

Topic Modeling

Topic modeling was performed at an attempt to gain insight into what twitter users are really tweeting about, if there is a hidden topic, and if any clusters can be identified. Latent Dirichlet Allocation within topic modeling was used as an algorithm from the Python package ‘Gensim’.

It was discovered that several different groups of terms were found to be overlapping with similar topics. In fig. 12, at least two groups were found to be of contrasting topics. This visualizes the different views (subjectivity) twitter users may have about topics pertaining to the Hurricane, according to their tweets. Within natural language processing, it provides a different outlook of how people feel towards certain topics, similar to sentiment analysis.

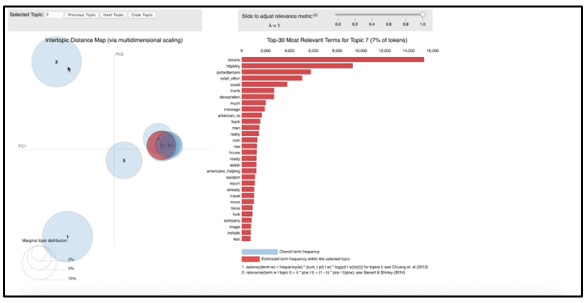

Fig. 12. [partial] results from topic modeling

Results

A few questions that can be answered from the preliminary findings are somewhat expected results. What are some keywords that are most frequently used? What are the main topics of concern? Do the hashtags raise awareness (post-event) or preparedness (pre-event)?

Through the word cloud and clustering techniques, keywords that were frequently used were ‘hurricane’, harvey’, ‘texas’, ‘safe’, ‘category’, etc. These keywords were expected in that they all relate back to the disaster event, Hurricane Harvey in 2017. With safe assumption of these keywords, many were more likely popular hashtags created during the event, with intent of raising awareness and public safety preparing for the storm. Overall, most of the tweets were neutral, possibly just informative tweets to let people know a significant hurricane is approaching and action needs to be taken.

The main topic of concern seems to be raising awareness that everyone within the impacted vicinity is aware of an imminent threat and should take action for their safety. In summary, the frequently used keywords related back to the hurricane itself and regarding the storm event (landfall, category 4, coast, rain, wind, evacuate) in contrast to aid-related keywords such as: donate, relief, food, water, donations, etc. This can provide the conclusion that many of the tweets in this dataset were generated

immediately before and/or during the storm rather than post-hurricane relief efforts.

It was also found that there was a higher amount of Twitter activity in the late evening hours (after 9pm) and less activity in the morning hours (8am). This trend in activity could be due to the fact that the storm made landfall at 10pm; people could tweet to raise awareness that the hurricane made landfall, seek safety/shelter, etc. Twitter activity seemed to decrease as more time elapsed after the storm.

Other questions we were able to gain insight to answer were: Can we identify any groups? Can any predictions be made? Classifying the tweets between certain groups proved to be more challenging with this limited dataset. We were not able to classify specific groups within the tweets outside of binary classification. The task is very attainable but more data is needed; likewise, with predictive tasks. Even more so, demographic data would assist in this task.

Again, rather than the tweets expressing a sentiment towards aftermath relief efforts, many of the tweets itself were on the informative side, such as: landfall, flooding, safety, take shelter, etc. Some tweets seem to express sentiment of what could have been *during* the hurricane. That could be likely due to many of the tweets having been collected around August 25th, the day the hurricane made landfall in Texas.

Collectively, this dataset can still help in improving disaster management, at least in the preparedness stages of an incoming event. Raising awareness is still essential in emergency preparedness because if a storm is not taken seriously, more lives of the general public are in danger. Through this research, we have found that many of the tweets suggest the sentiment and focus were possibly within the hours leading up to the storm. This can be informative to organizations, large and small (government to local), that raising awareness especially through social media, can motivate locals to get actionable towards an impeding, life-threatening events.

Challenges

Typically, the preprocessing stages can be known to be the most time-consuming. Complications began at initial approach of loading the [Univ. North Texas] 46GB dataset in JSON format. It was decided that MongoDB will be used to ingest, truncate unnecessary attributes and convert the data file into a more approachable format such as comma separated value for future use with analytical tools for preprocessing, analytics and visualizations. Applying MongoDB proved to be beneficial with the dataset to convert from key-value pairs to a csv file type.

However, between the preprocessing stages and preliminary findings, it was realized that the dataset was not thoroughly cleaned. For example, in an original word cloud, it was filled with random words, almost none pertaining to a hurricane (Appendix, Fig. 24). Some random words were, for example: emotions, happy, hello, morning, etc. These hurricane-irrelevant words were more likely most used words in tweets in general. Other challenges included stripping random characters from tweets and conversions between different data file types.

Another challenge was also during the preprocessing and preliminary findings stages; it was realized that the original dataset used did not contain enough data for the purpose of our research. The dataset was given by a publication group from the University of North Texas with a size of approximately 46GB and 130 attributes. Out of that 46GB dataset, roughly 3,600 tweets were found to be hurricane related. With less than 10% of the tweets being hurricane-relevant, it was decided that with such a large dataset and very few records applicable to the research, this dataset should not be used solely for the research and instead, a different dataset (Kaggle) will be used for primary analysis. It was unfortunate as the Univ. Texas dataset contained attributes such as time and location which could be used for other analytics but it would have violated data integrity issues by combining two unequal datasets. For example, if the location data (from Univ. N. Texas) dataset were to be concatenated with the tweets (from Kaggle), we would be falsifying the locations of those specific tweets, since they came from two separate datasets.

Another complication with the original dataset was the size for analysis. A 46GB dataset is time consuming to perform a task on a conventional machine. Again, this contributed to the deciding factor that this dataset should not be the primary

source. Ultimately, after time and effort spent in this 46GB dataset, it was decided that the Kaggle dataset was more suitable for the research, even though the dataset size was much smaller in comparison.

Improvements

While reflecting on the analytical findings and applying results to answer questions in the proposed approach, it was realized that perhaps a specific hurricane in general was not necessary. Rather, a combination of hurricanes could be more beneficial. For example, if tweets were mainly the attribute to be analyzed, a combination of tweets pertaining to hurricanes Irma, Sandy, Harvey, etc. could be used. It all relates back to disaster management in general. Any or all hurricanes are appropriate to a specified timeframe i.e., all hurricanes in 2017. This research aims at future applications and improvements for any or all cases disaster management.

Using Hurricane Harvey made it applicable to assess disaster management in August 2017, however using overall hurricanes in general can make it applicable to future events by applying wisdom gained in learning from past events. In another found publication, the authors use just that: a combination of hurricanes to learn from the past (Alam, 2018). That publication aims at analyzing the different types of information made available on social media based on “situational awareness” and “actionable information” using a combination of three different hurricanes. Situational awareness would be applied to bigger organizations such as FEMA or the National Guard, while actionable information would be applicable to first responders, such as rescue teams and emergency medical services.

What differentiates this publication from others, and beneficial to future research is that it also aims at developmental attempts of future automated systems for disaster management. By isolating actions needed from either bigger organizations or first responders, this research moves closer to overcoming a big data velocity challenge in real-time processing. This research also overcomes other big data challenges in volume and variety by analyzing different types of data such as tweets and multimedia content using a variety of artificial intelligence techniques. The research by Alam et al. is a very sophisticated in-depth approach in findings ways of improving disaster management. The research concluded that it is still challenging to make sense of large amounts of disaster/crisis related data on social media, especially real-time computational methods. Likewise, our research is challenged by real-time processing.

As a result of dropping the original dataset, location data was not included for analysis. The addition of location data would be beneficial in mapping visualizations of where the tweets were coming from. Other visualizations from mapping could also include a comparison of where the tweets were originating in comparison to the areas hit hardest during the hurricane. Figure 22 (in appendix) provides an example of mapping potential.

The Kaggle dataset was limited in size, also meaning it lacked desirable attributes. With a larger dataset and more attributes, a multiclass-classification is possible. Adding demographic information about twitter users [who tweeted about the hurricane] would provide more analytics, both in descriptive and predictive. Twitter can provide this information to an extent but with the time constraints and external factors, the limited Kaggle dataset was ultimately the primary source. Essentially, more data for the dataset is desirable.

Conclusion

Social media contains many popular platforms for all age groups to raise awareness. Using twitter to express sentiment, inform or raise awareness can provide real-time updates. Although social media platforms can create big data challenges such as volume, velocity and variety, the data generated by users can provide useful information to disaster management groups from levels ranging from federal to local efforts. In some cases, it can be beneficial to groups needing to take action leading up to the disaster event and to others, it can benefit the aftermath with relief efforts. Using advanced computational methods to analyze user generated data can assist future research and we consider this as a next step in overcoming the big data challenges that face disaster management.

Dataset information:

1. Dan, [last name unknown]. 2017. *Hurricane Harey Tweets: recent tweets on hurricane harvey*, dataset. Retrieved March 2019 from https://www.kaggle.com/dan195/hurricaneharvey.

2. Phillips, Mark Edward. Hurricane Harvey Twitter Dataset, dataset, 2017-08-18/2017-09-22;(digital.library.unt.edu/ark:/67531/metadc993940/: accessed February 23, 2019), University of North Texas Libraries, Digital Library, digital.library.unt.edu;

References (including tutorials)

1. Alam, Firoj, et. al. 15 May 2018. A Twitter Tale of Three Hurricanes: Harvey, Irma, and Maria. Retrieved April 15, 2019. Retrieved from https://arxiv.org/pdf/1805.05144.pdf

2. Amadeo, Kimberly. (20 January 2019). Hurricane Harvey Facts, Damage and Costs. The Balance. Retrieved April 06, 2019 from https://www.thebalance.com/hurricane-harvey-facts-damage-costs-4150087.

3. Augustyn, Adam. (February 2009). Twitter, Microblogging Service. Encyclopedia Britannica. Retrieved February 23, 2019. Retrieved from https://www.britannica.com/topic/Twitter.

4. Dennis, B. (2016, March 11). Why your tweets could really matter during a natural disaster. Retrieved February 23, 2019, from https://www.washingtonpost.com/news/energy-environment/wp/2016/03/11/twitter-might-help-pinpoint-worst-damage-after-a-natural-disaster-study-finds/?noredirect=on&utm_term=.aaac1e2c651e

5. Heisler, Yoni. (February 2018). Twitter's 280 character limit increased engagement without increasing the average tweet length. Retrieved April 13, 2019. Retrieved from https://bgr.com/2018/02/08/twitter-character-limit-280-vs-140-user-engagement/

6. Imran, Muhammad, et al. (2013 May). Practical Extraction of Disaster-Relevant Information from Social Media. Accessed March 09, 2019. Retrieved from https://mimran.me/papers/imran_shady_carlos_fernando_patrick_practical_2013.pdf

7. Irfan, Umair. 2017. "Megadisasters Devastated America in 2017. And They're Only Going to Get Worse." Vox. December 28, 2017. https://www.vox.com/energy-and-environment/2017/12/28/16795490/natural-disasters-2017-hurricanes-wildfires-heat-climate-change-cost-deaths.

8. Jayasekara, Dilan. 2019. "Extracting Twitter Data, Pre-Processing and Sentiment Analysis Using Python 3.0." Towards Data Science. April 3, 2019. https://towardsdatascience.com/extracting-twitter-data-pre-processing-and-sentiment-analysis-using-python-3-0-7192bd8b47cf.

9. Kohorst, Lucas. 2018. "Basic Data Analysis on Twitter with Python." FreeCodeCamp.Org. April 17, 2018. https://medium.freecodecamp.org/basic-data-analysis-on-twitter-with-python-251c2a85062e.

10. Lamb, Alex, et al. (2013 June). Separating Fact from Fear: Tracking flu infections on Twitter. Accessed March 15, 2019. Retrieved from http://www.aclweb.org/anthology/N13-1097

11. Li, S. (2019, January 9). Having Fun with TextBlob. Retrieved May 15, 2019, from Towards Data Science website: https://towardsdatascience.com/having-fun-with-textblob-7e9eed783d3f

12. Li, Susan. 2018a. “Topic Modelling in Python with NLTK and Gensim.” Towards Data Science. March 30, 2018. https://towardsdatascience.com/topic-modelling-in-python-with-nltk-and-gensim-4ef03213cd21.

13. Li, Susan. 2018b. “Named Entity Recognition with NLTK and SpaCy.” Towards Data Science. August 17, 2018. https://towardsdatascience.com/named-entity-recognition-with-nltk-and-spacy-8c4a7d88e7da.

14. Michailidis, Marios. 31 August 2017. *Social ML: Hurricane Harvey*. Accessed on 09 March 0219. Retrieved from https://github.com/h2oai/social_ml
15. Mims, C. (2012, October 22). How Twitter Helps in a Disaster. Retrieved February 23, 2019, from https://www.technologyreview.com/s/419368/how-twitter-helps-in-a-disaster/

16. Nazrul, Syed Sadat. 2018. “Multinomial Naive Bayes Classifier for Text Analysis (Python).” Towards Data Science. April 9, 2018. https://towardsdatascience.com/multinomial-naive-bayes-classifier-for-text-analysis-python-8dd6825ece67.

17. No Author/Contributor. “A Comprehensive Guide to Understand and Implement Text Classification in Python.” 2018. Analytics Vidhya (blog). April 23, 2018. https://www.analyticsvidhya.com/blog/2018/04/a-comprehensive-guide-to-understand-and-implement-text-classification-in-python/.

18. No Author/Contributor. “Document Clustering with Python.” n.d. Accessed May 12, 2019. http://brandonrose.org/clustering.

19. No Author/Contributor. “Generating WordClouds in Python.” 2018. DataCamp Community. August 7, 2018. https://www.datacamp.com/community/tutorials/wordcloud-python.

20. No Author/Contributor. “Gensim Topic Modeling - A Guide to Building Best LDA Models.” n.d. Accessed May 12, 2019. https://www.machinelearningplus.com/nlp/topic-modeling-gensim-python/.

21. No Author/Contributor. “Get and Work With Twitter Data in Python Using Tweepy.” 2018. Earth Data Science - Earth Lab. February 5, 2018. https://www.earthdatascience.org/courses/earth-analytics-python/using-apis-natural-language-processing-twitter/get-and-use-twitter-data-in-python/.

22. No Author/Contributor. “Stack Overflow - Where Developers Learn, Share, & Build Careers.” n.d. Stack Overflow. Accessed May 12, 2019. https://stackoverflow.com/.

23. No Author/Contributor. “Twitter Data Analysis Using Python – Begin Analytics.” n.d. Accessed May 12, 2019. https://beginanalyticsblog.wordpress.com/2018/02/07/twitter-data-analysis-using-python/.

24. No Author/Contributor. “Twitter Sentiment Analysis Using Python.” 2017. GeeksforGeeks (blog). January 24, 2017. https://www.geeksforgeeks.org/twitter-sentiment-analysis-using-python/.

25. No Author/Contributor. Databases and Collections — MongoDB Manual. (n.d.). Retrieved May 15, 2019, from https://github.com/mongodb/docs/blob/v4.0/source/core/databases-and-collections.txt website: https://docs.mongodb.com/manual/core/databases-and-collections

26. No Author/Contributor. MongoDB - Export Data. (n.d.). Retrieved May 15, 2019, from https://www.quackit.com/mongodb/tutorial/mongodb_export_data.cfm

27. No Author/Contributor. Natural Language Processing for Beginners: Using TextBlob. (n.d.). Retrieved May 15, 2019, from https://www.analyticsvidhya.com/blog/2018/02/natural-language-processing-for-beginners-using-textblob/

28. No Author/Contributor. Parsing a large JSON file efficiently and easily. (n.d.). Retrieved May 15, 2019, from NGDATA website: https://www.ngdata.com/parsing-a-large-json-file-efficiently-and-easily/

29. No Author/Contributor. Text Classification: A Comprehensive Guide to Classifying Text with Machine Learning. (13:15). Retrieved May 15, 2019, from MonkeyLearn website: https://monkeylearn.com/text-classification

30. No Author/Contributor. TextBlob: Simplified Text Processing — TextBlob 0.15.2 documentation. (n.d.). Retrieved May 15, 2019, from https://textblob.readthedocs.io/en/dev/

31. No Author/Contributor. tutorialspoint.com. (n.d.-a). MongoDB Create Database. Retrieved May 15, 2019, from www.tutorialspoint.com website: https://www.tutorialspoint.com/mongodb/mongodb_create_database.htm

32. No Author/Contributor. tutorialspoint.com. (n.d.-b). MongoDB Tutorial. Retrieved May 15, 2019, from www.tutorialspoint.com website: https://www.tutorialspoint.com/mongodb/

33. Paruchuri, V. (2016, March 1). Tutorial: Working with Large Data Sets using Pandas and JSON in Python –. Retrieved May 15, 2019, from Dataquest website: https://www.dataquest.io/blog/python-json-tutorial/

34. Saltelli, A., Ratto, M., Tarantola, S., & Campolongo, F. (2005). Sensitivity Analysis for Chemical Models. Chemical Reviews, 105(7), 2811–2828. https://doi.org/10.1021/cr040659d

35. Shaikh, Javed. 2017. “Machine Learning, NLP: Text Classification Using Scikit-Learn, Python and NLTK.” Towards Data Science. July 23, 2017. https://towardsdatascience.com/machine-learning-nlp-text-classification-using-scikit-learn-python-and-nltk-c52b92a7c73a.

36. Stowe, Kevin, et al. (2016 November 01). *Identifying and Categorizing Disaster-Related Tweets*. University of Colorado, Boulder, Colorado. Accessed February 23, 2019. Retrieved from http://www.aclweb.org/anthology/W16-6201

37. Tseng, G. (2017, August 13). Summarizing Tweets in a Disaster – Towards Data Science. Retrieved February 23, 2019, from https://towardsdatascience.com/summarizing-tweets-in-a-disaster-e6b355a41732

38. Vallantin, Wilame Lima. 2018. “Mining Twitter for Sentiment Analysis Using Python .” Medium (blog). October 16, 2018. https://medium.com/@wilamelima/mining-twitter-for-sentiment-analysis-using-python-a74679b85546.

Packages

Python: NLTK, Word_tokenize, TextBlob, Gensim
R: readr, kmeans, ggplot

Appendix: (Additional plots and visual references)

```
#Removing the  punctutations and unuseful characters
data['Tweet']=data['Tweet'].str.replace('[^\w\s]','')
data['Tweet'].head()

0    if decide drive coldplayhouston prepared stay ...
1    as hurricane harvey fast approaching time prep...
2    is jerryjordan_ktt providing live hurricanehar...
3                 im waiting steve harvey hurricane meme
4               the name hurricane harvey  steve harvey
Name: Tweet, dtype: object
```

Fig. 13. Screenshot of preprocessing phase, removing punctuations

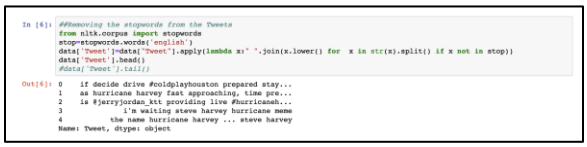

Fig. 14. Screenshot of preprocessing phase, removing stop words

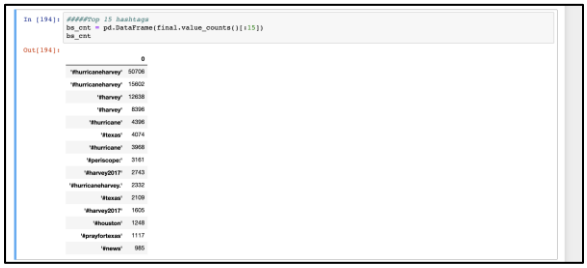

Fig. 15. Hashtag analysis, Python

```
'harvey',                308445
'hurricane',             250999
['hurricane',             92941
'texas',                  90891
'hurricaneharvey',        46529
'via',                    32033
'category',               31472
'storm',                  31197
'4',                      26045
'landfall',               20819
'trump',                  18431
'i',                      16994
'safe',                   16426
'hurricaneharvey']        16342
'harvey']                 16320
'coast',                  15803
'path',                   15277
'news',                   15261
'the',                    14933
'winds',                  14430
'live',                   13858
'houston',                13148
```

Fig. 16. word count, Python

```
plot3 <- ggplot(harvey, aes(x=polarity)) +
  geom_histogram(color="darkblue", fill="lightblue", bins=10) +
  xlab("Polarities") +
  ylab("Tweet Count") +
  ggtitle("Sentiment Analysis of Tweets on Polarity") +
  theme(plot.title = element_text(hjust = 0.5))
plot3
```

```
plot2 <- ggplot(data=harvey, aes(x=polarity, y=subjectivity)) +
  geom_point(alpha=1/20,position=position_jitter(h=0),
             aes(colour= factor(kmeans(scale(cbind(polarity,subjectivity)), centers=3)$cluster))) +
  labs(title="Polarity vs. Subjectivity",
       color="Clusters") +
  theme(plot.title = element_text(hjust = 0.5))
plot2
```

```
plot1 <- ggplot(data=harvey, aes(x=polarity, y=subjectivity)) +
  geom_point(alpha=1/20,position=position_jitter(h=0)) +
  ggtitle("Polarity vs. Subjectivity") +
  theme(plot.title = element_text(hjust = 0.5))
plot1
```

```
plot4 <- ggplot(harvey, aes(x=subjectivity)) +
  geom_histogram(color="darkblue", fill="lightblue", bins=10) +
  xlab("Subjectivity") +
  ylab("Tweet Count") +
  ggtitle("Sentiment Analysis of Tweets on Subjectivity") +
  theme(plot.title = element_text(hjust = 0.5))
plot4
```

Fig. 17. screenshots of R-Studio, code for plots

Fig. 18. screenshot of Python, code for initial steps in cluster analysis

```
Cluster 1:
 houston
 harvey
 hurricane
 flooding
 coldplay
 hurricaneharvey
 texas
 deluge
 handle
 postpone

Cluster 2:
 category
 hurricane
 harvey
 landfall
 storm
 texas
 winds
 strengthens
 makes
 mph

Cluster 3:
 safe
 stay
 prayers
 hurricaneharvey
 path
 thoughts
 praying
 texas
 affected
 hurricane
```

Fig. 19. cluster groups

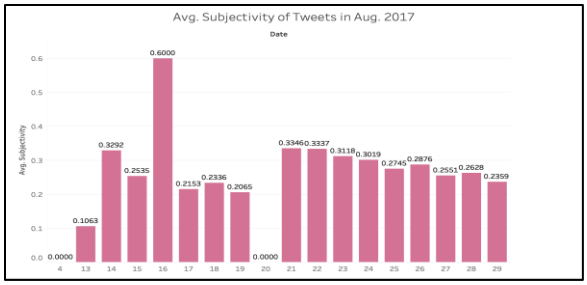

Fig. 20. Plot average subjectivity over time (days in August)

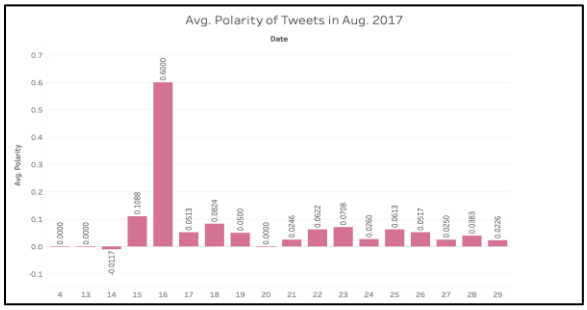

Fig. 21. Plot average polarity over time (days in August)

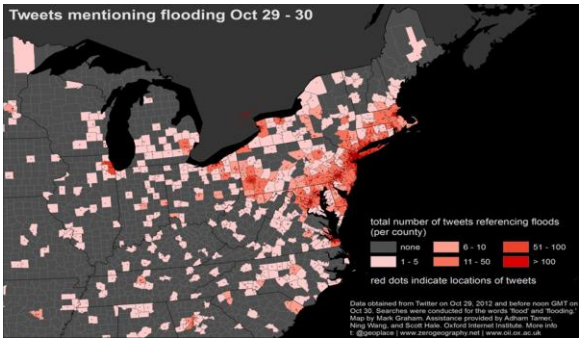

Fig. 22. Mapping potential with location data

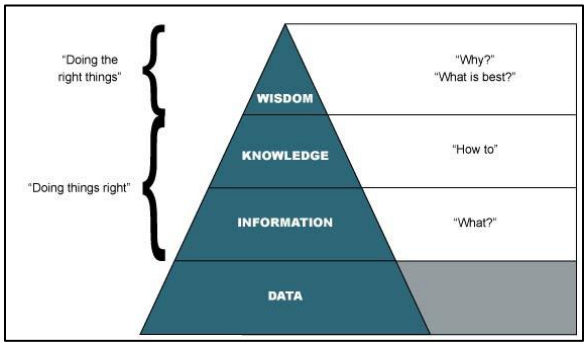

Fig. 23. Data wisdom pyramid

Fig. 24. Initial word cloud, indicating the data is irrelevant to hurricanes

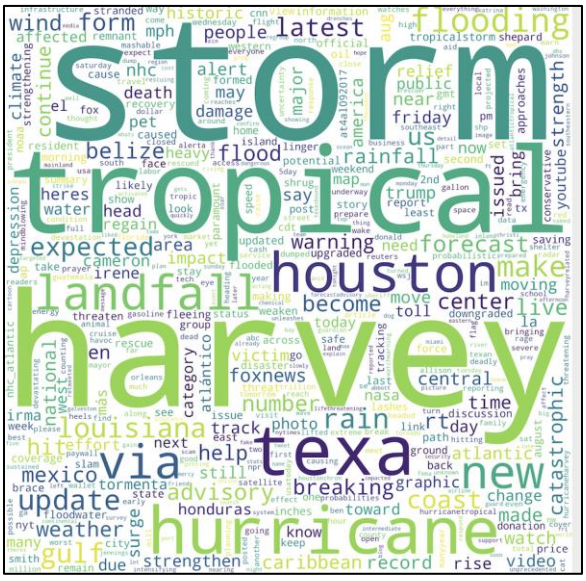

Fig. 25. Word cloud from [downgraded to] Tropical Storm Harvey dataset.