Project Harvey: A Classification of Twitter Data Surrounding Harvey Weinstein and Hurricane Harvey Tweets

Kristin Robinson*
University of Colorado Boulder

Joel Marquez*
University of Colorado Boulder

Franklin Harvey*
University of Colorado Boulder

Jocelyne Agboglo*
University of Colorado Boulder

ABSTRACT

Project Harvey is about taking a look at the differences between certain tweets on Twitter during the occurrences of two main events in 2017. The two main events are the devastation of Hurricane Harvey, and the Harvey Weinstein sexual harassment allegations. We set out to visualize the key differences between the two events. Ultimately with our data, we were looking to see if the conversations about the Hurricane crossed over with the conversations about the Weinstein scandal. Our goal was to create a classifier that discerns between a hurricane Harvey tweet and a Harvey Weinstein tweet, and visualize these differences. That classifier was used to create two visualizations on both Harvey Weinstein tweets and Hurricane Harvey tweets. The first visualization is a text based analysis that distinguishes the features between a Harvey Weinstein tweet and a Hurricane Harvey tweet. The second visualization is a geolocation map, plotting locations of tweets sent out about both Harvey Weinstein and Hurricane Harvey.

1 INTRODUCTION

Twitter is popular microblogging site that allows for users to engage in real-time conversations. These conversations can pertain to the users, or they can follow current events going on all around the world. One important aspect of Twitter is the use of hashtags to follow events, celebrities, tv shows, etc. Along with the use of hashtags there are also key words that follow certain popular topics. To identify certain tweets on the platform, it can be easy to navigate through topics using hashtags and keywords.

In working with Twitter data it can be hard to collect in terms of looking at the Twitter API and compiling the tweets that you might need. We got access to millions of tweets already collected and set out to visualize what we needed. The tweets collected pertained to Hurricane Harvey when the event occurred, and months after. The problem that occurred with the collection was that the Harvey Weinstein scandal broke. With the two events containing similar names, this meant that there were now tweets about both events.

Hurricane Harvey originally made landfall on August 25th, 2017 along Middle Texas Coast. The hurricane started as a tropical storm earlier in the month, evolving into a major Hurricane in about 40 hours. With an enormous event like this occurring, it gained traction and a lot of news and media coverage. This disaster garnered a lot of conversation for months.

* krro4986@colorado.edu
* joma6391@colorado.edu
* Franklin.Harvey@colorado.edu
* joag6491@colorado.edu

On October 5th, 2017 is when the New York Times posted an article on Harvey Weinstein, which included decades of allegations of sexual assault by numerous women. This event also gained a lot of news and media coverage, with a lot of discussions on Twitter. With these conversations taking place including the name “Harvey”, we were looking for any cross-overs between the two events on Twitter. With 41 days in between these events, there is a good chance of having discussions about both the hurricane and the scandal occurring at the same time.

In this paper, we set out to create a classifier that separated Harvey Weinstein tweets from Hurricane Harvey tweets. We then visualized these tweets to show the key differences between the two events.

2 RESEARCH

In taking on the task of creating a classifier and creating these visualizations we looked towards research surrounding Twitter data.

2.1 Related Works

One aspect of Twitter data when it comes to visualizing tweets is location. Because Twitter gives users the option of geotagging their tweets, location data can be pulled from the Twitter Streaming API and mapped out visually. In collecting a large sample of Twitter data though, not all of the tweets are geotagged. In looking at the API, it was found that only a small fraction of about 1% of all tweets contain location information and are geotagged [1]. Although this is a small percentage it is something that you can work with. Because these tweets have locations that lends itself to a geolocation map visualization. In terms of the Twitter API, it takes three key parameters. It takes keywords which consist of words, phrases or hashtags. It also takes geographical boundary boxes and user ID if it is needed [2]. The data collected that was used for our visualizations were also taken from the streaming API.

As you can run into the problem of not having enough geo-tagged tweets there are other options to try and get locations. One example found was a framework for estimating a user’s location, specifically their city, based on the content of the tweets. So this does not include or factor in any geodata. The thought process here to combat this problem of having a lack of geo-tagged tweets was that a user’s tweet could contain some location-specific content. So it would include specific words or phrases that are more likely associated with certain locations than others [3]. This framework could be very helpful if you are working with a Twitter dataset and need more geographical information. In working with big data though, there might be enough geo data to look at and plot on a map.

One important aspect of Twitter mentioned before is its real-time nature. Especially in times of crisis, devastation, or important social events Twitter is a main hotspot for these conversations. To focus on this real-time aspect there was research done on tracking and detecting events in real-time on the platform. The research done proposed an event notification system that monitors tweets and then also delivers notifications to the user [4]. To do this the content of a tweet is looked at, and certain keywords pulled out. But the context of the keywords will be looked at also. So that the keywords are about the correct subject. So as an example you can be tracking an earthquake and pull out the keyword 'earthquake'. Training data will be prepared and applied and a classifier made based on certain features. These features of the tweet could include specific keywords, the number of words, and the context of the word. This can then be applied to a set of tweets about a certain event.

Continuing to look at the real-time aspect of Twitter that also includes behavior, interactions, and conversations during a crisis. So that includes earthquakes, hurricanes, and other disaster situations. Because of how current the platform is, it is accurate in measuring the impacts of certain events that happen. There was research done in exploring the attitudes of users during disasters that occurred. In looking at trends about positive and negative attitudes there was a proposal of a visual framework to visualize attitudes of geo-tagged twitter data. There is an entropy based metric created to model sentiment on social media data. And also, the sentiment that was taken out was placed in a visualization framework so that the uncertainty of public opinion could be explored [5].

In times of crisis and disasters, with so many conversations going on it is a lot of information for people to consume. In searching through so many tweets and data, users have to be careful in looking at what they are consuming to see the accuracy of it. In looking at a disaster that occurred there is disseminated confirmed and unconfirmed news. There was an analysis done on how information spread throughout Twitter. The analysis showed that the spread of tweets that corresponded to rumors was different from tweets that spread news. This is because rumors were questioned more than news by users on Twitter [6].

Along with looking at the type of information that is sent out during a disaster, we can zone in on the differences of the users sending out the tweets. Situational awareness is an individual and cognitive state of understanding "the big picture" during critical events and situations [7]. With Twitter as a platform and situational analysis, the research on this is trying to identify and measure features that could support certain technology in analyzing mass emergency situations [8]. Looking at this information can provide the differences when looking at users who are "on ground" during disasters verses those who are far away from the event.

3 PROJECT HARVEY

Our project consists of two visualizations on the Twitter data collected, using the classified tweets for both Harvey events.

3.1 Data Collection and Process

The data used for our visualizations were taken from the Twitter Streaming API. It was collected at the University of Colorado Boulder by NSF Grant funded Project EPIC. There were approximately 22 million tweets collected, split up into different files. Each of those files consisted of about 1 million tweets, with no particular order to them. We chose to work with one data file, and process that to train our data. The data file we worked with contained about 1.24 million tweets. Our data set consisted of tweets about Hurricane Harvey between the last week of August and April 10th, 2018. The original keywords for all of the tweets contained about 50 terms pertaining to Hurricane Harvey (see Figure 1).

To process our data, we had to create a classifier that differentiated the Harvey Hurricane tweets from the Harvey Weinstein tweets. We used the Naive Bayes equation in our classifier to train our data (see Figure 2). Our first step in processing was gathering some tweets that we knew were referring to each of the classifications. We did this through filtering for keywords like hurricane or Weinstein. The total amount of training tweets consisted of about 66,156 tweets. The amount of tweets on Hurricane Harvey included in the total was about 51,374 tweets. The amount of tweets on Harvey Weinstein included in the total was about 14,782. Then using those tweets, we trained our Naive Bayes classifier. Once our model was trained and text features were weighted, we used the classifier to predict the remaining data which consisted of 1,171,276 tweets. The total number of tweets predicted as Hurricane Harvey tweets were 971,421. The total number of tweets predicted as Harvey Weinstein tweets were 199,855. We were then able to retrieve the top features for each class from the model and select tweets that contained those features to display in our visualization.

#hurricaneharvey
#bridgecity
#corpuschristi
#flood
#flooding
#harveyrelief
#harveyrescue
#harveysafe
#harveysos
#harveystorm
#harveystorm2017
#houston In
#houstonflood
#houwx
#hurricanestrong
#khou11
#laflood
#landfall
#louisianaflood
#louisianafloods #noonecares
#nederland
#nexstarharvey
#orange
#orangefield
#portarthur
#portneches

	#sabine
	#southtexas
	#stormharvey
	#stxwx
	#tornado
	#tropicalstormharvey
	#txwx
	#vidor
	harvery
	harvey
	harvey2017
	harveyflood
	harvy
	houstonstrong
	hurricaine
	hurricaineharvey
	hurricane harvey
	hurricaneharvy
	port arthur
	san antonio harvey
	sosharvey
	storm surge
	stormsurge
	texasstrong

Figure 1: Hurricane harvey keywords

Naive Bayes methods are a set of supervised learning algorithms based on applying Bayes' theorem with the "naive" assumption of independence between every pair of features. Given a class variable y and a dependent feature vector x_1 through x_n, Bayes' theorem states the following relationship:

$$P(y \mid x_1, \ldots, x_n) = \frac{P(y)P(x_1, \ldots x_n \mid y)}{P(x_1, \ldots, x_n)}$$

Using the naive independence assumption that

$$P(x_i|y, x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) = P(x_i|y),$$

for all i, this relationship is simplified to

$$P(y \mid x_1, \ldots, x_n) = \frac{P(y) \prod_{i=1}^{n} P(x_i \mid y)}{P(x_1, \ldots, x_n)}$$

Since $P(x_1, \ldots, x_n)$ is constant given the input, we can use the following classification rule:

$$P(y \mid x_1, \ldots, x_n) \propto P(y) \prod_{i=1}^{n} P(x_i \mid y)$$

$$\Downarrow$$

$$\hat{y} = \arg\max_y P(y) \prod_{i=1}^{n} P(x_i \mid y),$$

and we can use Maximum A Posteriori (MAP) estimation to estimate $P(y)$ and $P(x_i \mid y)$; the former is then the relative frequency of class y in the training set.

Figure 2: Naive bayes equation

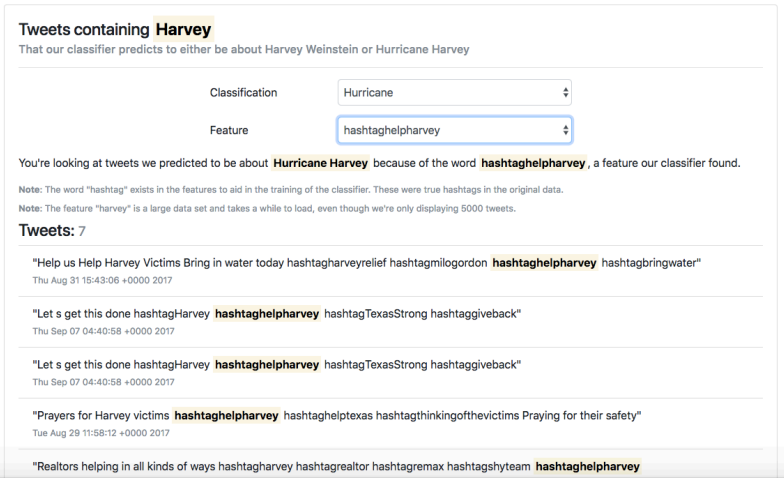

Figure 3: Text based vsualization

Blue = Harvey Weinstein

Red = Hurricane Harvey

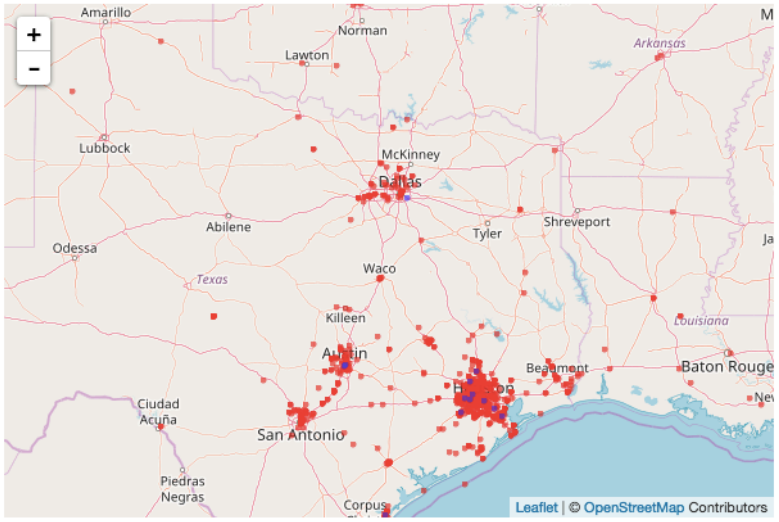

Figure 4: Geolocation visualization

3.2 Text Based Visualization

The first visualization is a text based analysis on the Twitter data. It portrays what distinguishes a Harvey Weinstein tweet from a Hurricane Harvey tweet. The visualization displays what our classifier found as the top features and the tweets that contain those features (with the feature highlighted in each displayed tweet). This visualization allows the user to look at either tweets about Harvey Weinstein or Hurricane Harvey and then decide which feature to look at. The interactivity allows for the user to pick which subject they want to explore, and then pick a feature to search. The search results bring up all of the tweets containing the specific feature word that was first searched (see Figure 3). The tweets are comprised of the text content inside, and the date the tweet was sent out. The back end of this visualization is a scikit-learn classifier fed about 60,000 tweets filtered on the words "weinstein" and "hurricane" as training data.

3.3 Geolocation Visualization

The second visualization is a geolocation map, plotting both the classified Harvey Weinstein and Hurricane Harvey tweets. The points were plotted based off of the geo-tagged tweets in our data. 2610 of the hurricane classified tweets were geocoded and while 198 of the Weinstein related tweets were geocoded.

3.4 Design Elements

The design process for visualization one was pretty simple. Our goal was to just create an interactive text based visualization that the user could sift through. For visualization two, we distinguished Hurricane Harvey tweets for Harvey Weinstein tweets by different colors. Tweets about the hurricane are in red, and tweets about Weinstein are in blue (see Figure 4).

4 DISCUSSION

For our text based visualization, the top 10 Hurricane Harvey features were as followed: 'hashtagtexassearchandrescue', 'hashtaghelpandhopeforhouston','hashtaghelpforharvey', 'hashtaghelpforhouston','hashtaghelpharvey','hashtaghelphousto n','hashtaghelpincrisis','hashtaghelping','hashtaghelpacripple' ,'hashtaghelpinghand'. For the top 10 Harvey Weinstein features, they were as follows: 'men' 'trump' 'rt' 'spacey' 'kevin' 'did' 'brad' 'harvey' 'weinstein' 'fuck'. In looking at our geolocation visualization, you can see clusters of Hurricane Harvey tweets in Texas, around San Antonio, Houston, and Dallas. There is also some tweets spread out in the United Kingdom. As for Harvey Weinstein tweets, those seem to be spread on the East and West coast.

REFERENCES

[1] Shamanth Kumar , Xia Hu , Huan Liu, A behavior analytics approach to identifying tweets from crisis regions, Proceedings of the 25th ACM conference on Hypertext and social media, September 01-04, 2014, Santiago, Chile

[2] F. Morstatter, J. Pfeffer, H. Liu, and K. M. Carley. Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose. ICWSM, 2013

[3] Zhiyuan Cheng , James Caverlee , Kyumin Lee, You are where you tweet: a content-based approach to geo-locating twitter users, Proceedings of the 19th ACM international conference on Information and knowledge management, October 26-30, 2010, Toronto, ON, Canada

[4] Takeshi Sakaki , Makoto Okazaki , Yutaka Matsuo, Earthquake shakes Twitter users: real-time event detection by social sensors, Proceedings of the 19th international conference on World wide web, April 26-30, 2010, Raleigh, North Carolina, USA

[5] Yafeng Lu , Xia Hu , Feng Wang , Shamanth Kumar , Huan Liu , Ross Maciejewski, Visualizing Social Media Sentiment in Disaster Scenarios, Proceedings of the 24th International Conference on World Wide Web, May 18-22, 2015, Florence, Italy

[6] Marcelo Mendoza , Barbara Poblete , Carlos Castillo, Twitter under crisis: can we trust what we RT?, Proceedings of the First Workshop on Social Media Analytics, p.71-79, July 25-28, 2010, Washington D.C., District of Columbia

[7] Sarter, N.B. and D.D. Woods. Situation Awareness: A Critical but Ill-Defined Phenomenon. The International Journal of Aviation Psychology 1, 1 (1991), 45-57.

[8] Sarah Vieweg , Amanda L. Hughes , Kate Starbird , Leysia Palen, Microblogging during two natural hazards events: what twitter may contribute to situational awareness, Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, April 10-15, 2010, Atlanta, Georgia, USA