Bad Weather Coming: Linking social media and weather sensor data

Shane Halse
Pennsylvania State University
State College, USA
seh297@ist.psu.edu

Aurélie Montarnal
IMT Mines Albi.
Albi, France
aurelie.montarnal@mines-albi.fr

Andrea Tapia
Pennsylvania State University
State College, USA
atapia@ist.psu.edu

Frédérick Bénaben
IMT Mines Albi.
Albi, France
benaben.emac@gmail.com

ABSTRACT

In this paper we leverage the power of citizen supplied data. We examined how both physical weather sensor data (obtained from the weather underground API) and social media data (obtained from Twitter) can serve to improve local community awareness during a severe weather event. A local tornado warning was selected due to its small scale and isolated geographic area, and only Twitter data found from within this geo-locational area was used. Our results indicate that during a severe weather event, an increase in weather activity obtained from the local weather sensors does correlate with an increase in local social media usage. The data found on social media also contains additional information from, and about the community of interest during the event. While this study focuses on a small scale event, it provides the groundwork for use during a much larger weather event.

Author Keywords

Twitter; weather; sensor data; social media;

INTRODUCTION

With the evolution of technology, we have empowered everyday people to become expert data collectors. This data comes in many forms; through the messages they post on social media, the geo-coded images they share, or simply the information such as traffic data. All this information has the potential to provide critical data to those who seek to leverage it.

The concept of humans as sensors is not one that is new to the field of crisis research (Goodchild, 2007). Often people have been placed in strategic places to report observations back to those in command. However, more recently there has been an upsurge in attempting to understand how we may be able to take data produced by people in the field and integrate it into a situational awareness model (Vieweg et al. 2010). Using humans as sensors is not without its flaws as we are left with the questions of; where does it come from? Who provided it? And how much of it is representative of factual data? Even without these questions answered the value of the information may outweigh our doubt of its source.

During a crisis there are also additional sources of information that can be gained from the community. One example is that of weather sensor data found through the weather undeground API (Wunderground, 2017)[1]. This data is comprised of both personal weather stations and officially managed weather stations such as those found at airports. While these sensors provide raw weather data such as wind and gust speed, temperature and rainfall, they do not provide reactionary data about how it has, or is, affecting the community. Thus we are left with questions of how can this data be used, what does it say and how is the community affected by a severe weather event.

In order to enhance both the information gained from weather sensors and that found on social media, this paper will examine how physical weather sensor data correlates with how people react within social media. This serves two purposes. The first is to improve the validity of social media messages during an event. That is, through outlining the timeline of a local event through the use of weather sensor data, we can begin to understand how

[1] https://www.wunderground.com/weather/api/d/docs - A weather API designed for developers

severely and area is affected by the event in question. For example, it would allow us to know when the weather event peaks in its severity. Establishing this timeline thus allows us to know when a given area is experiencing the greatest impact. Second, through investigating social media data and how it corresponds to this timeline we can begin to understand how the community of this area is reacting to the event. Furthermore through a contextual analysis of the local social media data from the affected area, we can develop an understanding of how the said community is reacting to the event. In doing this we can further the information gained from digital sources. That is, the raw sensor data is elaborated upon through the use of social media messages and the social media data is supported by the raw sensor data through establishing a factual timeline.

For this study we selected the college town of State College located in central Pennsylvania, USA population of approximately 100,000 people. This location is not known for being highly tornado active, as such we believe that a severe weather warning would have an effect on the amount of weather related Tweets found from that area. In addition on May 1st, 2017 there was a severe weather alert issued for the area that corresponded to tornado and high wind watches. The storm did hit the area later that evening but only caused minor damage to houses and other infrastructure. There were no fatalities reported. While this event may have been only minor in comparison to other historical storms, it was selected because of this. That is, we wanted to find an event that was representative of common occurrences across the country as our focus is to see how even minor weather events can affect how people react on social media.

RELEVANT LITERATURE

Social media has it all. Any topic, any idea and any issue is likely being discussed somewhere. This makes it an ideal candidate for data mining however, due the abundance of information available, makes mining it problematic. While this does hold true in most areas of interest, in crisis response there is an immense effort to obtain only the data related to a particular crisis. It is important to realize that social media has become a crucial tool in communication that has allowed both the general public and governmental agencies to reach populations of an area that may be or is being affected by a severe event (Kavanaugh et al., 2012). Previous works by Crooks, Croitoru, Stefanidis, & Radzikowski, (2013) and Sakaki, Okazaki, & Matsuo, (2010) have shown that Twitter can indeed be used as a social sensor to detect earthquakes. In some cases providing information about the quakes where sensors may not have been available. Furthering this line of thinking, Takahashi, Abe, & Igata, (2011) asked and suggested that Twitter data may actually be able to replace that of traditional sensors. Their work found that Twitter could indeed be used to detect the pollen count for a given area through detecting certain keywords related to hay fever. While their goal was to replace the use of sensors for detecting pollen count by observing Twitter data feed, we feel that rather than replace sensors we could instead complement them. That is, in our paper we examine weather sensors which provide wind speed. Knowing that a weather event is occurring in a given area allows us to determine what impact it is having through the use of social media feeds. In addition, it may also provide insights to first responders should the impact be severe. A further example that can be very quickly applied to the domain of crisis response is that of detecting traffic delays (Daly, Lecue, & Bicer, 2013). While the author's focus was outside this domain, the examples of social media usage to extend traffic diagnosis could be used to provide valuable information about evacuation routes during a crisis event.

In a study by Mccormick, (2016) the author discusses the obstacles of use and implementation of social media tools for the crisis responder. She outlines four major barriers of the adoption of social media data which include the verification of data, security of the information, the liability issues and the subjectivity of the information. It is described in the article that data cannot always be verified or fact checked. As such, carefully constructed rumors have a tendency to propagate the network as quickly as the truth. To combat this, research has attempted to address this issue through the use of trust and credibility detection. As described by Mendoza, Poblete, & Castillo, (2010), there is a prevalence of false tweets and rumors. While, this false information does propagate through the network differently, it still requires it to spread before it can be detected. In addition, Gupta, Kumaraguru, Castillo, & Meier, (2014) developed a tool to detect real-time credibility called TweetCred. This leverages machine learning and examines 45 key features such as account age, tweet age, word usage, follower to followee ratio, etc. Research on filtering social media topic data, specifically in relation to a crisis, can be found in the works by Olteanu, Castillo, Diaz, & Vieweg, (2014). In this work they describe the development of a lexicon used to filter and mine data from social media blogs with the goal of reducing the amount of human effort required to separate relevant information from noise. While reducing the amount of information is an important first step in utilizing social media data within a crisis, the next step becomes one of determining how accurate this information is.

One major drawback to detecting credibility or trust within a social media network is being able to compare data thought to be valid. Gupta et al., (2014) used a method of crowdsourcing and labeled data to establish this "ground truth". Jurgens, Finnethy, McCorriston, Xu, & Ruths, (2015) used geo-locational meta-data to establish the ground truth of a user's location as they found self-report to be less than adequate. Yin & Tan, (2011) argue that ground

truth for a large dataset can be deduced from a small sample set, but that the small set must still be verified. Unfortunately, all these methods require a great deal of human effort in their implementation. In this paper we aim to show that through the use of sensor data this ground truth can be established quickly and efficiently. In order to improve the concept of finding credible tweets we take a first look at how sensor data can be linked to messages found on social media sites such as Twitter. Through showing that they can indeed be linked we provide another method by which tweets can be programmatically fact checked. For example, if a sudden surge in the amount of information concerning a storm comes out of a certain area, checking this against weather sensor data from that area allows us to determine if these messages have validity.

In addition to utilizing weather sensor data for credibility verification, researchers have shown the importance of weather data to emergency managers. Baumgart, Bass, Philips, & Kloesel, (2006) show that these emergency managers have trouble integrating advanced radar imagery into their decision making process due to issues of interpreting velocity. Furthermore, most weather related decisions are made through the use of data provided by the National Weather Service (NWS) which is discussed by Miller, Black, Williams, & Knox, (2016). Here they describe how the National Weather Services (NWS) wind warning systems may be too conservative. Leveraging social media data within this decision making process could help emergency management better understand when and where the weather event is at a particular time.

To investigate how both weather sensor data and social media data can complement each other we developed the following research questions:

- Is there an association of a weather event described by sensors, and social media behavior?
- How important is the amount of tweets that can be of further help, in providing decision makers with information that is complementary to the sensors' data?

Our findings indicate that there is indeed a strong correlation between the data found on social media and the timeline generated by data found from the weather sensors for a given area. Furthermore, the contextual analysis of the social media data did provide further information usable by those responding to the crisis.

Overall, while our analysis was based on a severe weather warning, it is believed that utilizing similar methods of analysis during a severe event that has a much greater impact would yield a larger correlation and provide much more contextual data within social media.

METHODS

Data Collection - Social Media

Data was live collected from Twitter using the Twitter API and customized software for the event on 05/01/2017 (0:01-23:59). In total there were 5,143 tweets contained one or more of the keywords or key phrases found in the crisislex limited to an area 50 miles around the center point of State College. This was achieved using the Twitter "near this place" call seen on the advanced search page. The timeline for this event can be seen from both Twitter data and the weather sensor data. Throughout the morning many warnings indicating a potential storm were issued, and at approximately 14:09 the NWS issued an official tornado watch. The highest rate of weather related tweets was 65 between 17:10 and 17:15 with the wind speeds peaking at approximately 18:24.

As we collected all the tweets for this time period and location, we began the process of refining them. In order to ensure only crisis related words were filtered we leveraged the lexicon found at http://crisislex.org/crisis-lexicon.html[2] (Olteanu, 2017) which consisted of 380 terms. These terms included words such as "flood, storm, tornado, torrential rains, etc" and filtered out tweets that did not contain these words. This left us with a total of 5,143 tweets over the 24 hour period. The frequency of these tweets is shown in Figure 1 below in a 5 minute interval.

[2] http://crisislex.org/crisis-lexicon.html - A recommended lexicon for Twitter querying contains an automatically-generated and human-curated list of terms found to be frequently related to disasters

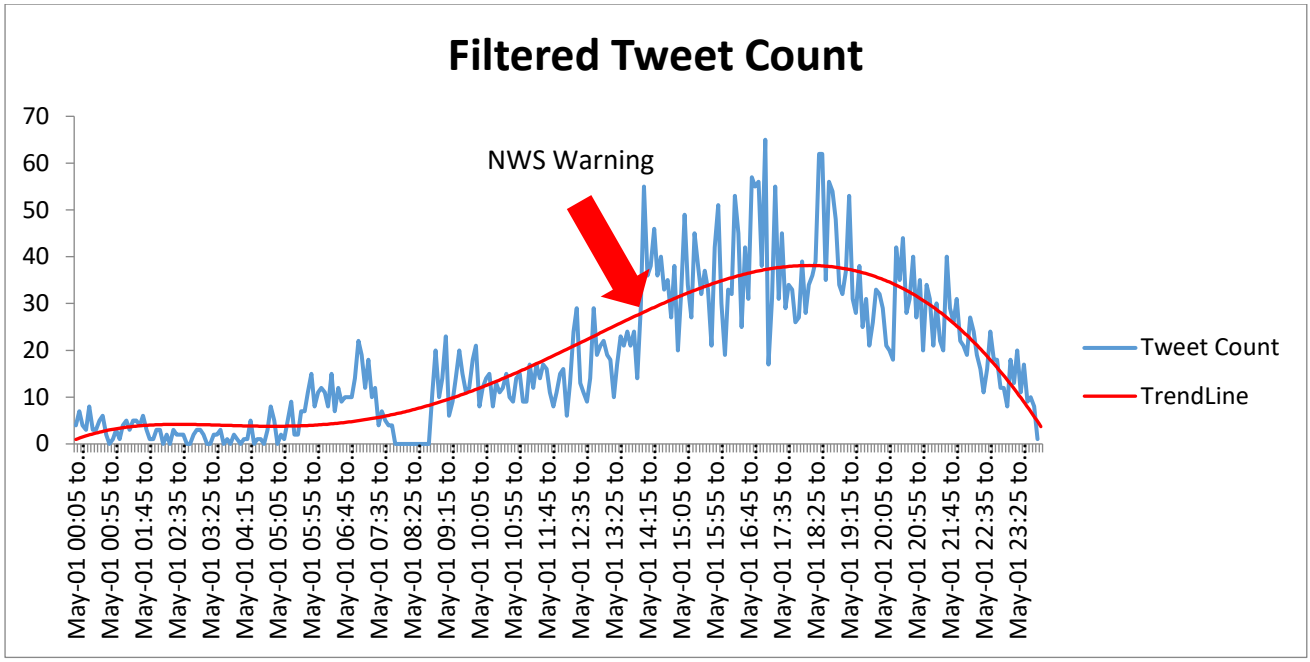

Figure 1 - Filtered Tweet Count

Data Collection - Weather Sensor

Weather sensor data was collected from 34 weather stations located in the area using the Weather Underground API (Wunderground, 2017). The locations of these weather sensors can be seen in Figure 2 - Weather Sensor Locations.

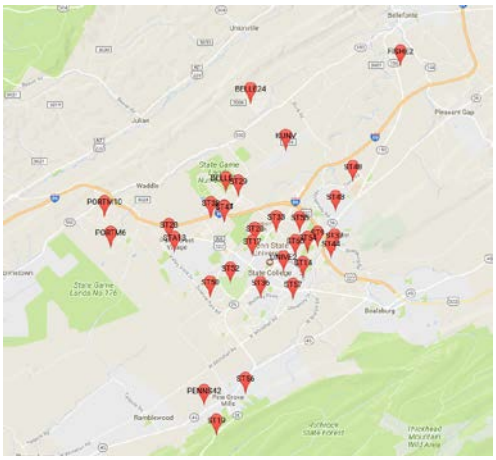

Figure 2 - Weather Sensor Locations

Twelve weather stations were removed from our data source due to inactivity or not having the ability to measure wind speed. From each of the weather sensors, we collected wind speed (in miles per hour) from the remaining 22 active sensors in 5 minute intervals. 4all sensors to give us the wind speed data for the area of interest. This data is shown in Figure 3 below.

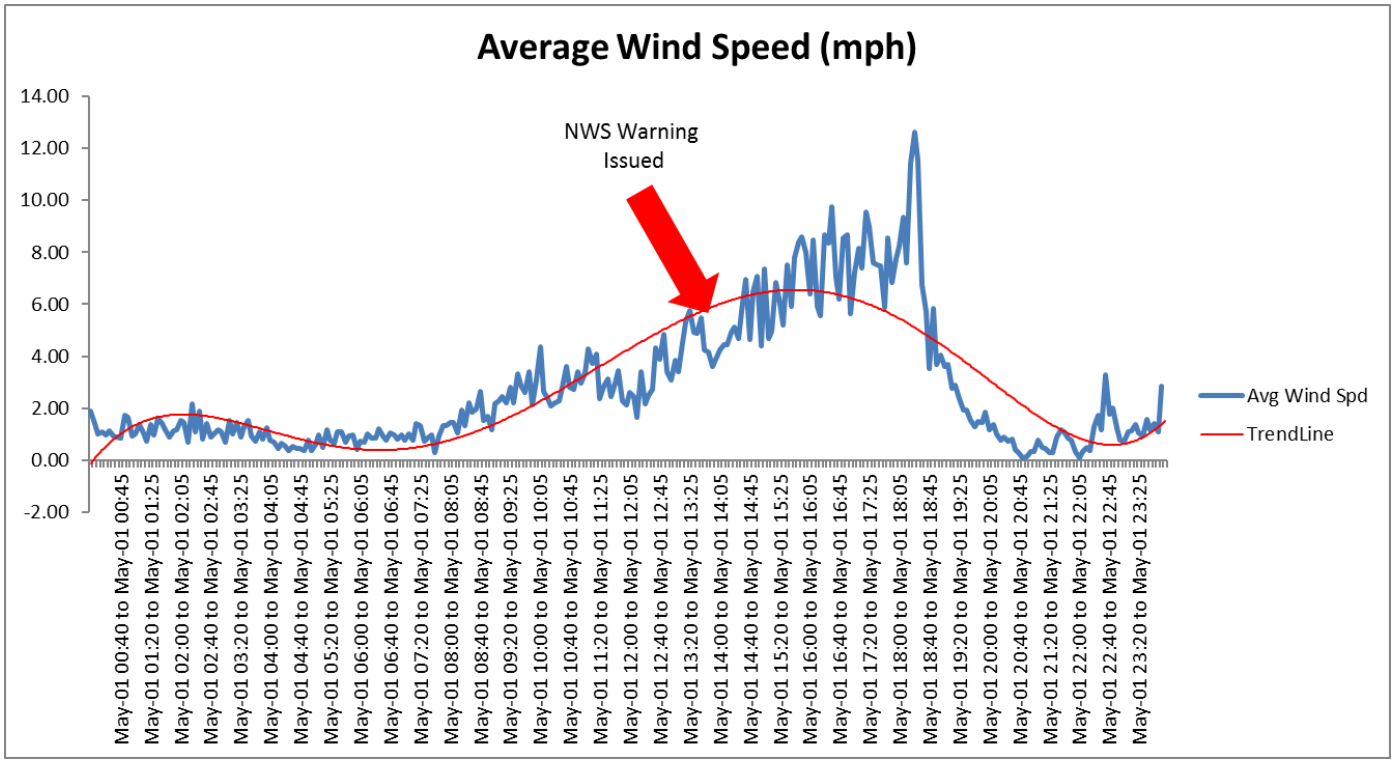

Figure 3 -Average Wind Speed (mph)

Statistical Analysis

To examine our first research question of how closely social media data correlates with weather sensor data for a day in which a severe weather warning occurred, we utilized a simple linear regression. To set this analysis up we used the number of Tweets related to weather data as our dependent variable with the wind speed as our independent variable. We postulated a positive correlation between the wind-speed and the number of tweets produced. To examine this, we used the R studio with the built in "lm()" functionality (R Core Team, 2015).

To examine our second research question, we performed a contextual analysis of the tweets for two time periods. The first was eight hours before the peak wind conditions from 10:54 am to 11:54 am and the second between 17:54 and 18:54 which corresponded to half an hour before and after the peak winds. This data was then coded to fit into one of 4 categories;

- Weather warning tweet or retweet for example: "*Thunderstorm warning and tornado watch in full effect. Finals remain on.*"
- Informational tweet for example: ("*Power flickering on Vairo Blvd in State College. Ton of rain &wind but not much thunder & lightning. Best storm in a long time! #PAwx*"),
- Tweet related to the event but containing no information for example: "*Well this is a freaky storm*"
- Unrelated tweets for example: "*One person suffered a nonlife threatening gunshot wound inside Norrell Annex. Three others were inside. Un-injured.*"

FINDINGS

For 05/01/2017 a total of 5,143 tweets contained one or more of the keywords or key phrases found in the crisislex dataset. From this data, as a first analysis, the Pearson correlation coefficient was calculated between the number of Tweets and the average windspeed, and, as intuitively observed when comparing Figure 1 and Figure 3, the result seems promising with a positive correlation of 0.63. In line with this correlation between the two features, a simple linear regression was calculated to predict the number of weather related Tweets based on the wind speed gathered from personal weather stations. A significant regression equation was found ($F(1,285) = 191.4$, $p<0.0001$), with an R^2 of 0.3997. The number of weather related Tweets is equal to (wind speed) miles per hour when tweets are counted. The number of tweets increased 3.770 for each mile per hour of wind speed. A scatter plot of these results can be seen in Figure 4 - Tweets vs Wind speed.

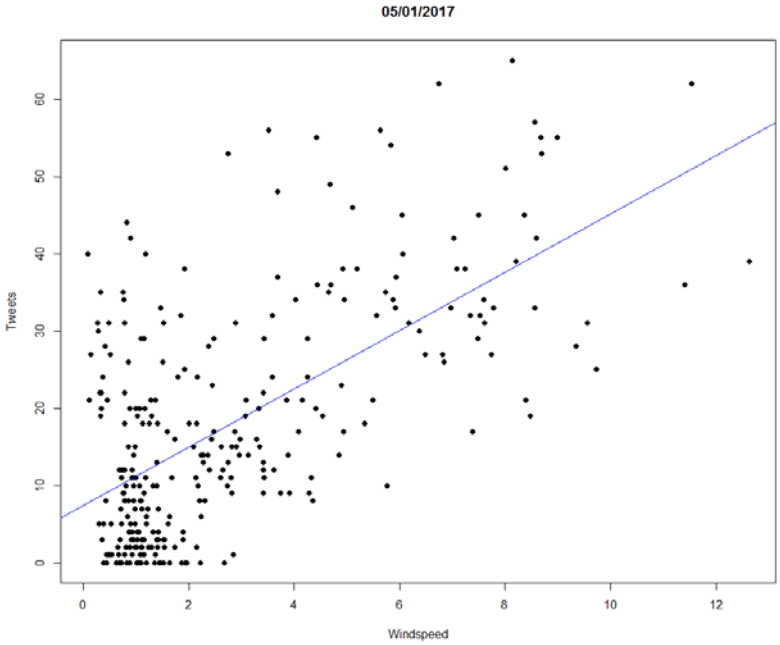

Figure 4 - Tweets vs Wind speed

For the second part of this study we examined the context of the tweets to determine what information, if any, was being distributed during the peak winds. This was done by having a set of coders look through the tweets and determine if the tweet had any information related to weather such as mentioning, rain, wind, clouds, sunshine, etc. As some of the tweets could belong in multiple categories it was decided that the most representative category should be used. In the case of the "warning" categories, if a message contained both a warning and informational data such as "Severe Thunderstorm Warning in effect in Beaver Greene and Washington Counties until 215 PM #pawx", but was in the format above we placed it in the warning category due to the standard format i.e. starts with "Severe Thunderstorm Warning in effect….". An example of an informational message would be that of "Police scanner: Tree down on Camp Meeting Road near Leet/Leetsdale. Just up the hill from Beaver Street."

One hour was selected for both coders to label to determine the inter coder reliability. A Cronbach alpha score of 0.92 was achieved. Due to this high inter coder reliability we then separated the rest of the data for labeling in which each coder was given a partial dataset. For the first interval (interval 1), there were 159 weather related tweets found the morning of the storm between 10:54 am and 11:54 am. In Figure 5 we can see a chart of the breakdown of tweet context for this time period.

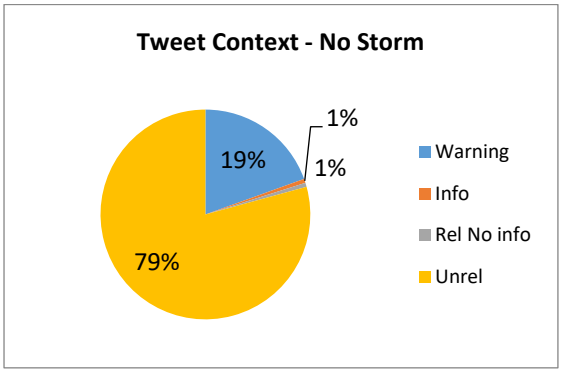

Figure 5 - Tweet Context - No Storm –Interval 1

It is important to note that while there was no active storms in the area of interest during this time period, there were tweets related to storms and flooding in Eustace Texas and DC Metro that were able to bypass our search filter. For the second interval (interval 2), there were 510 weather related tweets found half an hour before and after the storm event (17:54 through 18:54) in our area. In Figure 6 we can see a chart of the breakdown of tweet context for the half an hour before and after the storm.

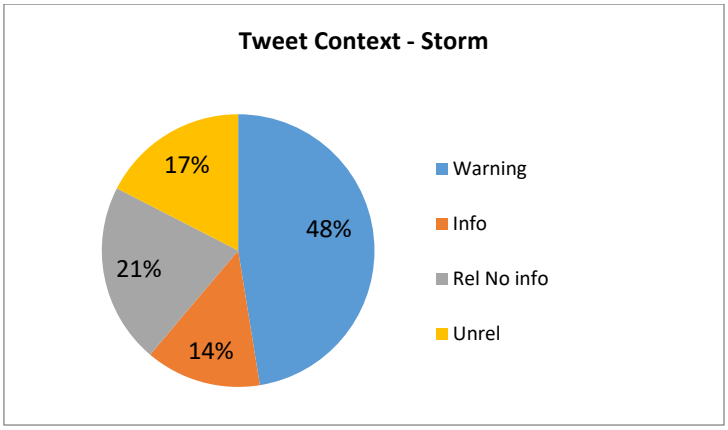

Figure 6 - Tweet Context - Storm

DISCUSSION

Leveraging the power of citizen sourced data is one way in which crisis response managers can gain insights into an effected area during a crisis event. In the case of this study we leveraged citizen sourced data in two ways; the first is through the posts of local users found in Twitter, and the second is from personal weather stations provided by local community members through wunderground.com[1].

Our findings indicate that there was indeed a positive correlation between the amount of weather related posts and an increase in wind speed for a local area. This supports the idea that when an event, such as a severe weather warning, occurs, people will turn to Twitter to discuss and share information about it. While our study is limited to only a particular area, and a particular event, further work could leverage similar methodologies to determine if our findings are supported in other geographic locations. From the contextual analysis we can see that despite using a well-known lexicon, we are still left with tweets that are unrelated to the event we are interested in.

Rather than relying on a "fuzzy" data source such as Twitter, to detect the occurrence of an event, using sensors data allows a more efficient and accurate result. In turn, the study shows that social media are still an important source of data, mostly because it brings complementary information. However, it was observed that even during this small scale event we can see that a majority of the weather related tweets are those that share the weather warnings, followed by related tweets to the event without information as well as unrelated tweets. Only small proportion of the tweets contain informational data which would be invaluable to enhance the situational awareness of the decision-makers. Further work would thus be needed to refine our filtering methods to exclude the weather warnings, non-informational and unrelated data allowing us to obtain only data with actionable information.

The second part of this research was to examine the categorical makeup of tweets both before and during a particular event. It is hypothesized that that the output of information that is related to the event would increase during the peak wind speeds. While we acknowledge the defined categories can be further refined and expanded, our findings show that a much larger percentage of Tweets are event related when a particular event is occurring. In the case of this research we found that before an event only 21% of Tweets contained information that could be linked to the storm, however during the event 83% of the Tweets contained some information related to the storm. While most of these messages were warning related Tweets and retweets making up 48% of the total filtered Tweets for this time period, 14% of these did have some informational component associated with them and 21% contained no information but were still storm related. Further work would be needed to examine how these findings could be used. One possible avenue could be in determining when a particular threshold has been reached that could then be leveraged for automating the detection of weather related events. Additionally, by expanding the categories beyond related and unrelated to an event, one could begin categorizing different types of Tweets to serve as markers for determining event type, impact and scale. In addition, while we did find that the weather related Tweets did increase during and event further work would be needed to examine the scale of the event impacts what proportion of Tweets are related to it.

For example, this small scale event contained 5,143 locally filtered Tweets half an hour before and after a storm. Of those, 510 were weather related. Of the 510 weather related tweets, only 70 contained actionable information. Even within these 70 tweets 39 were retweets leaving only 31 original tweets. This means that even through rigorous filtering we are left with very few tweets which would be useful to crisis responders.

CONCLUSION

This paper illustrates the importance of taking into account both sensors and social media data, because of the complementarity of their intrinsic nature – rational but limited in the value it can take in the first case, and fuzzy but rather open in terms of value in the second case – and of their reliability rates – rational sensor data is most likely to be more reliable and objective than human-written data.

In this sense, we have shown in this paper that utilizing physically located sensor data allows us to gain an understanding of what is happening in an area. That is, through the use of weather sensors we'd be able to determine when a weather event of anomaly has occurred. However, this information does not allow us to understand in what way a particular event affects the community. In combining weather sensor data with data found on a social media website, we can begin to understand how the community is reacting to the event.

Utilizing Twitter as a source to elaborate on the data found from the sensor data we can provide much more granularity about the impact of an event. Not only does it provide us with more details for a given area, but it opens the door for time sensitive analysis to take place. This is especially true in the crisis related domain were time is a precious commodity. Sensor data can be regarded as factual information provided the sensor is working accurately. In our work utilizing multiple sensors within an area insures that the probability of false data is minimized. This, combined with the Twitter data, allows us to determine the severity of the event and the impact it has on the community. In the case study provided in this paper we saw that the impact was minimal resulting in only minor damage such as flooding and downed trees. However as the scale and magnitude of the event increases, it is predicted that this would be reflected in the social media data along with the weather sensor data.

While, the event we selected was small in scale, a larger scale event may yield more fruitful results. That is, while there were severe weather warnings, no actual weather event occurred in a populated area. Further research would be needed to examine a local event that did actually affect a higher populated area. While increasing the scale of an event would also yield a larger number of informational tweets, it adds additional complexity to the filtering process. In addition, further work would also be needed to determine if these findings hold true beyond a particular geographic region. Also, examining how language or cultural differences influence Twitter usage during an event could also provide valuable insight. Using weather sensor data could provide the ability to pinpoint the location and time of an event, which could in turn be used to further investigate and improve filters.

FUTURE WORKS

This paper begins to examine the potential for providing additional data to elaborate on traditional sensors through the use of social media. That is, sensor data may provide an indication that an event has or is occurring, whereas social media data could potentially provide more detailed information concerning that event. This work takes a first look at how these two data sources are linked laying the ground work for future work. The first area where this work could prove useful is in establishing some sort of ground truth. That is, through using sensor data which is believed to be truthful, we can verify and vet information found of social media. While this paper focusses on weather data, many other sensors could provide valuable data which then in turn could be used to validate social media data. Social media data however, would allow us to expand beyond just the information gained from the sensors.

In addition, another area of future work is to further expand upon the contextual analysis. While we only focused on if the data was related to the storm or not, this can be expanded to provide much more granularity with respect to the categories a Tweet may belong to. For example, Tweets could be categorized into those which provide information related to damage, impact or scale. These in turn could be used to provide insights into how the community has or is being affected by the event.

REFERENCES

Baumgart, L. A., Bass, E. J., Philips, B., & Kloesel, K. (2006). Emergency Management Decision-Making during Severe Weather. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* , *50*(3), 381–385. https://doi.org/10.1177/154193120605000336

Crooks, A., Croitoru, A., Stefanidis, A., & Radzikowski, J. (2013). #Earthquake: Twitter as a Distributed Sensor System. *Transactions in GIS*, *17*(1), 124–147. https://doi.org/10.1111/j.1467-9671.2012.01359.x

Daly, E. M., Lecue, F., & Bicer, V. (2013). Westland row why so slow? *Proceedings of the 2013 International Conference on Intelligent User Interfaces - IUI '13*, (March), 203. https://doi.org/10.1145/2449396.2449423

Goodchild, M. F. (2007). Citizens as sensors : the world of volunteered geography, (November), 211–221.

https://doi.org/10.1007/s10708-007-9111-y

Gupta, A., Kumaraguru, P., Castillo, C., & Meier, P. (2014). TweetCred: Real-time credibility assessment of content on Twitter. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *8851*, 228–243. https://doi.org/10.1007/978-3-319-13734-6_16

Jurgens, D., Finnethy, T., McCorriston, J., Xu, Y. T., & Ruths, D. (2015). Geolocation prediction in Twitter using social networks: A critical analysis and review of current practice. *The 9th International Conference on Weblogs and Social Media (ICWSM)*, 1–10. https://doi.org/10.1002/0471264385.wei0223

Kavanaugh, A. L., Fox, E. A., Sheetz, S. D., Yang, S., Tzy, L., Shoemaker, D. J., … Xie, L. (2012). Social media use by government : From the routine to the critical. *Government Information Quarterly*, *29*(4), 480–491. https://doi.org/10.1016/j.giq.2012.06.002

Mendoza, M., Poblete, B., & Castillo, C. (2010). Twitter Under Crisis: Can we trust what we RT? *Workshop on Social Media Analytics*, 9. https://doi.org/10.1145/1964858.1964869

Miller, P. W., Black, A. W., Williams, C. A., & Knox, john A. (2016). Maximum Wind Gusts Associated with Human-Reported Nonconvective Wind Events and a Comparison to Current Warning Issuance Criteria. *Weather and Forecasting*, *31*(2), 451–465. https://doi.org/10.1175/WAF-D-15-0112.1

Olteanu, A., Castillo, C., Diaz, F., & Vieweg, S. (2014). CrisisLex: A Lexicon for Collecting and Filtering Microblogged Communications in Crises. *Proc. of the 8th International Conference on Weblogs and Social Media*, 376. https://doi.org/10.1.1.452.7691

Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). Earthquake shakes Twitter users: real-time event detection by social sensors. *WWW '10: Proceedings of the 19th International Conference on World Wide Web*, 851. https://doi.org/10.1145/1772690.1772777

Takahashi, T., Abe, S., & Igata, N. (2011). Human-Computer Interaction. Towards Mobile and Intelligent Interaction Environments. In *Human-Computer Interaction. Towards Mobile and Intelligent Interaction Environments: 14th International Conference, HCI International 2011, Orlando, FL, USA, July 9-14, 2011, Proceedings, Part III* (Vol. 6763, pp. 240–249). https://doi.org/10.1007/978-3-642-21616-9

Vieweg, S., Hughes, A. L., Starbird, K., & Palen, L. (2010). Microblogging during two natural hazards events. Proceedings of the 28th International Conference on Human Factors in Computing Systems - CHI '10, 1079. http://doi.org/10.1145/1753326.1753486

Wunderground.com. (2017). API | Weather Underground. [online] Available at: https://www.wunderground.com/weather/api [Accessed 30 Jul. 2017].

Yin, X., & Tan, W. (2011). Semi-supervised truth discovery. *Proceedings of the 20th International Conference on World Wide Web - WWW '11*, (c), 217. https://doi.org/10.1145/1963405.1963439