Assessing the Stability of Tweet Corpora for Hurricane Events Over Time: A Mixed Methods Approach

Estella Z. Xin
University of Texas at Austin
Department of Computer Science
Austin, TX 78712 U.S.A.
estellaxin@utmail.utexas.edu

Dhiraj Murthy
University of Texas at Austin
Moody College of Communication
Austin, TX 78712 U.S.A.
dhiraj.murthy@austin.utexas.edu

Nandhini S. Lakuduva
University of Texas at Austin
Department of Computer Science
Austin, TX 78712 U.S.A.
nlakuduva@utmail.utexas.edu

Keri K. Stephens
University of Texas at Austin
Moody College of Communication
Austin, TX 78712 U.S.A.
keristephens@austin.utexas.edu

ABSTRACT

When natural disasters occur, various organizations and agencies turn to social media to understand who needs help and how they have been affected. The purpose of this study is twofold: first, to evaluate whether hurricane-related tweets have some consistency over time, and second, whether Twitter-derived content is thematically similar to other private social media data. Through a unique method of using Twitter data gathered from six different hurricanes, alongside private data collected from qualitative interviews conducted in the immediate aftermath of Hurricane Harvey, we hypothesize that there is some level of stability across hurricane-related tweet content over time that could be used for better real-time processing of social media data during natural disasters. We use latent Dirichlet allocation (LDA) to derive topics, and, using Hellinger distance as a metric, find that there is a detectable connection among hurricane topics. By uncovering some persistent thematic areas and topics in disaster-related tweets, we hope these findings can help first responders and government agencies discover urgent content in tweets more quickly and reduce the amount of human intervention needed.

CCS CONCEPTS

• **Human-centered computing → Collaborative and social computing;** *Social media*; Social content sharing • **Computing methodologies → Machine learning;** Topic modeling

KEYWORDS

Natural disasters, Twitter, qualitative data, longitudinal research, topic modeling, machine learning

ACM Reference format:

Estella Z. Xin, Dhiraj Murthy, Nandhini S. Lakuduva, and Keri K. Stephens. 2019. Assessing the Stability of Tweet Corpora for Hurricane Events Over Time: A Mixed Methods Approach. In *Proceedings of the International Conference on Social Media & Society*, Toronto, Canada (SMSociety). DOI: https://doi.org/10.1145/3328529.3328545

1 INTRODUCTION

When Hurricane Harvey hit land, it caused floods of over four feet in parts of Texas, displaced an estimated 30,000 people from their homes, and resulted in at least 72,000 people requiring rescue [6]. As the storm left residents stranded without traditional information sources such as television or radio, people increasingly relied on social media to get real time information [20]. Some also opted to use social media to request or give aid due to overwhelmed 9-1-1 emergency phone lines. In one case in the aftermath of Hurricane Harvey, a Twitter post of elderly residents in a flooded nursing home went viral and helped bring rescuers to that area. Rescuers were then able to locate the nursing home and rescue the residents based on the information from the tweet [17].

During natural disasters, many relief efforts and organizations use social media to understand the situation of victims, rescuers, and others. People affected by the disaster also use social media to seek help. These relief-oriented frontline services and providers have a specific need for real-time identification of users in need of assistance. But, this remains a challenging problem for official emergency services as well as unofficial responders. Specifically, the high-volume of noise (i.e., content not directly related to a disaster) in social media streams makes it difficult to find legitimately relevant information on social media. Twitter, in particular, has a high volume of data, but much of it is considered noise.

Social media posts providing or requesting aid can be of great value for people in times of natural disasters. However, not all posts requesting or offering aid go viral and most are lost within the noise. Therefore, developing methods to classify whether the social media post is signal (relevant content to identifying urgent disaster-related social media content such as requesting or offering urgent help) or noise (less- or non-relevant content to the disaster) is an important task [15]. It should also be noted that there is no consensus on gold standards of relevant tweet-associated keywords and phrases to search for on Twitter during hurricane events and this introduces a lot of bias as researchers scramble to collect data as an event unfolds. Post-event, researchers might be willing to compromise and work with whatever event-related data they can secure.

Previous work indicates that images collected from private social media posts during Hurricane Harvey can be a great medium for determining if a post is signal or noise [15]. In this study, we ask whether social media text collected from private social media posts (obtained through interview methods) resemble both recent and historical disaster-related tweets, or whether the former, gold standard remains qualitatively different from this more easily mineable public Twitter data. If tweet data is both persistent over time and has some similarity to private data, tweet data could provide some guidance for stakeholders tasked with mining social media data during disasters.

We leverage unsupervised learning by using LDA models—an unsupervised generative statistical model that clusters data into groups using words as attributes—to find relationships among topics (i.e., collections of statistically associated words) in the hurricane tweets. By using unsupervised learning, LDA, to identify topics that may be useful in finding signal tweets, ultimately, we plan to integrate this process into a pipeline that can automate the entire filtering process in real time. Our hypothesis is that there is some textual continuity between hurricane events over a longitudinal timeline. If this is the case, certain keywords and phrases could be monitored real-time as storm events unfold, rather than the current emphasis towards collecting data by hashtags. A pro-active rather than reactive approach has real benefits both for those affected by a disaster as well as relief organizations. Given there is no gold standard for social media collection during natural disasters, we seek to evaluate whether we can begin to establish loose standards to guide data collection beyond hashtags, which combined with existing practices of data sharing would be of real benefit to the field.

2 RELATED WORK

A significant amount of work has been done on classifying Twitter data as signal or noise during times of natural disasters [1, 12, 22]. In addition to filtering tweets, building a pipeline to accomplish this task real time during national disasters has also been a research focus. Existing pipelines including Artificial Intelligence for Disaster Response (AIDR), perform automatic classification of signal versus noise tweets for different disasters with the help of human intelligence to reach a higher accuracy [10]. AIDR was used to reach an accuracy of 80% for collecting signal tweets real time during an earthquake in Pakistan in 2013. AIDR is a versatile pipeline approach that can be used for all disasters, but it relies heavily on human intelligence because the classification process requires volunteers to provide labels for tweets in order to achieve high accuracy levels.

Similarly, Tweedr [2] is a pipeline that extracts actionable information from Twitter to be used by disaster relief workers through processes of classification, clustering, and extraction. Tweedr, like AIDR, uses supervised learning and requires human labeling to achieve its results [2].

Most work in classifying social media posts in times of disasters has used public data on platforms such as Twitter that uses APIs that allow researchers to scrape data easily [4]. Due to the accessibility of public data, private posts that are inaccessible via APIs are often overlooked as training data (i.e., data used to train an algorithm). However, previous work indicates that social media posts collected via public data have high amount of noise. During times of disaster, news reporting and celebrity related posts have been found to dominate social media platforms [15]. Approaches to collect private data through people who were directly affected by the disaster to use for training machine learning algorithms are not standard practice. This is due to the high levels of time and effort needed to obtain these data, as well as the lack of field-based data collection skillsets amongst some crisis informatics teams.

However, such data are highly valuable as they can be used to better train models while working with smaller sample sizes. The latter is not only important as it ipso facto reduces the likelihood of more noisy patterns inherent to most big data approaches, but also benefits from the value of ‘small data’ [9] that can be more easily and accurately interpreted.

A natural language processing technique, Latent Dirichlet Allocation (LDA) has been used extensively for studying crisis-related tweet corpora [11]. This algorithm is an unsupervised, probabilistic topic modeling technique that has proven to be useful in identifying topics, clusters or assemblages of statistically related words from large corpora [7]. LDA can help identify the topics present in each dataset and successfully categorize new tweets that are not currently categorized to a topic. Each document results in a distribution of topics from which words are drawn. LDA also provides a document-probability distribution over each topic that highlights the level of importance of each topic to the document [21].

There is also a significant amount of work establishing the effectiveness of LDA in finding topics [14, 19, 21]. There have been numerous implementations of LDA on platforms such as Tweedr to help analyze the differences between signal and noise data during natural disasters [1, 2]. However, there is a lack of research in comparing clusters across multiple hurricanes to find useful topics with multiple datasets in similar natural disaster scenarios.

3 METHODS

3.1 Data

We collected data from six hurricanes/storm events ranging from 2010 to 2018 – Bonnie, Sandy, Harvey, Lane, Florence, and Michael (as detailed in Table 1). In addition, we collected interview data in the field, and used noisy, random Twitter data. Given the span of these different events, our data collection methods varied over time. For Bonnie and Sandy, we made calls to the Twitter REST API during the events using php scripts. For Lane, Florence, Michael, we used the social media data collection platform Netlytic [8] to make calls to the Twitter REST API in 15-minute intervals using various hashtags and keywords such as #Florence or "storm" to get relevant hurricane data. For Harvey and the noise dataset, we extracted tweets from an existing python-based 'Spritzer' Twitter STREAM API call running on Amazon Web Services, which was returning approximately 1% of tweets worldwide. We also uniquely conducted 33 in-person field interviews with rescuees and rescuers in the immediate aftermath of Hurricane Harvey in the greater Houston area over a three-month period by working with local community and relief organizations. As part of this process, we elicited photos, videos, and text-based posts from respondents' private Facebook, Nextdoor, and Snapchat feeds that we manually curated (detailed as Private (Harvey) in Table 1 and hereafter also referred to as the "signal dataset"). We had respondents take screenshots of their feed and send them to our research team. All of our private social media data was obtained through individuals who volunteered to give their private posts to contribute to this study following an approved Institutional Review Board protocol. We used Google Cloud Vision API to then extract the text present in these images into individual text files (i.e. individual 'documents' that are similar to a tweet in length) to compare with our public API-derived Twitter data.

In addition, we conducted 33 semi-structured qualitative interviews of rescuees and volunteer rescuers with an average length of 37 minutes. Respondents were all interviewed and their interview transcripts were incorporated as documents as well as being used in full in the Private (Harvey) dataset outlined in Table 1. This private data reveals important information, such as types of damages, emotional reactions, etc. about the hurricanes that would be difficult to extract from Twitter due to the volume of noise. Table 1 details these data, which we used for LDA cluster analysis.

We chose these different methods of retrieving social media posts to reveal correlations among topic clusters for all hurricane posts and determine if these topics are differentiable from topics of non-hurricane related posts, or noise data. Following established practice [e.g., 13], we removed stop words as well as stemmed and tokenized the words.

Event	Year	Size (# of tweets/ statements)	Collection
Bonnie	2010	20,601	Twitter Spritzer
Sandy	2012	425,423	Hashtag-based
Harvey	2017	555,621	Twitter Spritzer
Noise	2017-2018	49,000	Twitter Spritzer
Private (Harvey)	2018	3,000	Qualitative Interviews
Florence	2018	5,483,030	Hashtag-based
Lane	2018	653,514	Hashtag-based
Michael	2018	3,265,706	Hashtag-based

Table 1. Data collection details

3.2 Overview

In this study, we focus specifically on hurricanes with the goal of achieving higher performance on classifying signal tweets than AIDR and Tweedr. We also aim to reduce the amount of human interaction in the data processing by analyzing patterns from different LDA topics of different hurricanes. This way we can reduce the latency even more, while requiring fewer human resources.

Using the social media text data collected from private interviews from Hurricane Harvey previously detailed, we first preprocessed the data by removing stop words and performing lemmatization on the raw dataset. The text was then fed into our LDA model which we iteratively tuned to derive clusters that we then manually identified as helpful signal-oriented topics to help further determine whether the post can be considered signal or noise. We then repeated the process for the other Twitter datasets (detailed in Table 1), applying the pipeline illustrated in Figure 1. The topics generated between the different hurricanes were then compared using Hellinger distance, which is a distance function to measure the similarity of two distributions, and has been successfully used with LDA analyses in the past [5], The goal of our methodological approach was to find correlations between topic clusters generated from several hurricanes over a long duration to test whether our LDA-derived model can successfully identify key words that help classify signal tweets across multiple hurricanes. This could provide exemplar topics that can be used real-time during future disasters.

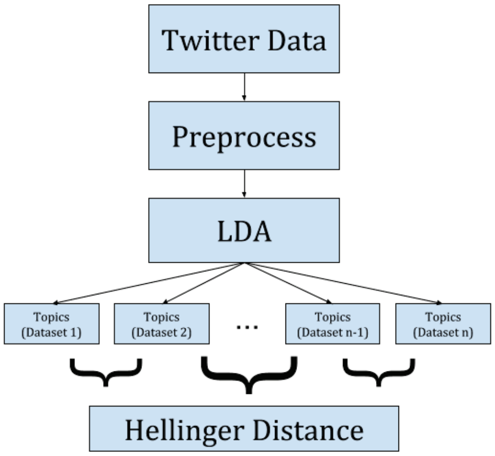

Figure 1. Data/Analysis Pipeline

3.3 Topic Modeling

We hypothesized that deriving clusters from tweets pertaining to a range of hurricanes over time using LDA would have some correlation between each hurricane's derived clusters. We also posited that the clusters between signal and noise tweets are distinct enough to distinguish. We evaluated whether LDA-derived topics from more recent hurricane events could have similar clusters as older events. Specifically, our LDA model returned five topic clusters per event, allowing us to use them to compare clusters and ultimately evaluate whether tweet-derived topics have some similarity with our ground truth interview-derived dataset.

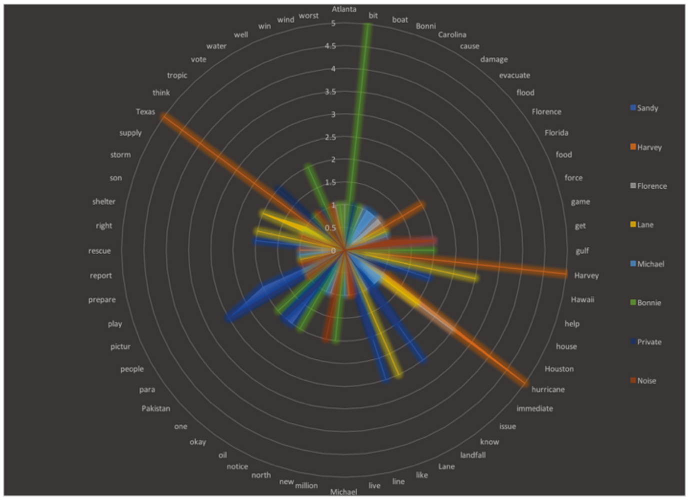

Figure 2. Top terms generated by LDA - all datasets

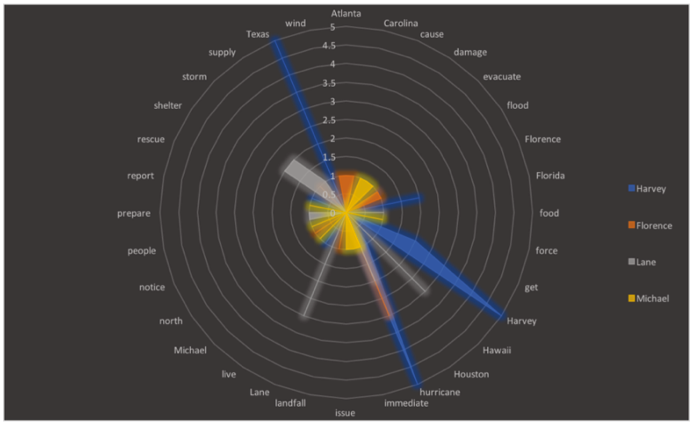

Figure 3. Top LDA terms for all 2017-18 Tweet data

3.3.1 LDA Implementation and Results. Topics often signify variable relationships that link words in a document's vocabulary based on their frequency of appearance in documents. A document may contain many topics, and these topics help identify themes throughout a collection of documents that in turn helps to label the corpus of data. We compared the themes we uncovered to test if the "hurricane" theme is prominently identified in each cluster from hurricane tweets. For our use of LDA, we employed Gensim, a library that uses Bayes approximations and Gibbs sampling to infer the distributions from observed documents [16]. For each hurricane event, we built a LDA model encompassing all the Twitter data, as well as qualitative interview-derived data from Hurricane Harvey. We then used the model to render five topics per hurricane event.

4 RESULTS

4.1 Topic Modeling Results

Figures 2-4 show the results of the LDA clusters over 6 hurricane events, the qualitative data, and one noise dataset. Figure 4 illustrates the top LDA-derived terms.

Figure 2 visualizes the top words of the LDA topic clusters for each storm, and Figure 3 does this for just the more recent 2017-2018 events. As Figures 3 and 4 illustrate, there is some overlap that appears in the top words for inter-event topic clusters, though the overlap is not always particularly identifiable.

Some of the terms derived from our work include the name of the hurricane or location (e.g., Texas and Houston for Hurricane Harvey). These terms are event specific and are not intended to be useful from a longitudinal perspective, but hurricane names and locations are important and recurrent parts of any Twitter-derived hurricane dataset. Moreover, persistent terms across the events as a whole (e.g. “storm”, “flood”, “hurricane”, “people”, and “get”) might not seem significant when considered individually, but when terms such as these are considered together, they have value. Specifically, our findings could also be used to create more targeted, Boolean searches (with inclusion and exclusion of terms) or be used to build classifiers that could produce more relevant or focused datasets than searching by hurricane event-specific hashtags, which is the current standard convention. Additionally, synonyms of persistent terms can be used for better real-time data collection during hurricane events.

4.2 Hellinger Distance

We used Hellinger distance to compute the distances between each LDA model. Hellinger distance quantifies the similarity between probability distributions and is the probabilistic analog of Euclidean distance [3]. For two probability distributions P and Q, Hellinger distance is defined as:

$$h(P,Q) = \frac{1}{\sqrt{2}} \cdot \|\sqrt{P} - \sqrt{Q}\|_2$$

The lower the output of Hellinger distance, the less difference there is between two hurricane events in terms of tweet text topics. The equation above shows how we take the difference between two distributions. Hence, more similar distributions result in a shorter absolute difference, which, in turn, results in shorter Hellinger distances. This distance is a relative measure. In other words, there is no threshold distance between two distributions that guides us to label either distribution as signal/noise.

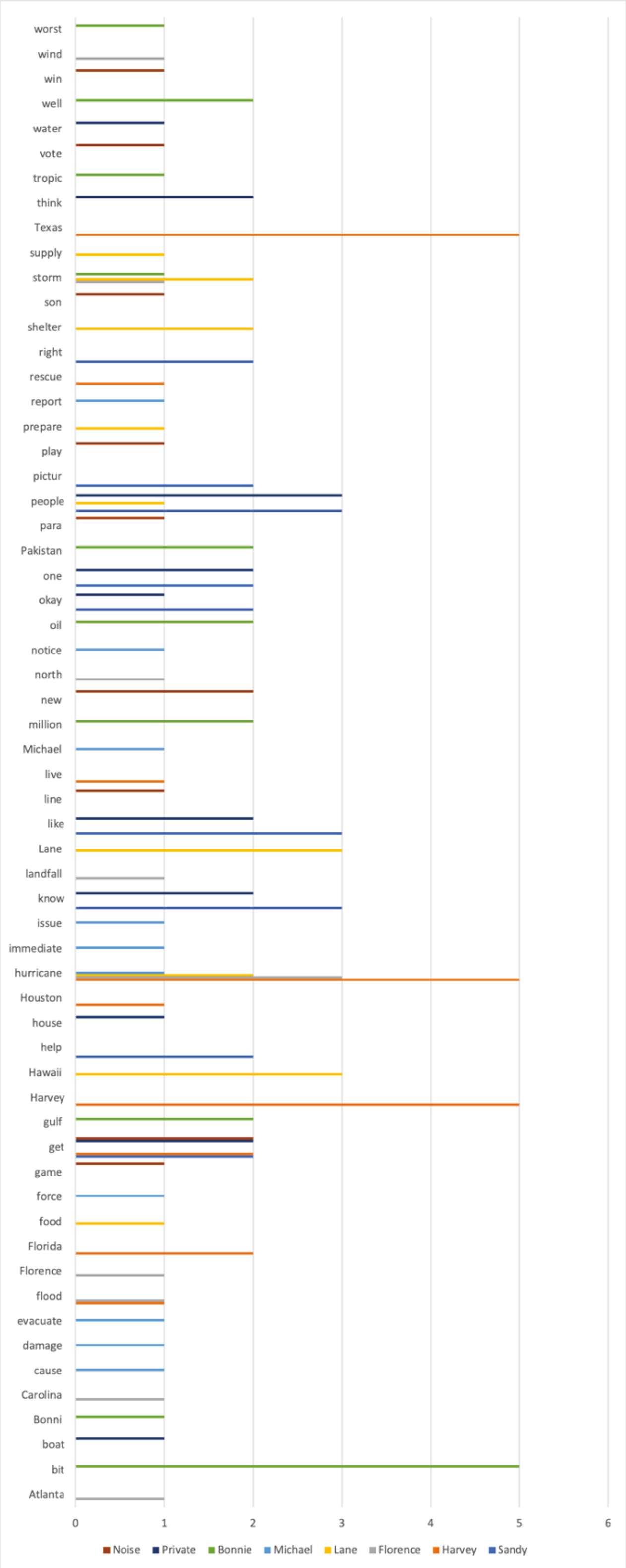

Figure 4. Top words from LDA among all datasets illustrated by frequency out of 40 (5 topics per dataset)

Shorter distances mean a more similar distribution of terms. Additionally, we are not using Hellinger distance to distinguish whether something is noise or signal, but rather to identify if there is some level of distance between signal and noise. Ultimately, our method is designed to compare each pair-wise distance with each other and analyze which distances are shorter than others

4.3 Hellinger Distance-based Results

The results of measuring the similarity between different hurricane clusters is shown in Figure 5. Each hurricane is represented as a circle proportional to the size of its Twitter dataset. The arrows in the Figure represent a hurricane's LDA clusters' closest distance to another hurricane's clusters. The solid lines represent closer distances and dotted lines are longer distances. The red circle represents the random noise data collected from Twitter and the small green circle on the right of the chart represents the qualitative, "ground truth" Harvey-derived signal data.

From the Hellinger distance results, all the hurricanes' closest distances are closer to other hurricanes or the signal cluster, with the exception of Hurricane Michael, which is closest to noise. Hurricane Harvey and Sandy have the most hurricanes in the closest proximity (along with the noise cluster for Hurricane Sandy). In the case of Hurricane Sandy, this may suggest that the Hurricane Sandy data set was diverse and scattered in content and therefore was close to many topics, regardless of whether the clusters were signal or noise

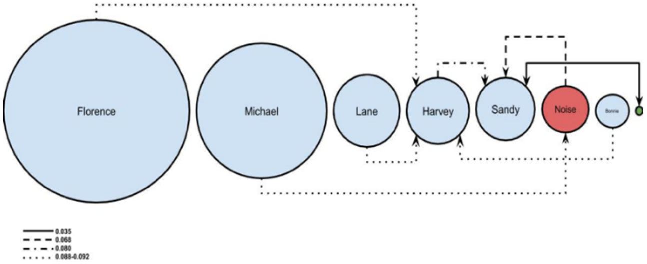

Figure 5. Hellinger distance between LDA models (The smallest circle at the far right signifies the Private (Harvey/'Signal') dataset; second to last circle is Bonnie)

Overall, the process of measuring the distance between LDA models for both hurricane tweets and noise data did not result in as strong of a distinction between signal versus noise data as hypothesized. Even though the majority of the hurricanes are closest to other hurricanes or the signal data, the distance between the noise cluster to a majority of the hurricane clusters is comparable to the distance to other hurricanes. For example, despite the fact that Hurricane Lane's clusters came closest to Hurricane Harvey's clusters, the second closest cluster in the output was the noise cluster.

The next closest hurricane to Lane was Hurricane Sandy, at 0.001 units further than the noise cluster. This suggests that although similar clusters exist among hurricanes, this relationship is inconsistent between multiple hurricanes. Therefore, LDA has difficulty differentiating the hurricane tweets as either signal or noise due to the lack of a strong similarity between topics across all the hurricane clusters.

5 LIMITATIONS AND FUTURE WORK

5.1 Limitations

Using keywords to extract Twitter data is convenient, but, many hurricane names such as "Michael", "Harvey" and "Florence" are also names of people or animals, which resulted in the inclusion of data that caused the hurricane cluster to present a greater number of topics similar to those within the noise dataset. This echoes issues found by others in collecting data based on keywords during crisis events, namely the inclusion of irrelevant tweets at high rates [18]. Additionally, though our qualitative interview-derived data was guaranteed to be ground truthed, long-length interviews are, of course, very different to the text found in brief tweets. These data were obtained through transcribed versions of verbal interviews and therefore contained far more verbose language compared to tweets. This difference in media for the signal dataset affected the performance of the LDA.

Furthermore, tweets are challenging to study in a comparative context with other media as they often contain higher levels of slang, URLs, language diversity, and other features compared with generic text or other social media posts.

5.2 Future Work

Future research can build on this work by improving on our process to better identify potential relationships between hurricane clusters. Further experimentation on the number of clusters to use for LDA and more intensive tuning of the LDA parameters could aid in finding similarities among clusters. Experimentation on metrics of measuring distances, such as Kullback-Leibler divergence, could also be used to calculate the distance between topic clusters and could potentially give a more holistic analysis of the cluster distances and relationships. Moreover, future work can use the terms we derived to build a tweet classifier and expand the application of our work. Of immediate value, future work can move away from collecting data on Twitter just by hurricane-related hashtags to methods where our identified terms are searched for (as a Boolean string) in lieu or in combination with relevant hashtags.

6 CONCLUSIONS

In this study, we investigate the effectiveness of an unsupervised machine learning model to classify disaster-related tweets either as signal (relevant content) or noise (less or irrelevant content) through a unique method of using Twitter data gathered from six different hurricanes alongside qualitative interviews conducted in the immediate aftermath of Hurricane Harvey. We find there is some level of stability across hurricane-related tweet topics over time. We used latent Dirichlet allocation (LDA) to derive topics, and using Hellinger distance as a metric found that while there is a detectable connection among hurricane topics, much further development is needed to operationalize this workflow “in the wild”. Though our findings could benefit from future work, this type of comparative work studying social media in crises is critically important and unique despite these limitations.

By comparing multiple storm/hurricane events over a longer time frame, our study makes a case that there are patterns between storm events on Twitter. A major implication of our work is that there is some persistence of terms in public social media corpora surrounding hurricane events. Therefore, generic training sets (at least in the context of US-based storms) could be used to reduce or even cut out the need for a human in the loop during the training of classifiers. Ultimately, we believe that our findings could not only be used to help develop real-time models to be used with future incoming hurricane data, but also break ground in a mixed methods approach which fuses computational work, qualitative interview data, and API-derived Twitter data.

ACKNOWLEDGMENTS

The authors would like to thank Koustav Rudra of the Indian Institute of Technology Kharagpur for assistance in extracting the Spritzer Harvey data from our Twitter data repository. We also thank interview respondents who shared their time and data with our team. This work was supported by a grant from the National Science Foundation [award # 1760453] RAPID: The Changing Nature of “Calls” for Help with Hurricane Harvey: 9-1-1 and Social Media. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] Alam, F., Ofli, F. and Imran, M. A Twitter Tale of Three Hurricanes: Harvey, Irma, and Maria. In *Proceedings of the 15th International ISCRAM Conference* (Rochester, NY, May 20-23, 2018). International Association for Information Systems for Crisis Response and Management (ISCRAM).

[2] Ashktorab, Z., Brown, C., Nandi, M. and Culotta, A. Tweedr: Mining twitter to inform disaster response. In *Proceedings of the 11th International ISCRAM Conference* (University Park, Pennsylvania, USA, May, 2014). International Association for Information Systems for Crisis Response and Management (ISCRAM).

[3] Beran, R. Minimum Hellinger Distance Estimates for Parametric Models. *The Annals of Statistics*, 5, 3 (1977), 445-463.

[4] Castillo, C. *Big crisis data: social media in disasters and time-critical situations.* Cambridge University Press, 2016.

[5] Dai, A. M., Olah, C. and Le, Q. V. Document embedding with paragraph vectors. *arXiv preprint arXiv:1507.07998* (2015).

[6] Griggs, B. *Harvey's devastating impact by the numbers. CNN*, 2017.

[7] Gross, A. and Murthy, D. Modeling virtual organizations with Latent Dirichlet Allocation: A case for natural language processing. *Neural networks*, 58 (2014), 38-49.

[8] Gruzd, A. Netlytic: Software for automated text and social network analysis. *Computer software, Available from:* http://netlytic.org/ *(accessed 25 May 2017)* (2016).

[9] Hsieh, C.-K., Alquaddoomi, F., Okeke, F., Pollak, J. P., Gunasekara, L. and Estrin, D. Small Data: Applications and Architecture. In *Proceedings of the Fourth International Conference on Big Data, Small Data, Linked Data and Open Data* (Athens, Greece, April 22,-26, 2018).

[10] Imran, M., Castillo, C., Lucas, J., Meier, P. and Vieweg, S. AIDR: Artificial Intelligence for Disaster Response. In *Proceedings of the Proceedings of the 23rd International Conference on World Wide Web* (Seoul, Korea, 2014). ACM.

[11] Kireyev, K., Palen, L. and Anderson, K. Applications of topics models to analysis of disaster-related twitter data. In *Proceedings of the NIPS Workshop on Applications for Topic Models: Text and Beyond* (Canada: Whistler, December, 2009).

[12] Li, H., Guevara, N., Herndon, N., Caragea, D., Neppalli, K., Caragea, C., Squicciarini, A. C. and Tapia, A. H. Twitter Mining for Disaster Response: A Domain Adaptation Approach. In *Proceedings of the 12th International ISCRAM Conference* (Kristiansand, Norway, May 24-27, 2015). International Association for Information Systems for Crisis Response and Management (ISCRAM).

[13] Liu, Z., Huang, W., Zheng, Y. and Sun, M. Automatic keyphrase extraction via topic decomposition. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing* (Cambridge, Massachusetts, 2010). Association for Computational Linguistics.

[14] Mehrotra, R., Sanner, S., Buntine, W. and Xie, L. Improving LDA Topic Models for Microblogs via Tweet Pooling and Automatic Labeling. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval* (Dublin, Ireland, July 28–August 1, 2013). ACM.

[15] O'Neal, A., Rodgers, B., Segler, J., Murthy, D., Lakuduva, N., Johnson, M. and Stephens, K. K. Training an Emergency-Response Image Classifier on Signal Data. In *Proceedings of the 17th IEEE International Conference on Machine Learning and Applications* (Orlando, FL, December 17-20, 2018). IEEE.

[16] Rehurek, R. and Sojka, P. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* (Valletta, Malta, May 17-23, 2010). University of Malta.

[17] Rhodan, M. *Hurricane Harvey: The U.S.'s First Social Media Storm. Time*, 2017.

[18] Saleem, H. M., Xu, Y. and Ruths, D. Novel Situational Information in Mass Emergencies: What does Twitter Provide? *Procedia Engineering*, 78 (2014), 155-164.

[19] Seroussi, Y., Zukerman, I. and Bohnert, F. Authorship attribution with latent Dirichlet allocation. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning* (Portland, Oregon, June 23-24, 2011). Association for Computational Linguistics.

[20] Smith, W. R., Stephens, K. K., Robertson, B. W., Li. J., & Murthy, D. Social media in citizen-led disaster response: Rescuer roles, coordination challenges, and untapped potential. In *Proceedings of the 15th International ISCRAM Conference* (Rochester, NY, May 20-23, 2018). International Association for Information Systems for Crisis Response and Management (ISCRAM).

[21] Tong, Z. and Zhang, H. A Text Mining Research Based on LDA Topic Modelling. In *Proceedings of the Sixth International Conference on*

Computer Science, Engineering and Information Technology (CCSEIT) (Vienna, Austria, May 21-22, 2016).
[22] Vieweg, S. Twitter communications in mass emergency: contributions to situational awareness. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work* (Seattle, Washington, February 11 - 15, 2012). ACM.