2019 IEEE International Conference on Smart Computing (SMARTCOMP)

Identifying the Context of Hurricane Posts on Twitter using Wavelet Features

1st Amrita Anam
University of Maryland
Baltimore County
Baltimore, USA
amrita1@umbc.edu

2nd Aryya Gangopadhyay
University of Maryland
Baltimore County
Baltimore, USA
gangopad@umbc.edu

3rd Nirmalya Roy
University of Maryland
Baltimore County
Baltimore, USA
nroy@umbc.edu

***Abstract*—With the increase of natural disasters all over the world, we are in crucial need of innovative solutions with inexpensive implementations to assist the emergency response systems. Information collected through conventional sources (e.g., incident reports, 911 calls, physical volunteers, etc.) are proving to be insufficient [1]. Responsible organizations are now leaning towards research grounds that explore digital human connectivity and freely available sources of information. U.S. Geological Survey and Federal Emergency Management Agency (FEMA) introduced Critical Lifeline (CLL)s which identifies the most significant areas that require immediate attention in case of natural disasters. These organizations applied crowdsourcing by connecting digital volunteer networks to collect data on the critical lifelines from data sources including social media [3], [4], [5]. In the past couple of years, during some of the deadly hurricanes (e.g., Harvey, IRMA, Maria, Michael, Florence, etc.), people took on different social media platforms like never seen before, in search of help for rescue, shelter, and relief. Their posts reflect crisis updates and their real-time observations on the devastation that they witness. In this paper, we propose a methodology to build and analyze time-frequency features of words on social media to assist the volunteer networks in identifying the context before, during and after a natural disaster and distinguishing contexts connected to the critical lifelines. We employ Continuous Wavelet Transform to help create word features and propose two ways to reduce the dimensions which we use to create word clusters to identify themes of conversations associated with stages of a disaster and these lifelines. We compare two different methodologies of wavelet features and word clusters both qualitatively and quantitatively, to show that wavelet features can identify and separate context without using semantic information as inputs.**

***Index Terms*—Hurricanes, Disaster Response, Wavelet Analysis, Clustering, Social Media**

I. Introduction

In the past decade, social media has proved to be an enormous source of real-time information during a crisis period. From SOS messages to alerts and guidelines from authorities, social media has been an easy and fast platform to broadcast valuable information in the times of need when the alternative to the traditional means [1]. Although social media text has been a popular platform of research, crisis analysis on social media still needs efficient and automated approaches that are alternative to the traditional text mining [6], [7]. The hurricanes past year brought many support groups together through social media. Volunteer networks were built to draw attention

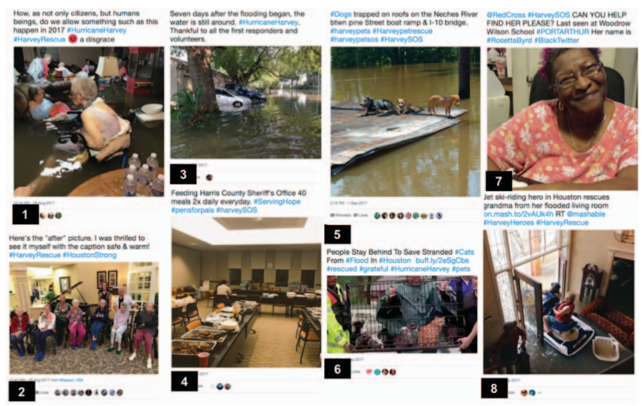


Figure 1: Tweets during hurricanes: 1-2 The before and after the status of a group of elderly people drowning in flood water who were rescued and put into a safe shelter, 3 the water level in a residential area after seven days after the flood hit Texas. 4 a relief centers to feed the flood victims two times daily, 5-6 dogs trapped in the water trying to float on remains and the captions gives the location and pets being rescued by a team, 7 Information about a missing woman is posted along with her photo, 8 a volunteer is rescuing an elderly woman using a jet ski.

for rescue request, finding safe shelters and relief inquiries [8]. People in need of immediate help wrote posts for themselves, their friends, families, and neighbors. Numerous social media groups were created to coordinate the search, rescue and relief process. Snapshots of a collection of tweets are shown in fig. 1 that were made during hurricane Harvey. The tweet snaps convey both the sides of the victims the and helpers. There are SOS messages including missing reports, search and rescue requests, flooded landscape, humans and pets stranded in water; as well as relief management and rescue operations handled by the government and non-government authorities and volunteers. The tweet snaps show that social media has been an efficient and effective communication platform for the exchange of real-time crisis information. FEMA has defined key aspects of information collection during natural disasters that are significant for efficient response. These aspects, termed

Table I: Example of tweets matching lifelines

Lifeline	Sub-topic	Tweet Example
Safety and Security	Evacuation	Heard one woman say this morning she decided not to evacuate Florida despite Hurricane Michael. She said she wasn't one to back away from a fight. Wondering if she realizes she can't exactly pull out an AR-15 and scare the storm into submission.
	Fatality	Thousands of people missing after Hurricane Michael; death toll at 17
	Rescue	A Florida couple was trapped by a collapsing wall while trying to escape Hurricane Michael. Neighbors sprang into action to rescue them
Food, water, and shelter		A few of our Noles loaded 54 trucks with 40,000 bottles of water and 125,00 cases of food with the local food bank to send to areas affected by Hurricane Michael.
		After Hurricane Michael, food is spoiling quickly, so residents,clamor for ice. Bottled water is a must. And gasoline is becoming,increasingly scarce.
Health and Medical	Hospitals	baymedicalctr in Panama City, FL had over 200 patients when Hurricane Michael made landfall. All have been transported to other hospitals. It is on generator power but ER is still open. Many of the staff working lost their homes in storm.
	Pharmacy	A mobile pharmacy is located in the Marianna Walmart parking lot,and will provide prescriptions, immunizations and general resources for those,impacted by Hurricane Michael. The pharmacy will temporarily operate 9 a.m.,to 7 p.m. daily.
Energy	Power	After riding Hurricane Michael out in a Panama City motel as the,walls shook & cracked, I had no power or cell service for days. I spent,two sleepless nights in a shelter & a rental car. I even injured my leg.
	Fuel	This island is now in ruins after Hurricane Michael More than 230,000 are still without power. Drivers are lining up for hours to get fuel. And residents in the hardest-hit areas are relying on airdropped food and water to survive. And you give them a website?
Transportation	Broken bridge	Houlihan Bridge closed to maritime traffic before and during Hurricane Michael
	Broken road	The road that leads into Cape San Blas Florida, lays broken in the sun on Saturday, October 13, three days after Hurricane Michael roared ashore.
Communications		Drones are helping provide cell signal in the panhandle. Network recovery operations ongoing in aftermath of Hurricane Michael

as Critical Lifelines (CLL) are: *i)* safety and security, *ii)* food, water, and shelter, *iii)* health and medical, *iv)* energy (power and fuel), *v)* communications, *vi)* traffic and transportation, and *vii)* hazardous waste [2]. Information that falls under these lifelines is crucial for emergency response. Some of these lifelines have their own sub-topics which are more specific to comprehend. These lifelines can be reflected in social media posts during a disaster. The lifelines along with their sub-topics and example tweets are presented in table. I. These critical lifelines have well-defined definitions and in some cases, subtopics. For example, the first CLL is 'safety and security', by definition this CLL captures information regarding evacuation, search and rescue processes, missing people, etc. The fifth CLL 'power and gas station' tracks information on power stations, availability electricity and gas stations. These sub-topics do not necessarily co-occur, for example, security issues and missing people are not likely to be part of the same tweet. We analyze these lifelines by projecting them on the wavelet features. Currently, the process of identifying information from various data sources, that are aligned with these lifelines, is done manually.

In this paper, we propose a method to identify and distinguish the context from tweets that are commonly used to reflect different stages and aspects during a natural disaster and compare them with the lifelines. We propose an unsupervised, multiresolution approach to extract *i)* wavelet features in reduced dimensions, and used them to explore *ii)* context distribution from before, during, and after the hurricane, and *iii)* context around the CLL. We show that our method presents a feature space to represent transient social media posts, which does not require large coherent context. The method can be converted into real-time and can sustain traditional text operations (e.g., semantic similarity). The end-to-end pipeline of the proposed framework is illustrated in fig. 2. The three key sections of the pipeline are a representing deconstructed tweets as temporal signals, creating wavelet features in reduced dimensions an applying the features in identifying the context around different stages and different CLL(s).

II. Data

We used the standard search API of Twitter to collect tweets related to hurricane Michael. The standard search API provides a collection of tweets that are most relevant to the query used. It does not provide the full list of users or tweets but a sample that fits the search filters. The mandatory filter is query and the rest are optional. The filters that we used in the search request to collect relevant tweets include *i)* query terms('q') ('hurricane + michael' and 'florida + hurricane'), *ii)* language('lang') which was set to English for convenience, *iii)* start date('since'), *iv)* end date('until') and *v)* number of tweets to retrieve ('count'). These filters return a relevant sample of the total tweets that match the search criteria. In this paper, we analyze tweets collected between October 10th and October 16th. The total number of tweets collected in these seven days is 95,406.

III. Related Work

Social media, especially Twitter is not a stranger to the researchers when it comes to analyzing those who are affected by disasters and those who are responsible for the management. One of the problems in this area is getting annotated data. In most cases, the authors manually annotate the tweets or use unsupervised learning to identify groups within the

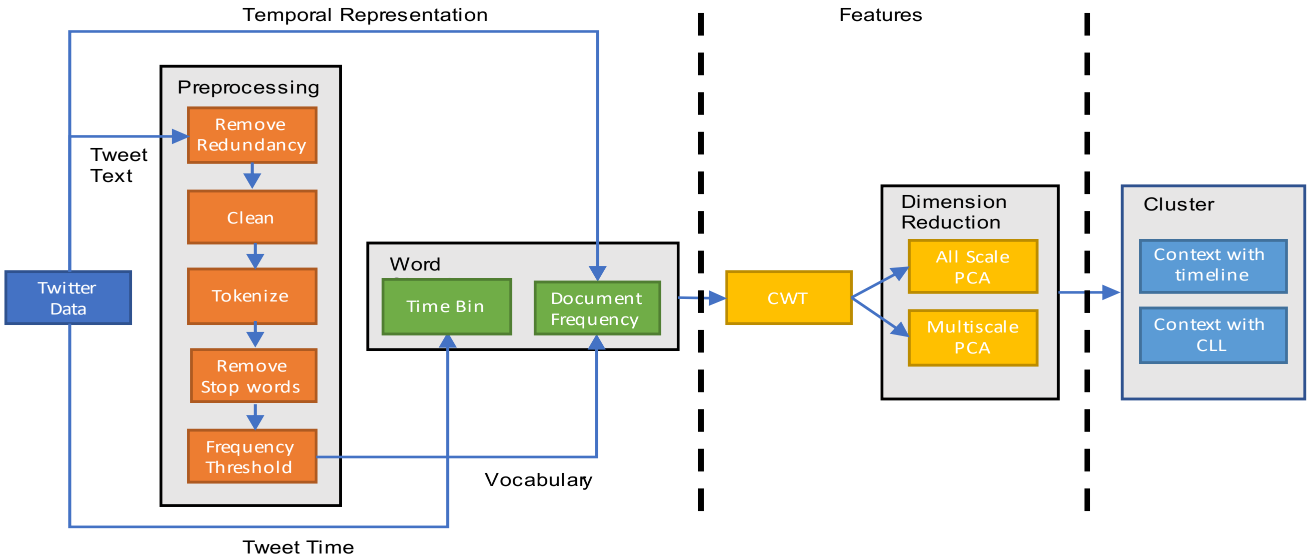


Figure 2: End-to-end pipeline for analyzing context distribution

data. In [9], the authors manually annotated a small sample (0.7%) of the twitter messages on hurricane Irene to classify them as concerned or unconcerned with MaxEnt Decision Tree, Naive Bayes algorithms. Tweets related to Hurricane Sandy was the most popular hurricane data to be researched with supervised learning [6], [10]–[13]. In [10], the authors analyzed three different disasters (e.g., hurricane Sandy and Moore, Oklahoma Tornado and Boston marathon bombing) to show that they differ in terms of foreknowledge, duration, severity, and news media coverage. A Bayesian approach was used to classify tweets during Hurricane Sandy with binary tags ”informational” and ”conversational” in [6]. The authors compared Naïve Bayes with a bag-of-words approach and validated with accuracy and F1 score. Credibility has always been a challenge for social media data. In [14], the authors identified credible users by applying classification with features extracted from user profile metadata. In the question of credibility, disaster-related images on Twitter have also been a popular research topic. The authors in [13] characterized identified patterns in temporal, social reputation and influence for the spread of fake images during hurricane Sandy. An analysis of the use of Twitter in local municipalities was explored with mixed-methods research effort in [12]. Another relevant research is presented in [15] where the authors created an unsupervised learning tool, CorEx that can extract and classify disaster-related words from Tweets during known emergencies. Other than hurricanes, disaster-related messages from floods [7], landfalls [16], earthquakes, wildfires were all explored by research communities. Features used in both supervised and unsupervised learning include the time of the tweet, hashtags, word frequency, information score, word embeddings, etc. A very similar research perspective to our is presented in [11] where the authors analyzed the content distribution of tweets during a crisis (e.g., hurricane Sandy) with hashtag trends and frequency, total impression, re-tweet rate, etc.

In this paper, we propose a methodology to create features using the wavelet transform text data. The method is an extension of the work presented in [17] where we introduce Continuous Wavelet Transform as a way to structure tweets. The use of wavelet analysis on Twitter data was also seen in [18] and [19] for event detection using Discrete Wavelet Transform where the authors reflect on the correlation of topics being reflected on different scales of a wavelet.

IV. Signal Representation

We build a vocabulary of English words extracted from all the tweets in the input sample. First, we tokenize the tweets into a list of unigrams after removing duplicate words, stop words, numbers, and symbols. The vocabulary is constructed by unigrams that are validated with English language corpus in Python's Natural Language Toolkit (NLTK). Each post is tokenized into a list of words (i.e., context words) that are in the vocabulary. We select words with frequency pruning and

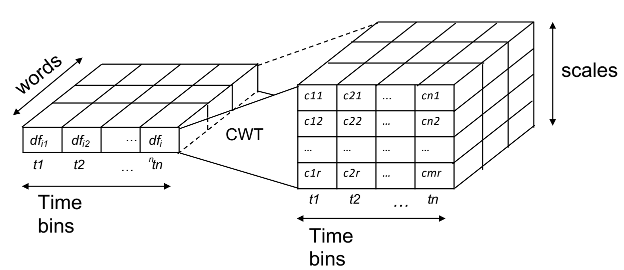


Figure 3: Data conversion structure by CWT

take the top 75% of the words. These context words are then represented by numeric vectors binned by a fixed duration of time. In this paper, we choose one hour as the bin duration. The fixed duration also represents sampling. For amplitude, we used document frequency which is defined by the number of documents (tweets) containing a word in a bin. We bin the context words in the vocabulary by a fixed duration. The process of converting tweets into tokens, binned by a fixed time duration is explained in our previous work [17].

V. Wavelet Features

After creating word signals from the input sample we extract features with CWT. These features are then reduced with two multiresolution dimension reduction techniques.

A. Continuous Wavelets Transform

We apply Continuous Wavelets Transform (CWT) on the temporal representations of words with Morlet wavelet [20], [21] as mother wavelet and pre-defined number of scales. Each word signal is converted into a matrix of wavelet coefficients as illustrated in fig. 3. The time bins of the word signal going into the CWT are represented by $t_1, t_2, ..., t_n$ and the document frequencies for a word signal s_i are represented by $df_1, df_2, ..., df_n$. CWT transforms the signal to a matrix C of size n x r where r represents the number of scales r. Wavelet coefficient $c_jk1(i)$ is the coefficient of i^{th} word in j^th bin st k^th scale.

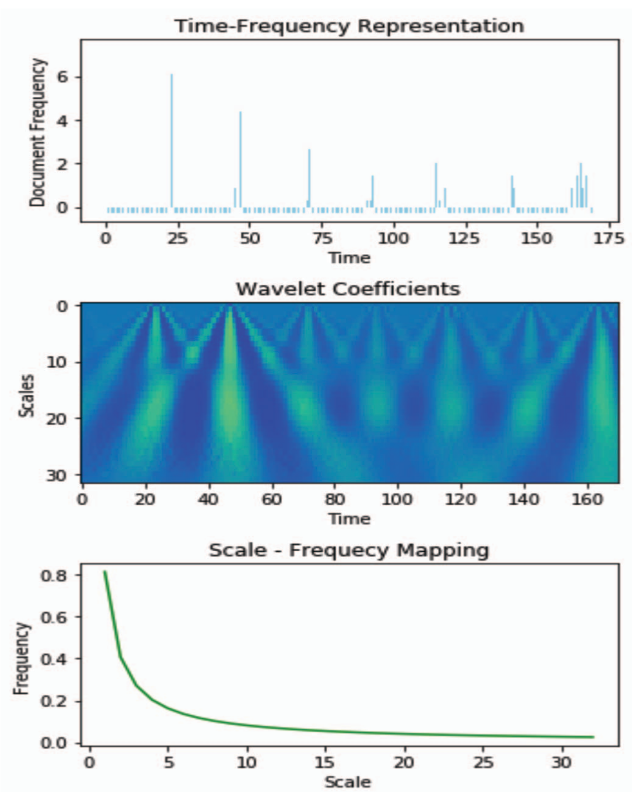


Figure 4: The signal representation, wavelet coefficients and mapping of pre-defined scales with obtained frequencies of the signals

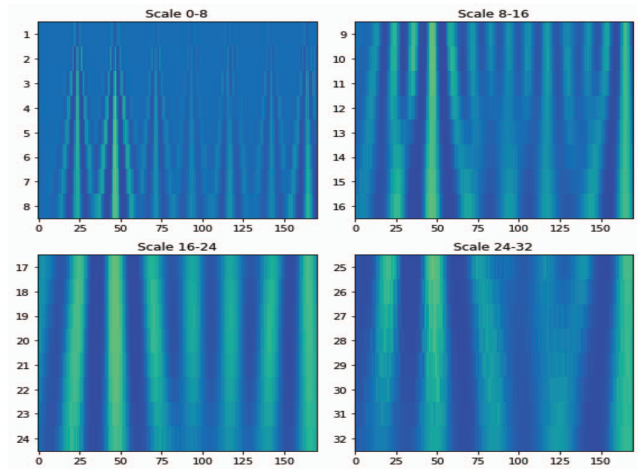


Figure 5: Four ranges of scales

$$C(u_i) = \begin{bmatrix} c_{11}(i) & c_{12}(i) & c_{13}(i) & \dots & c_{1n}(i) \\ c_{21}(i) & c_{22}(i) & c_{23}(i) & \dots & c_{2n}(i) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_{r1}(i) & c_{r2}(i) & c_{r3}(i) & \dots & c_{rn}(i) \end{bmatrix}$$

An example of the temporal representation and the scalogram of the wavelet transform of the word 'help' is illustrated in fig. 4. The x-axis shows the time bins and the y-axis shows the scales. Frequency and scale have inverse properties where the low-frequency components are obtained by the higher scales and high-frequency components are obtained by the lower scales. The characteristics of the signals are prominent and distinguishable for the high-frequency components and are fuzzy for the low-frequency components. Fig. 5 shows the scalograms of the same word signal in fig. 4 divided in four ranges of scales. The low scales (high-frequency components) demonstrates the strongest signals. In this paper, we consider wavelet coefficients that represent the first 32 scales for analysis.

B. Multiresolution Dimension Reduction

We applied two techniques of dimension reduction to reduce the coefficient matrices of word signals in order to create a feature set that can be used in further analysis. Vectorization and Principal Component Analysis (PCA) were used in both the methods for reducing the dimensions of the wavelet coefficients. The reduced principal components (PC)s are used in context analysis such as, clustering to find words that are similar in the time-frequency domain. We compare these methods of creating word features from wavelet coefficients by observing their performance in the cluster analysis. The captured variance of adding each principal component is shown in fig. 8. We compare these methods of creating word features

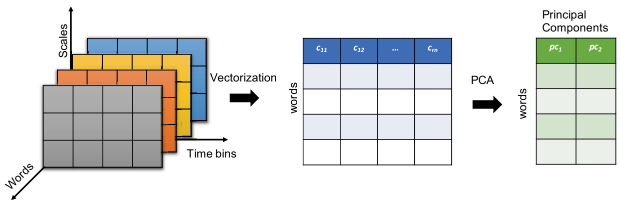


Figure 6: Data dimensions in all-scale PCA

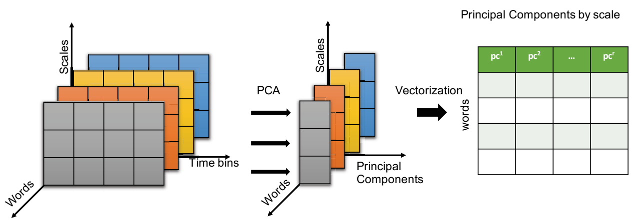


Figure 7: Data dimensions in multiscale PCA

from wavelet coefficients by observing their performance in clustering. The steps followed in the two methods are as follows:

1) All-Scale PCA: In this method, we first vectorize the wavelet coefficient matrix of each word with size $r \times n$, where r = the number of preset scales and n = the number of times bins. Then, we apply PCA on the vectors and select the top p, where $p \ll r \times n$. The top four principal components capture 95% of the total variance displayed in the top figure in 8. Fig. 6 illustrates the data dimensions in the pipeline.

2) Multiscale PCA: In this method, we create a principal component for every scale by apply PCA on all the features for each scale. Then, we take the top component for each scale. This method is used in [22] for statistically analyzing wavelet coefficients. The size of the new feature set would be $m \times r$ where m = the number of words and r = the number of preset scales. The variance captured by the increases with higher scales which means the low-frequency components capture more overall variance, displayed in the bottom figure in 8. The data dimension of this pipeline is given in fig. 7.

VI. Context Distribution around Stages with Clustering

In this experiment, we take three samples from before, during and after the hurricane hits the ground. We fuse the posts from all three samples and separate them time by adding a day in between. By applying K-means clustering algorithm, we compare the two features by observing their performances in building three identifiable word clusters from the fused tweet sample. Clustering highlight different themes of conversation on Twitter at different stages of the hurricane. K-Means clustering algorithm creates k clusters by computing distance between data and centroids through multiple iterations. K-Means initializes random co-ordinates as cluster centroids, separates clusters by comparing the distance between data points from centroids, calculates a loss function and updates the centroids for every iteration until the loss

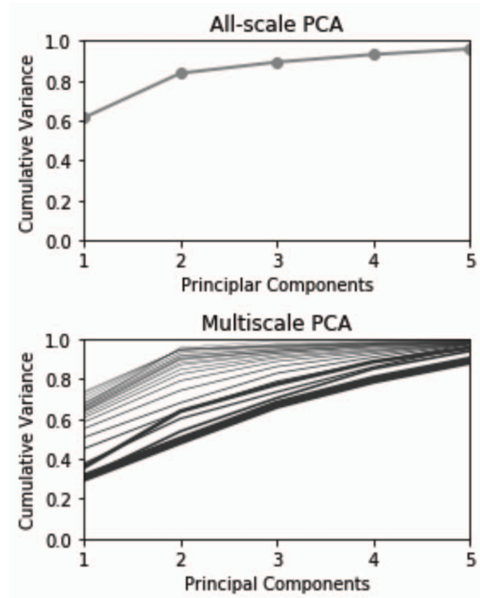


Figure 8: Captured variance by both methods of dimension reduction

function converges. We calculate the within-cluster sum of squared distances for $k = 1, 2, 3, ..., 10$ showed in fig. 9 to obtain the best k = number of clusters. In both the dimension reduction methods, the within-cluster sum of squared distances provides diminishing return after *k = 3*. We also observe that the within-cluster sum of squared distances for $k = 3$ using all-scale PCA is lower than that of multiscale PCA. To compare the performances of the features in correctly identifying the

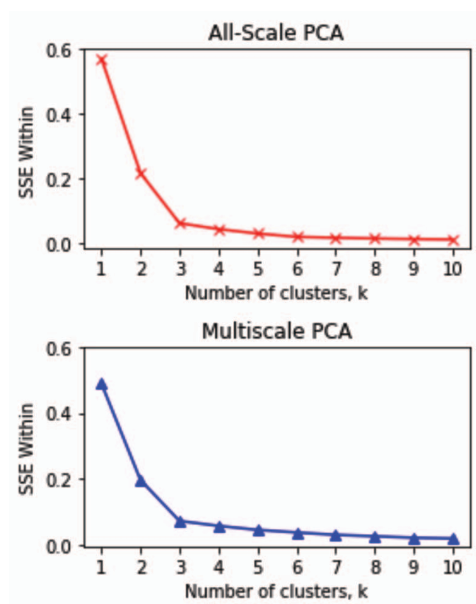


Figure 9: Sum of squared distances for different k

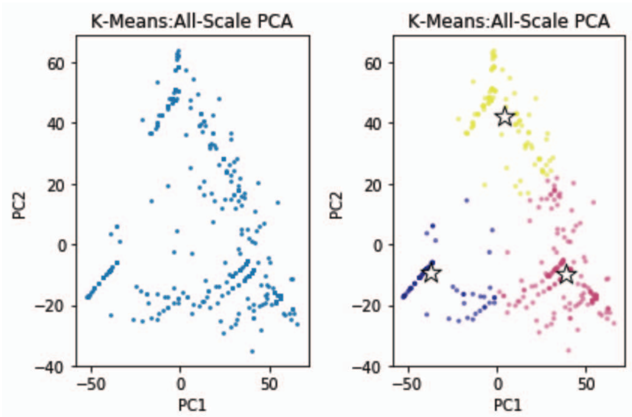


Figure 10: Word signals in the reduced space with all-scale PCA before (left) and after (right) clustering

initial samples, we labeled documents into our word cluster using the information bottleneck method [23].

$$Accuracy = \frac{\text{Number of correctly identified instances}}{\text{Number of total instances}}$$

The cluster separation for all-scale PCA features with K-Means is showed in In fig. 10 where the stars represent the centroids. For all-scale PCA, the x and y axes represent the

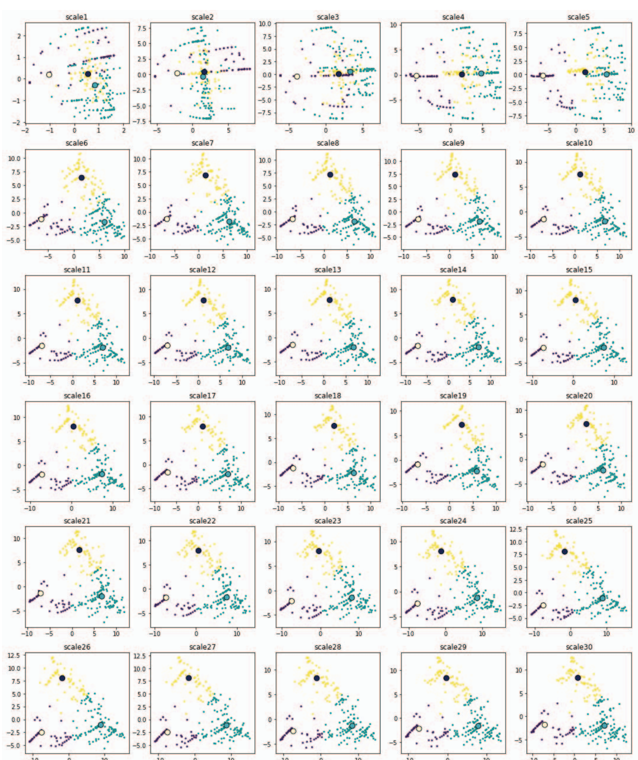


Figure 11: Data clusters for different scales with multiscale PCA

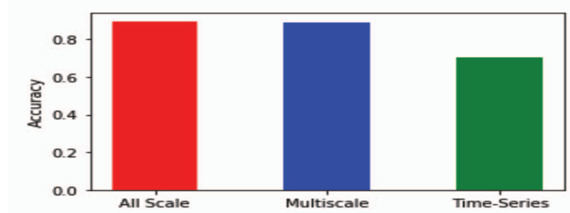


Figure 12: Accuracy of correctly labeled documents

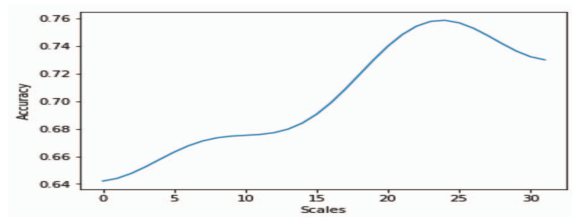


Figure 13: Accuracy of correctly labeled documents with scales using multiscale PCA

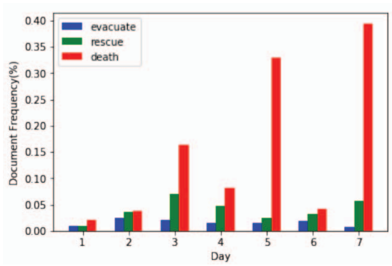
first and second principal components respectively. We can see in fig. 10 that K-means successfully separated the three clusters. Multiscale PCA reduces the data for every scale. The cluster representation is presented for every scale in 11. For multiscale PCA the x and y axes represent the first and second principal components respectively for scale. Cluster separation of the words for 30 scales is shown in fig. 11. We calculate accuracy to validate the document labeling with word clusters and compare it with the baseline PCA of the word signals (i.e., time-series PCA) displayed in fig. 12 with all-scale PCA having 89.5%, multiscale PCA having 88.7% and time-series PCA having 70.3% accuracy. The change of accuracy with scales is shown in fig. 12. Notice the accuracy is much higher between 20 to 25.

VII. Context Distribution around CLL(s) with Clustering

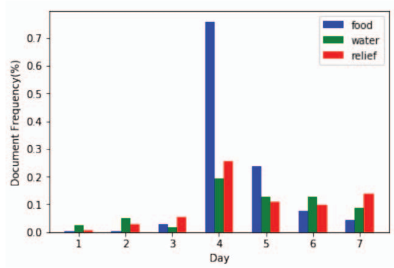
In this section, we compare our regular clusters with pre-defined centroids with selected words related to the Critical lifelines(CLL). CLL(s) are aspects of a disaster that are needed for efficient response. We implement a clustering experiment over seven days of hurricane Michael to find the closest topics to the CLL(s). We select candidate words that are best representatives of the CLL groups and are likely to co-occur. The candidate words are shown in table. II. The candidate words are selected from the definitions and different sub-

Table II: Candidate words for each CLL

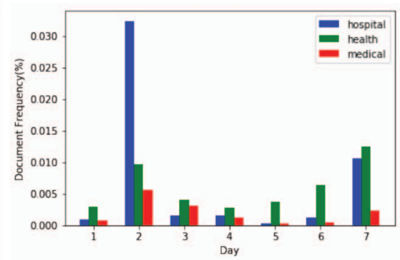
Critical Lifelines	**Candidate Words**
Safety and Security	'safety', 'evacuate', 'search, 'rescue'
Food, Water and Relief	'food', 'water','relief'
Health	'health', 'medical', 'hospital'
Power	'electricity', 'power', 'energy'
Communication	'phone', 'cell', 'network' '
Transportation	'traffic', 'road', 'flight'



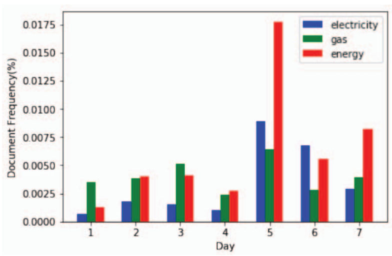
(a) Safety and Security



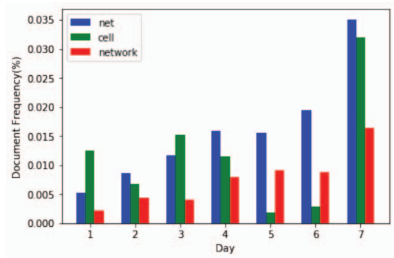
(b) Food, Water and Relief



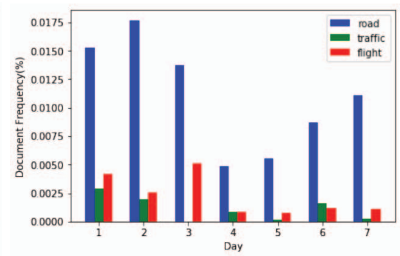
(c) Health



(d) Electricity



(e) Communication



(f) Transportation

Figure 14: Normalized frequencies of words reflecting Critical Lifelines

categories of the CLL(s). Fig. 14 illustrate the frequencies of some of the candidate words for the CLL(s) as a percentage of all the words by day. The bar charts reflect the dominance of the words in the first seven days after the hurricane hit the land. In fig. 14a, the frequency of the words ‘evacuate’ and ‘rescue’ increases in the middle and then decreases. The word ‘death’ dominates in this group and peaks on the seventh day. In fig. 14b and fig. 14c, the words ‘food’ which peaks on the fourth day, and ‘hospital’ which peaks on the second day, dominates over other words in their respective charts. In fig. 14d, the word ‘energy’ is the most frequent. In fig. 14e, all the words displayed peak on the seventh day. In fig. 14f, the word ‘road’ is the most dominant and the word ‘traffic’ is the least frequent.

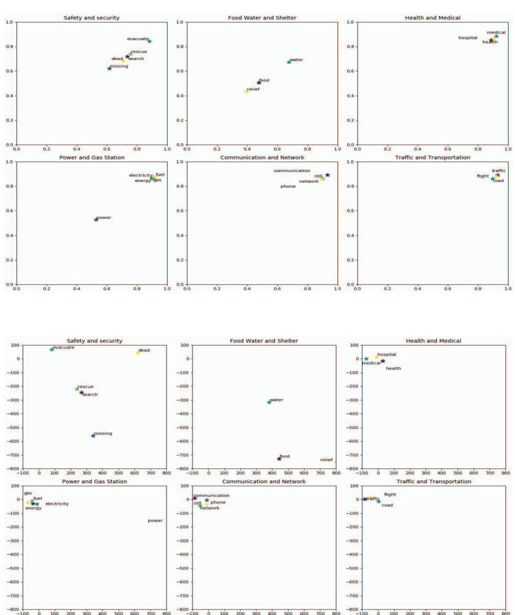


Figure 15: CLL words in all-scale PCA (top), and multiscale PCA (bottom)

Before applying clustering we represent the candidate words in the reduced feature space created by all-scale PCA and

Table III: Closest words to cluster centroids

assistance	sustained	big	footage	police	husband
group	landfall	law	apocalyptic	finds	unexpected
companies	stay	pieces	happened	surveys	copies
deserve	downgraded	remain	west	shocked	war
chairman	near	creates	bush	lady	octane
insurers	advisory	hopes	stock	technology	fighters
customer	politics	spoke	kid	visits	publicly
slow	radar	golf	towns	logs	targeted
number	ashore	email	journal	cats	hidden
climbed	bands	tests	cease	spells	gotten
homeowners	warn	spends	apple	unprepared	water
deaths	winds	golfing	yesterday	lawn	serving

Table IV: Closest words to CLL centroids

search	**food**	**health**	**electricity**	**phone**	**traffic**
rescue	gathered	assess	board	died	rejected
rubble	**children**	**officers**	efforts	counties	temporary
concern	answers	businesses	spent	available	**damage**
crews	government	talked	**days**	month	judge
teams	**drones**	**operating**	recovery	mission	definitely
survivors	struggling	things	inmates	meet	swept
idea	furious	**help**	**eastern**	**number**	sucks
article	**water**	office	story	media	comfort
homeless	provided	**thank**	helping	**students**	**bridge**
complete	looters	stealth	hand	department	waste
weekend	**victims**	losses	**residents**	mobile	shattered

Table V: Tweets representing words in CLL clusters

Tweets with matching CLL clusters
"On Tuesday, cadaver dogs found 2 bodies in the **rubble** of Mexico Beach, a captain of a FEMA Urban **Search** and **Rescue** Task Force,"
"**Search**-and-**rescue teams** rushed to reach communities that Hurricane Michael leveled, hoping to find **survivors** "
Attention all Home Depot employees our HD 0207 is having a **food** drive for the **victims** of Hurricane Michael.
Many residents said their **food** and **water** were running out. They were worried about the small **children** among them.
A big **thank** you to the UF **Health** ShandsCair Critical Care Transport Program teams and all first responders for **help**
We have a team of **officers** who are providing disaster behavioral **health** support for emergency **response**
Hundreds of thousands of **residents** are left without water or **electricity** for **days** in the wake of Hurricane Michael.
I survived hurricane Michael and after three and half **days** without **electricity**...
"So many organizations are helping after Hurricane Michael, even using **drones** to get **food** and phones to the people."
"**Traffic** Update: Residents of Bain **bridge**, Georgia experienced heavy **damage** from Hurricane , but they haven't let that put a damper on heir hopes"
Hurricane Michael caused 1.7 million **electricity** outages in the south-**eastern** US.
"Our shelter in Atlanta, GA is accepting **students** who are fleeing Hurricane Michael. The **phone number** and address are.."

multiscale PCA. The candidate words projected in the reduced space with all-scale PCA and multiscale PCA displayed in fig. 15. The interesting observation at this stage is that, the candidate words that are semantically connected are very closely placed (e.g., 'search' and 'rescue', 'health' and 'medical', etc.). Some words with strong ambiguity (e.g., 'power', 'communication', etc.) show a different characteristic. We examined the tweets and found that the word 'power' has been used in the sense of strength (e.g., 'political power', 'power of a community', etc.) and the word 'communication' has been used to convey verbal and physical interaction instead of the cellphone network.

We apply K-means clustering with pre-selected words that represent the CLL(s) centroids. The words that we chose to be centroids are 'search', 'food', 'health', 'electricity', 'phone' and 'traffic'. We compare the word clusters formed with randomized initialization with that of pre-defined CLL centroids. The words closest to the centroids of the cluster with the randomized initialization in affinity features are displayed in table. III. The words closest to the CLL centroids in each cluster are given in the table. IV. Notice that randomization is not able to pick up the CLL(s) as topics and hence we end up with clusters that are not very meaningful. We investigate the quality of the clusters by observing the word associations from actual. Some examples of tweets containing the words in the clusters with CLL centroids are shown in table. V. We observe that the clusters with the CLL centroids are much more meaningful in terms of word associations than the randomized clusters. Notice that, even though the features are created from time, frequency components, they reflect word co-occurrence and contextual similarity. We observe that words such as 'rescue' and 'teams' are placed in the cluster with centroid 'search'; 'water' is placed in the same cluster as 'food' which also contains words with strong association for the context, such as 'struggling', 'provided', 'looters'. Even the words that are not highlighted have semantic similarity, for example, 'mobile' is placed in the group with 'phone' as centroid, 'crews' is placed with 'teams', 'helping' is placed with 'hand'. We find that, even though the word clusters by randomized initialization show distinct themes of conversations, they are not necessarily good representative of the CLL(s). On the other hand, clusters created with CLL centroids are much more reflective of their individual CLL.

VIII. Conclusion

With the recent avalanche of natural disasters, we are in constant need of efficient and inexpensive solutions. During the hurricanes in the past couple of years, even with the developed disaster management systems, people struggled to convey their actual situations with the authorities. In their cry-for-help Twitter posts, they shared their helplessness publicly to get proper attention. The government's initiative to identify social media as a credible source of information is a massive motivation in this field of research. Unfortunately, the challenges of social media data are almost always resolved with human-in-the-loop solutions. Technologically advanced automated told and applications with social media data can assist the government and non-government authorities to broadcast their updates and identify disaster-related messages quickly and efficiently. Our work of analyzing tweets with signal processing tools like wavelets, open a different horizon in identifying context without semantic ontology. We show that wavelet features can reflect the transient nature of context in social media during a natural disaster. The cluster reflects the topic that co-exists in both time and frequency of their occurrence. The time-frequency features created through CWT show promising semantic coherence without using any linguistic inputs. The success of this approach will lie in adding efficient input from domain experts. With the help of domain knowledge, our method of wavelet-based features on social media can be used in multiple applications. A direct application of the process explained in this paper is tweet categorization using word clusters which we are currently working on. Another interesting application can be identifying the localization of specific topics in both time and frequency domain. These applications can be used to provide emergency responders with decision support tools from pre-labeled messages.

Acknowledgment

We thank Sophia B Liu, Innovatio Specialist at U.S. Geological Survey (USGS) for the information on FEMA and USGS. This work is partially supported by the NSF CNS grant 1640625 and NSF DGE grant 1515358

References

[1] "Social media provides critical information missed by fema," August 2018, [Online; 23 Aug. 2018]. [Online]. Available: https://phys.org/news/2018-08-social-media-critical-fema.html

[2] "Tropical storm florence lifelines." September 2018, [Online; 20 Sept. 2018, 7:00 AM EDT]. [Online]. Available: https://www.fema.gov/tropical-storm-florence-lifelines

[3] "Fema mobile app," December 2018, [Online; 11 Dec. 2018, 10:34 AM]. [Online]. Available: https://www.fema.gov/mobile-app.

[4] S. Roberts and T. Doyle, "Understanding crowdsourcing and volunteer engagement: Case studies for hurricanes, data processing, and floods," *Flood Damage Survey and Assessment: New Insights from Research and Practice*, vol. 228, p. 121, 2017.

[5] L. I. Besaleva and A. C. Weaver, "Crowdsourcing for emergency response," in *Proceedings of the International Conference on Frontiers in Education: Computer Science and Computer Engineering (FECS)*. The Steering Committee of The World Congress in Computer Science, Computer . . . , 2016, p. 248.

[6] B. Truong, C. Caragea, A. Squicciarini, and A. H. Tapia, "Identifying valuable information from twitter during natural disasters," *Proceedings of the American Society for Information Science and Technology*, vol. 51, no. 1, pp. 1–4, 2014.

[7] C. Caragea, A. Silvescu, and A. H. Tapia, "Identifying informative messages in disaster events using convolutional neural networks," in *International Conference on Information Systems for Crisis Response and Management*, 2016, pp. 137–147.

[8] M. Rhodan, "'please send help.' hurricane harvey victims turn to twitter and facebook," August 2017. [Online]. Available: http://time.com/4921961/hurricane-harvey-twitter-facebook-social-media/.

[9] B. Mandel, A. Culotta, J. Boulahanis, D. Stark, B. Lewis, and J. Rodrigue, "A demographic analysis of online sentiment during hurricane irene," in *Proceedings of the Second Workshop on Language in Social Media*. Association for Computational Linguistics, 2012, pp. 27–36.

[10] H. M. Saleem, Y. Xu, and D. Ruths, "Effects of disaster characteristics on twitter event signature," *Procedia engineering*, vol. 78, pp. 165–172, 2014.

[11] B. Wang and J. Zhuang, "Crisis information distribution on twitter: a content analysis of tweets during hurricane sandy," *Natural Hazards*, vol. 89, no. 1, pp. 161–181, 2017.

[12] Y. Zhang, W. Drake, Y. Li, C. W. Zobel, and M. Cowell, "Fostering community resilience through adaptive learning in a social media age: Municipal twitter use in new jersey following hurricane sandy." in *ISCRAM*, 2015.

[13] A. Gupta, H. Lamba, P. Kumaraguru, and A. Joshi, "Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy," in *Proceedings of the 22nd international conference on World Wide Web*. ACM, 2013, pp. 729–736.

[14] R. Pandey, H. Purohit, J. Chan, and A. Johri, "Ai for trustworthiness! credible user identification on social web for disaster response agencies," *arXiv preprint arXiv:1810.01013*, 2018.

[15] N. O. Hodas, G. Ver Steeg, J. Harrison, S. Chikkagoudar, E. Bell, and C. D. Corley, "Disentangling the lexicons of disaster response in twitter," in *Proceedings of the 24th International Conference on World Wide Web*. ACM, 2015, pp. 1201–1204.

[16] K. Stowe, J. Anderson, M. Palmer, L. Palen, and K. Anderson, "Improving classification of twitter behavior during hurricane events," in *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media*, 2018, pp. 67–75.

[17] A. Anam, A. Gangopadhyay, and N. Roy, "Evaluating disaster time-line from social media with wavelet analysis," in *2018 IEEE International Conference on Smart Computing (SMARTCOMP)*. IEEE, 2018, pp. 41–48.

[18] J. Weng and B.-S. Lee, "Event detection in twitter." *ICWSM*, vol. 11, pp. 401–408, 2011.

[19] X. Dong, D. Mavroeidis, F. Calabrese, and P. Frossard, "Multiscale event detection in social media," *Data Mining and Knowledge Discovery*, vol. 29, no. 5, pp. 1374–1405, 2015.

[20] R. Büssow, "An algorithm for the continuous morlet wavelet transform," *Mechanical Systems and Signal Processing*, vol. 21, no. 8, pp. 2970–2979, 2007.

[21] C. Torrence and G. P. Compo, "A practical guide to wavelet analysis," *Bulletin of the American Meteorological society*, vol. 79, no. 1, pp. 61–78, 1998.

[22] B. R. Bakshi, "Multiscale pca with application to multivariate statistical process monitoring," *AIChE journal*, vol. 44, no. 7, pp. 1596–1610, 1998.

[23] N. Slonim and N. Tishby, "Document clustering using word clusters via the information bottleneck method," in *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2000, pp. 208–215.