Event Characterization and Separation using Wavelet Signatures

Amrita Anam
amrita1@umbc.edu
University of Maryland Baltimore County
Baltimore, MD

Aryya Gangopadhyay
gangopad@umbc.edu
University of Maryland Baltimore County
Baltimore, MD

Nirmalya Roy
nroy@umbc.edu
University of Maryland Baltimore County
Baltimore, MD

ABSTRACT

In this paper, we present a method for developing event signatures from social media data. Social media posts like tweets contain signals from social sensing. Our method describes how such signals can be extracted and analyzed to create the underlying event signatures. We compared our proposed methodology with other contemporary methods such as *word2vec* and time series analysis.

CCS CONCEPTS

• **Information Systems** → **World Wide Web**; • **World Wide Web** → *Web applications*; • **Web application** → Social Networks.

KEYWORDS

Wavelet transform, social media, emergency response, event characterization

ACM Reference Format:
Amrita Anam, Aryya Gangopadhyay, and Nirmalya Roy. 2018. Event Characterization and Separation using Wavelet Signatures. In *MileTS '19: 5th KDD Workshop on Mining and Learning from Time Series, August 5th, 2019, Anchorage, Alaska, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

We propose a method for analyzing social sensing with wavelet signatures for natural disasters based on social media posts. The data we collected and analyzed are related to several recent natural disaster events in the United States. In this paper, we demonstrate that information belonging to every physical event sensed from social media posts can be represented by the frequency distribution of a list of 'context words' over a finite length of time. The proposed approach transforms these 'context words' into time-dependent signals containing a spectrum of frequencies and amplitudes that vary with time. The process of creating word signals from tweets explained in our previous work [1] where we also demonstrate how the wavelet representation can be used in useful applications such as real time prediction of event characteristics and trajectory. By combining signal processing techniques with data science, we intend to substantiate the hypothesis that a physical event on social media leaves an identifiable signature in the wavelet space. The wavelet signatures are homogeneous for the same kind of events and are different for events of different kinds. We use Twitter as the platform for social sensing because tweets are sensitive to both time and length of the context. We establish our findings with four datasets from two types of natural disasters - hurricanes and flash floods, collected from the recent events that occurred in the United States: i) hurricane Michael *(HM)*, ii) hurricane Florence *(HF)*, iii) flash flood in Arizona swimming hole *(AZ)*, and iv) flash flood in Cummins Falls state park *(CF)*. We characterize both large scale national disasters such as hurricanes and small scale local disasters such as flash floods with signatures in time and frequency domains.

Subsequently, we demonstrate that multiresolution analysis with Continuous Wavelet Transforms (CWT) can be used to identify the signature of a physical event. We show that different events can be separated through clustering the data in the wavelet space. Although events of similar types are harder to separate as opposed to those of different events, in both cases, we achieve superior accuracy over baseline methods such as word embedding and the basic word signals.

2 TIME-FREQUENCY REPRESENTATION

We collected tweets, each tweet was associated with a date and time of its creation. The pre-processing of the data comprised of a five-step process to build the vocabulary of context words that are reflective of the event. Twitter is similar to a noisy sensor. The pre-processing stage separates noise from the signals by adding a filter to limit the data to meaningful inputs. We pre-process the data by removing duplicate tweets, numbers, symbols, URLs and stop words. We tokenized the filtered tweets into unigrams keeping words only valid in English vocabulary. We use the time information to create time-dependent word signals by binning them with a fixed duration (Δt). The start time of the j+1$^{\text{th}}$ bin was calculated by $t_{j+1} = t_j + \Delta t$, where $j = 1 \ldots n$. We chose document frequency (df) as the amplitude of the word signals[8]. The df of the i^{th} word for the j^{th} bin is the number of tweets in bin j that contain u_i. Δt is determined by observing the data variation as explained in [1]. The word signals created from for flash flood datasets and hurricane datasets are presented in fig. 1. CWT was applied on the

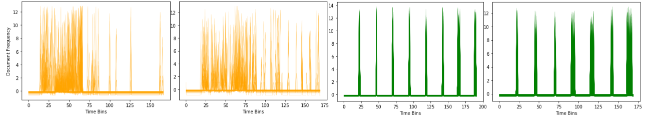

Figure 1: Signal representation of context words

MileTS '19, August 5th, 2019, Anchorage, Alaska, USA

ACM ISBN 978-1-4503-9999-9/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

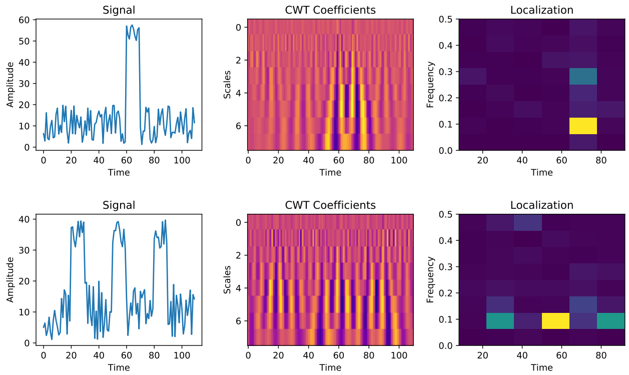

Figure 2: Time-Frequency Localization of two signals

word signals with *Morlet* wavelet [2], [7] as mother wavelet and a preset number of scales.

3 EVENT CHARACTERIZATION WITH WAVELET SIGNATURES

A key contribution of this paper lies in finding the characteristics of physical events using wavelets. To instantiate the idea of event characterization, we present two cases with different characteristics of non-stationary signals in both frequency and time domains illustrated in fig. 2. The left two images represent the signals containing 100 samples each. They are sampled with the same sampling frequency and transformed with the same mother wavelet with 32 scales. The figures in the middle column reflect the wavelet coefficients and the figures on the right are the spectrograms of the signals to understand how the spectrum of frequencies vary with time. The signal in the top row has a single peak in the time axis which is anomalous to the rest. This results in higher energy in the wavelet coefficients captured by the bright yellow vertical cone in the middle. The spectrogram highlights two signals in the frequency axis reflecting the low-frequency components and the high-frequency components. The low frequency components reflected by the rectangle between frequency 0 to 0.1 are stronger. The figures in the top row of fig. 2 is an example of a signal localized in time. The signal showed in the bottom row of fig. 2 contains periodically recurring peaks. The Fourier transform identifies the main frequency which is shown by rectangles between frequency 0 to 0.1. The time, however, shows recurring instances when the signal has is stronger. The spectrogram depicts that signal with the recurring peaks is localized in frequency but not in time.

Our methodology using wavelet signatures allows us to identify a pattern in time-frequency localization of signals which can be used to uniquely characterize their respective events. We performed an exploratory analysis using Shannon Wavelet Entropy (SWE) with scale and time on all four datasets to identify the signatures that can characterize hurricanes and flash floods. The flash flood datasets are collected from local, small scale disasters which caused flooded withing minutes of hitting the land which lasted for a few days. On the contrary, the hurricanes datasets are collected from large scale national disasters which impacted multiple states and the flooding lasted for weeks. Comparing with the word signals in fig. 1, the flash flood datasets follow the characteristics the example in the top row figures in fig. 2 and the hurricane datasets mimic the example in the figures in the bottom row of fig. 2.

3.1 Scale Entropy

We calculate SWE for each scale over all the time components to find the distribution of entropy over all scales. Let, j, k be the scale and time components of a wavelet and $C_j(k)$ be the wavelet coefficient of signal s at time k and scale j, then the wavelet energy E_j can be calculated by,

$$E_j = \sum_k |C_j(k)|^2 \tag{1}$$

Next, we calculate the The Relative Wavelet Energy (RWE) at scale j by normalizing the energy at every scale by total energy. The total energy is calculated by summing the energy over all scales,

$$E_{total} = \sum_j E_j \tag{2}$$

RWE represents the distribution of wavelet energy across different scales. RWE is at scale j can be retrieved by,

$$\rho_j = \frac{E_j}{E_{total}} \tag{3}$$

SWE at scale j is then calculated by,

$$Sswe_j = -\rho_j \cdot log\rho_j \tag{4}$$

3.2 Time Entropy

To find a significant time duration where the signal has maximum entropy we calculate SWE at every time bin. Let, j, k be the scale and time components of wavelet and $C_k(j)$ be the wavelet coefficient at time k and between scales $j1$ to $j2$, then the wavelet energy E_j can be calculated by,

$$E_k = \sum_{j=j1}^{j=j2} |C_k(j)|^2 \tag{5}$$

The total energy can be calculated by,

$$E_{total} = \sum_k E_k \tag{6}$$

The Relative Wavelet Energy (RWE) at scale j can be retrieved by,

$$\rho_k = \frac{E_k}{E_{total}} \tag{7}$$

RWE represents the distribution of wavelet energy across different scales. We can calculate SWE of a signal s at for a window w by summing over the entropy of the RWE values of all the scales.

$$Tswe_k = -\rho_k \cdot log\rho_k \tag{8}$$

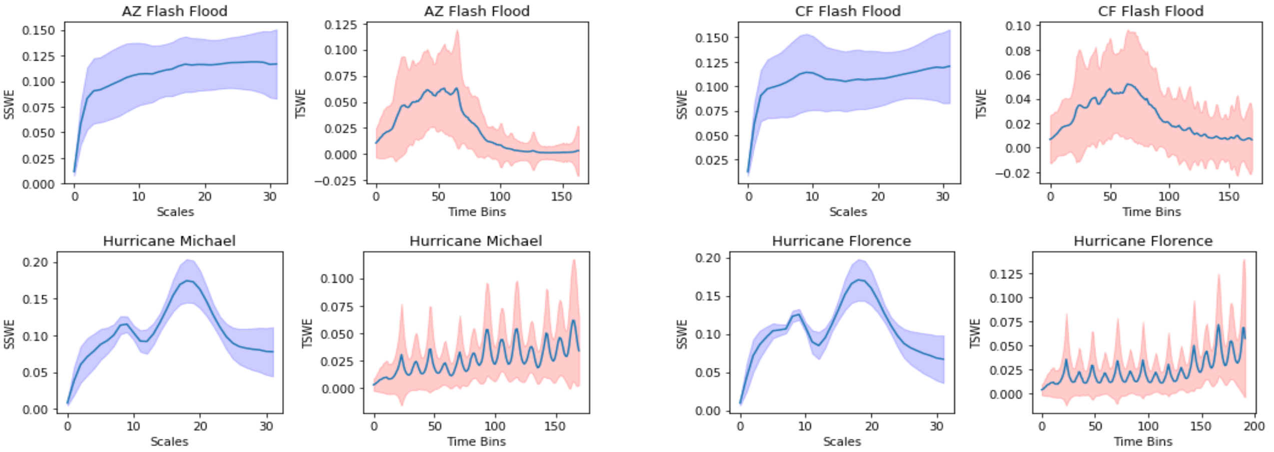

Figure 3: Event signature

3.3 Observations

We observed the mean and standard deviation of *Sswe* for all four datasets. There was a noticeable difference in behavior behaviors between the hurricanes and the flash flood, as expected. From both our hurricane datasets we observe that choosing a scale range is important because the RWE and SWE are impacted by scale. The signals in both HM and HF datasets, have significantly high energy between scale 15 and 21 and hence can be localized in scale. The entropy *Sswe* has a sharp increase between scale 15 to 21. The *Sswe* has a downward trend after scale 21. On the other hand, the flash flood data sets have no significant band with high intensity in energy and entropy and hence cannot be localized in scale. They have a sharp increase at the initial scales, 0 to 2 then stays steady throughout.

Along with RWE, SWE for all the observations, we also analyze the localization of a disaster in time by analyzing the mean and the standard deviation of the *Tswe*. Following our previous assumption that the local small scale disasters are localized in time, we observe a sharp rise in the RWE and SWE in AZ and CF dataset between day 2 and day 4. However, for the hurricane datasets, we do not see any time localization. Even though there is a slight bump on the *Tswe* in the first couple of days in hurricane Michael, it did not give us any meaningful range because for large scale disasters the tweet flow stays high beyond a week.

4 EVENT SEPARATION WITH CLUSTERING

In this section, we follow the process flow given in fig. 4 to distinguish events from a fused sample. As shown in the architecture we simulate the cases by taking samples from each event and aligning them at the same time or with a time shift. For similar event samples, we apply the pipeline on both the hurricane and flash flood data sets. For different events, we take a hurricane sample and a flash flood sample as the inputs. The samples from two events are then fused together to create one list of vocabulary. The vocabulary is then converted into word signals with bins created from the first and last date of the simulated time line. The bin duration is kept the same for all four cases for comparison. The setup for the problem is the four different situations to identify and separate two ongoing events from their respective tweets. The assumption behind this design is that each event sample will have their own word distribution which follows the distribution by the actual incidents. These distributions are time-sensitive which is why the shift in time will have an impact on the clustering. The shift will also have an impact in the energy of the coefficients which will separate them. These cases are:

(1) **Same Time - Similar Events** We align the two random samples from similar events on the exact same timeline. This is the hardest of the four to separate because there is no distinction coming from the time components and the word distribution from similar events are also very similar.
(2) **Time Shift - Similar Events** We align the two random samples from similar events with a time shift less than the total duration of the samples so that there is some overlap. In this case, the factor that can create good clusters mostly come from time. The context words that are unique to an event will drive the force of the cluster separation.
(3) **Same Time - Different Events** We align two random samples from two different events on the same timeline. The

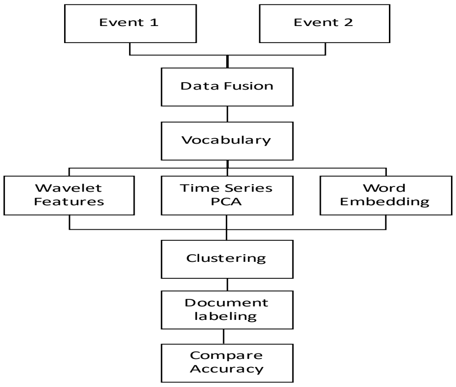

Figure 4: Process flow for event separation

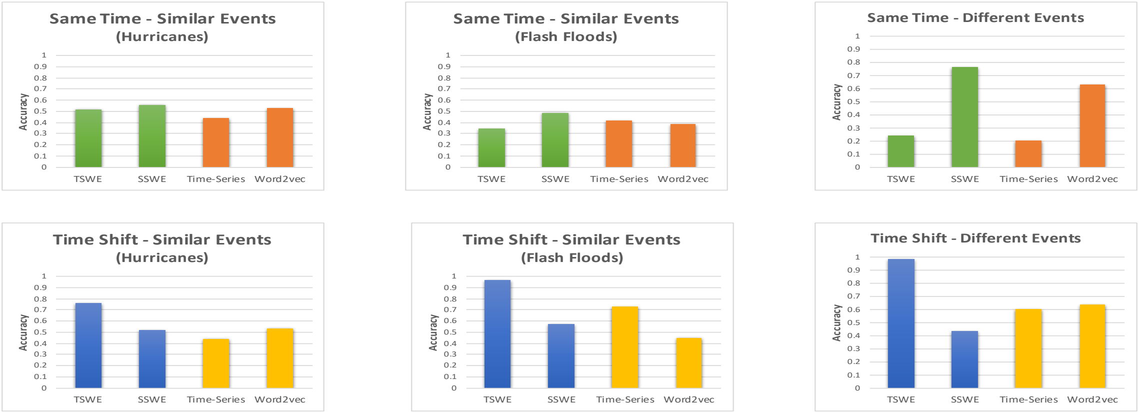

Figure 6: Event separation

distinct factors for this case can come from the frequency spectrum of each event along with the word distribution.

(4) **Time Shift - Different Events** We align two random samples from two different events with a time shift with some overlap. If the events have completely different word distributions, then this is the best case for cluster separation because we would get two completely different set of words being reflected in two different time settings and words belonging to the same event will have similar representations.

4.1 K-means Clustering

K-means clustering algorithm is a simple, robust algorithm that partitions observations into k clusters by comparing the distance between the features with the cluster centroids. The cluster centroids are adjusted through an iterative refinement technique so that all observations that are clustered in one group is nearest to the centroid of that cluster. We apply the K-means algorithm with Euclidean distance as the distance metric to compare our feature sets (i.e., time entropy, scale entropy) and the baselines (basic word signals and *word2vec* [5]) for efficient context separation by word clusters. Word2vec is a shallow neural network based probabilistic model wich creates word embeddings. We compute word embeddings for length of fifty. We reduce the dimensions of the time entropy, scale entropy and baseline word signals with Principal Component Analysis (PCA). For validation, the documents were assigned to our word cluster using a modification of information bottleneck method [6]. Instead of building a clustering algorithm from probabilities, we just use the clusters achieved from K-means. We calculate the accuracy of the labels of the assigned documents as the performance metrics to compare the features, where accuracy $= \frac{\text{Number of correctly identified instances}}{\text{Number of total instances}}$

4.2 Results

The results of the four cases are presented in fig. 6. In the first case of 'same time -similar events', none of the features achieve good accuracy because the word distributions are very similar in both the samples. In the second experiment, 'time shift - similar events' - the time entropy for both hurricanes and flash floods performs best where the flash floods achieve 96% accuracy outperforming the baseline word signals which achieves 70% accuracy. In the third case, 'same time - different events', our feature scale entropy outperforms the best performing baseline *word2vec* with 60%. In the last case, 'time shift - different events', time entropy achieves 98% accuracy outperforming *word2vec.* It is obvious that among the baselines, the time shift has very little impact on word2vec. However, features such as time entropy and baseline time series PCA are sensitive to time and present much better results when the data is shifted.

5 RELATED WORK

We were particularly motivated by the work in [8], where the authors used a mash-up of signal processing techniques with text mining techniques on twitter posts for real-time event detection. In social-media analysis, it is important to utilize the information with both time and frequency which motivated researchers to analysis social-media data with wavelets. Content-based clustering was used in [3], [4] to find temporal patterns in social media. The authors of [3] find a diversity of content where diversity is defined by a change in entropy in a different spectrum of the wavelets created from time dependents signals of content. The authors converted time-dependent signals of clicks, hash-tags, and phrases to wavelets and developed a clustering algorithm K-Spectral Centroid (K-SC) to cluster them to find temporal pattern [4].

6 CONCLUSION

This paper contains a novel method that demonstrates that physical events can be characterized by wavelet signatures. We show that wavelet signatures can be used to distinguish events from a fused dataset. Two co-occurring events are hard to separate because of the lack of distinct features whereas different events with a time shift are the easier to separate. It is also visible that most of the wavelet features work better than other current approaches. Our future work includes finding more signatures for different events, using signatures for transfer learning and building hybrid models mixing probabilistic features with wavelets.

REFERENCES

[1] Amrita Anam, Aryya Gangopadhyay, and Nirmalya Roy. 2018. Evaluating Disaster Time-Line from Social Media with Wavelet Analysis. In *2018 IEEE International Conference on Smart Computing (SMARTCOMP)*. IEEE, 41–48.

[2] Richard Büssow. 2007. An algorithm for the continuous Morlet wavelet transform. *Mechanical Systems and Signal Processing* 21, 8 (2007), 2970–2979.

[3] Munmun De Choudhury, Scott Counts, and Mary Czerwinski. 2011. Find Me the Right Content! Diversity-Based Sampling of Social Media Spaces for Topic-Centric Search.. In *ICWSM*.

[4] Xiaowen Dong, Dimitrios Mavroeidis, Francesco Calabrese, and Pascal Frossard. 2015. Multiscale event detection in social media. *Data Mining and Knowledge Discovery* 29, 5 (2015), 1374–1405.

[5] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.

[6] Noam Slonim and Naftali Tishby. 2000. 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, 208–215.

[7] Christopher Torrence and Gilbert P Compo. 1998. A practical guide to wavelet analysis. *Bulletin of the American Meteorological society* 79, 1 (1998), 61–78.

[8] Jianshu Weng and Bu-Sung Lee. 2011. Event detection in twitter. *ICWSM* 11 (2011), 401–408.