Predicting Earthquake Epicenters with Twitter

Marley Beckett
Department of Electrical and Computer Engineering
University of British Columbia
marleybeckett@ece.ubc.ca

Graham AD Archibald
School of Biomedical Engineering
University of British Columbia
graham.archibald@ubc.ca

Abstract

Earthquakes often occur with no warning making disaster response and subsequent management daunting, if not impossible, tasks. Social media messages can be an important source of information during natural disasters, as they can provide real-time, *in-situ* details much faster than traditional sources such as news stations and government warning systems. Further, areas with less sophisticated infrastructure and disaster response resources are rapidly gaining near universal internet and social media access. This project presents a machine learning approach using topic modelling to predict the epicentre of an earthquake for real time analysis and classification. The results illustrate that with relevant social media data, the epicenter of an earthquake can be predicted with great accuracy.

1 Introduction

Natural disasters present all-encompassing threats to human society, and loom even as we move through the twenty-first century. For disasters such as earthquakes, it is critical that emergency response teams are able to quickly assess the magnitude and geographical location of such disasters. In doing so, an appropriate response can be initiated. Such response is crucial for providing support to those impacted, as well as mitigating loss of life and damage to infrastructure.

Advancing social-media based disaster interventions is of crucial importance as the world becomes increasingly connected through the internet. When 8000 Nepalese were killed by an earthquake in April of 2015, social media tools were flooded with requests for help as well as information related to specific aspects of the damage. Unfortunately, there was little capacity to gather and arrange the data into an interpretable form. When this occured, a mere 40% of Nepalese were online [1], a number which has since risen to above 90% in just seven years. Since 2005, nearly four billion people have gained access to the internet, with most new users being in the developing world which broadly lacks sophisticated seismology equipment and infrastructure [2]. Advances in technologies such as 5G and Device to Device (D2D) are making real-time social media interventions even more feasible for disaster scenarios [3]. Developing machine learning (ML) based tools which can leverage this rapidly increasing 'dataset' is crucial for improving global disaster response equity.

While previous works have analyzed social media behaviours during crises, they have generally lacked in quantifying at least one of the spatial, temporal, or semantic nature of the data. In this project, we use real time social media data, namely 'tweets', as input to an ML model for prediction of an earthquake epicentre by analyzing the tweets surrounding the 2014 Napa California earthquake.

2 Related Work

Widespread adoption of social media over the past decade in particular has brought light to the concept *User Generated Data* (UGD). UGD has proven to be a good source of information for applications that aim to understand public opinions and feelings relating to a specific topic or event.

Capturing UGD from Twitter, for both sentiment analysis and topic modeling, became popularized in 2008 with the release of the public Twitter API the year prior [4]. Previous works have shown that in comparison to other social media platforms, Twitter prevails as a top method for capturing UGD, as the 140-character limitation on tweets makes it similar to English language sentences [5].

There is no *de-facto* standard for the topic modelling of Tweets during a crisis. Several applications of crisis topic modelling, such as virus outbreaks [6], political crises [7], natural disasters [8][9][10], [11], and social interactions [12] were reviewed; the algorithms used in each are displayed in Table 1. For applications where geographical filtering was used for classification tasks, Latent Dirichlet allocation (LDA) model prevails as the most common approach. This unsupervised algorithm has proved to be an effective probabilistic classifier for Twitter topic modelling.

One example, where LDA is applied to Tweets in the context of a natural disaster, is a study by Kireyev *et al.* [13] where LDA is leveraged for the extraction of disaster-related Tweets. Their focus is on improving the algorithm itself by using a term weighting function, therefore addressing the small information contained in single Tweets. Topics covering single events like such as earthquakes tsunamis are successfully extracted, but no geospatial analysis is presented, which is crucial when intended to be used as an intervention.

In a 2022 publication by Rachunok et al.[14], location based filtering resulted in higher accuracy for topic modeling of UGD if the event in consideration is limited to a geographical area. Additionally, with the steady decline of hashtag usage within tweets since 2016, keyword filtering is becoming a less popular method. The dataset used in this project was obtained using location based filtering.

For this report, we referenced a primary academic article that focused on topic modelling of the 2014 Napa (CA) earthquake. Resch et al.[8] proposed a Bag of Words (BoW) approach for feature selection in conjunction with an ensemble Latent Dirichlet Allocation (LDA) classification model. Tweets classified with high probability under the "earthquake" topic from the primary LDA were cascaded into a secondary LDA for further topic breakdown. A geographical analysis of the earthquake damage was reported, but this work did not consider epicentre prediction. Additionally, no justification was provided as to why a LDA model was selected.

Author	Year	Model	Description
Kireyev [13]	2009	LDA	Improving LDA for short documents
Mishler [7]	2015	STM (LDA variation)	Cluster Twitter users in the Ukrainian crisis
Do [6]	2016	Emsemble SVM	Analyzing emotions in virus outbreak
Resch[8]	2017	LDA (BoW)	Predicting disaster footprint (Napa 2014)
Alam [9]	2019	Random Forest	Visual summaries of disaster events using AI
Ahadzadeh [10]	2021	SVM (GRBF Kernel)	Earthquake damage assessment (Napa 2014)
Kim [12]	2021	LDA (BoW)	Detecting Asian-based hate crimes
Madichetty [11]	2021	SVR (PoS)	Detecting earthquake related tweets

Table 1: Comparison methods of topic modeling for Twitter data during a crisis.

3 Problem Formulation

We focused our research project on the highly documented 2014 South Napa (CA) earthquake. At 3:20 AM local time (10:20:44 UTC) on August 24, 2014, residents of Napa, California witnessed a magnitude 6.0 earthquake; the largest to take place in the San Francisco Bay area since 1906. The earthquake epicentre was located in a densely populated area, making the extraction of key information a crucial task. The raw data used in this study is comprised of 998,719 geotagged tweets from August 17, 2014 to August 31, 2014 [8], and is summarized in Table 2.

This event took place at a time when social media usage, namely Twitter, was increasing in popularity. The prevalence of Twitter usage in the hours immediately following the earthquake are visualized in Figure 7 below. On August 24, the term "earthquake" was the highest *tweeteed* keyword, as shown in Figure 1.

Dataset Description	
Time Period	August 16, 2014 - August 31, 2014
Number of Tweets Collected	998,719
Geographical Bounds	[-123.5 °W, 37.19 °N] , [-121.04 °W, 38.99 °N]
Number of Tweets Post-Processing	798,196

Table 2: Overview of data from the 2014 Napa earthquake [8].

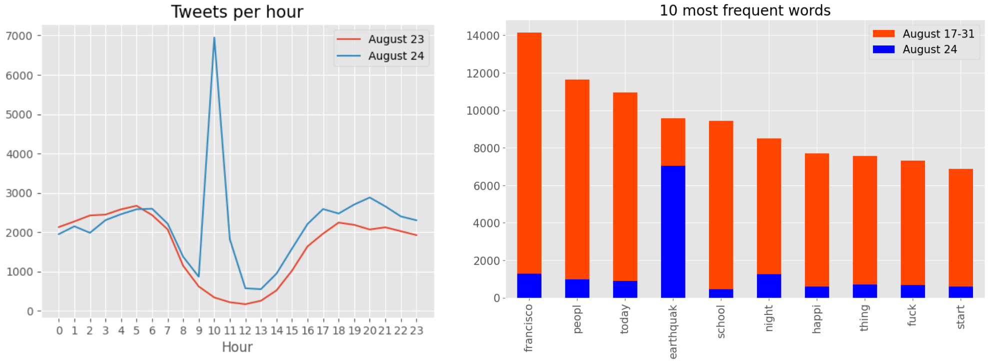


Figure 1: (a) Total Tweets per hour on the days of August 23, 2014 and August 24, 2014 (UTC). (b) Most common words for the entire dataset compared to on the day of the earthquake.

This project aims to predict the epicentre of an earthquake using Twitter data alone. Previous approaches using this dataset included predicting a damage map from the data, however epicentre location was not included in their scope. Additionally, these approaches focused on using data collected on August 24, however the approach followed in this project makes use of the entire dataset. Predictions will be made by allowing the LDA to discover if an "earthquake" topic is among one of the N most common topics overall.

4 Methodology

4.1 Pre-processing

To analyze the sentiment of tweets, the text-based data must be presented in a format robust to outliers and repetitions. URLs are considered noise in our topic modeling approach because they contain unspecific and largely uninterruptible information and are therefore removed. Numbers are removed, because they usually do not contain semantically valuable information. One exception could be the magnitude description of the earthquake, however the word ‘magnitude’ will convey this meaning and is represented in the topic model. Similarly, special characters are removed to preserve the focus on words in the dataset. While we do assume that emoticons and emojis are frequently used to express emotion, they are not standardized and have been removed as a result. Additionally, duplicate words within individual tweets and duplicate tweets have been removed to account for *spam* messages, which are prevalent. Although this step was not taken in the reference paper, we have include it as it bettered our results [8].

Stopwords are commonly used words which do not carry distinct semantic meaning, such as auxiliary verbs, conjunctions, and articles. These terms appear in almost every tweet and would form a topic of their own as they co-occur frequently [15]. Thus, we removed stopwords using a predefined list from the NLTK Toolkit.

Words with three characters or less were also removed as the literature suggests that such words contain little semantic meaning [16]. To deal with outliers, words which appear less than 3 times in the text corpus have been removed as they do not contribute significantly to a topic and drastically inhibit the algorithm’s performance.

4.1.1 Stemming Tokenization

For grammatical reasons, documents use various forms of a word with the same ‘stem’, such as ‘look’, ‘looks’, and ‘looking’. Further, there are families of words with similar meanings, such as ‘democracy’, ‘democratic’, and ‘democratization’. In many situations, it is useful for a search for one of these words to return documents that contain another word in the set. We used a Porter stemmer that reduces single words to their root, resulting in a condensation of the text corpus which in turn increases significance of the word-topic association.

After all pre-processing steps took place, individual words were separated into *tokens* for BoW referencing. After the described steps have been applied, the majority of the *cleaned* tweets are only a few tokens in length.

4.1.2 Vectorization

A BoW method of vectorization was used to convert text-based data to numerical values. *CountVectorizer* embedding was used for the text analysis. This method converts tokenized words into sparse vectors that represent the vocabulary of tweets within the dataset. To address *coupon-collecting*, words were only added to the *corpus* if they were in the top 50,000 words to occur. Additionally, words synonymous with earthquake, such as *shake*, *tremble*, and *quake*, were manually added to a dictionary, and were referenced as the single word *earthquake*. Figure 1 displays the most common words in both the entire training set and on the day of the earthquake.

4.2 Implementation

Based on the assumption that each tweet can be classified as a mixture of N topics, the LDA model determines underlying topics and the relative topic distribution of each tweet. Using the conventional description of LDA models, an individual word within a tweet is a *token* and each individual tweet is a *document*. The entire collection of documents is referred to as a *corpus* [17].

Adjustable model parameters of LDA include N, the number of topics, α, Dirichlet prior on per-document topic distributions, and β, the Dirichlet prior on per-topic word distributions [17]. Additionally, corpus parameters *noAbove* and *noBelow* control the frequency of tokens and documents included. The LDA model as it relates to this project is shown in Figure 2, where θ is the distribution of topics in documents, Z is the identity of topic of all words in all documents, and W is the identity of all words in all documents.

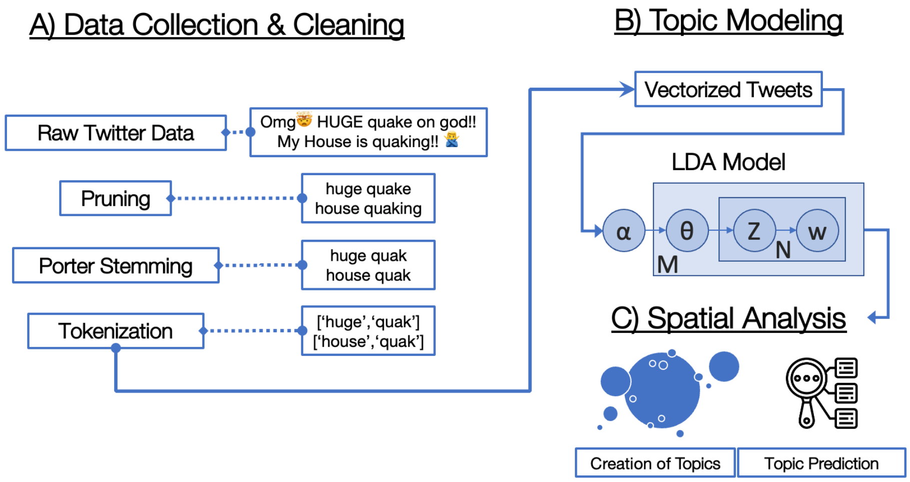


Figure 2: Workflow from pre-processing to analysis. **A)** begins with raw Twitter data and outputs cleaned, vectorized documents. In **B)**, we use a LDA model to convert the vectorized tweets into salient topics. For the spatial analysis component in **C)** we are able to view the created topics as well as predict how new documents should be classified.

The LDA model is evaluated through both coherence and perplexity scores. Coherence enables topics to be interpreted as being coherent if the majority of the terms used are related [18]. Perplexity is an intrinsic evaluation measure that describes how well the LDA can reproduce the statistics of the previously omitted data; perplexity is the normalized log-likelihood of a test set [9]. Coherency is often a better measurement of the accuracy of LDA topic modeling for Tweets, as it only measures the probability of a specific observation, and the internal representation of the Tweet is ignored [19].

Implementation and evaluation of the LDA model was done in the following steps using the *gensim* LDA library:

1. Divide processed and tokenized Tweets into 70% *training*, and 30% *testing* datasets. Further divide training data into 5 folds for model cross-validation.

2. Create a corpus instance for each of the training folds.

3. Train several LDA models using different values of *N* on 4-folds of the training data.

4. Evaluate the LDA model and select *N* by:

   (a) Observing the word distribution for the topic that contained the highest distribution of the word "earthquake".

   (b) Produce a coherence score for the training data.

   (c) Produce a perplexity score for each of the models on the remaining fold of training data.

5. Create a corpus from the reserved testing data, and evaluate the above metrics by implementing an iteration of the highest scoring LDA model from (4).

During training, we observed the word distribution for keywords within the "earthquake" topic increasing with the number of topics modelled as seen in Table 3. Between training iterations of $N = 20, 25$, a secondary topic relating to earthquakes emerged. To increase the probability of correctly classifying earthquake related tweets under a single topic, it was chosen to move forward with the LDA model trained using $N = 20$, although this model achieved worse perplexity and coherence scores than models with $N = 25$ and 30.

N	Word Distribution for Earthquake Topic on the Test Set
5	(0.157*"earthquak" + 0.023*"sleep" + 0.020*"california" + 0.011*"thought" + 0.011*"night")
10	(0.078*"earthquak" + 0.045*"california" + 0.028*"watch" + 0.022*"happen" + 0.020*"tweet")
15	(0.344* "earthquak" + 0.046*"night" + 0.045* "california" + 0.033* "damag" + 0.027* "morn")
20	(0.437*"earthquak" + 0.054*"california" + 0.019*"magnitud" + 0.015*"report" + 0.013*"aftershock")
25	(0.490*"earthquak" + 0.063*"california" + 0.047*"thought" + 0.039*"morn" + 0.024*"street"), (0.062*"damag" + 0.035*"aftershock" + 0.032*"point" + 0.031*"berkeley" + 0.027*"swear")
30	(0.730*"earthquak" + 0.045*"california" + 0.025*"drunk" + 0.024*"report" + 0.015*"weird"), (0.113*"morn" + 0.067*"parti" + 0.046*"aftershock" + 0.028*"cousin" + 0.023*"sonoma")

Table 3: Word-level distributions for the "earthquake" topic for different LDA models for different values of N.

5 Results

Following the model training mentioned in Section 4, the LDA model was applied to the reserved test set. Each datapoint was clustered according to the topic in which it obtained the highest probability of membership amongst the 20, and manual inspection of keywords determined the "earthquake" specific topic. In Figure 5 below, the datapoints classified under the "earthquake" topic during training are shown in yellow against all other 19 classes. Figure 5 shows this same result as a *1 vs. rest* approach on the testing data, where the blue points represent "earthquake" tweets. In both figures, a red marker was added to indicate the ground truth epicentre.

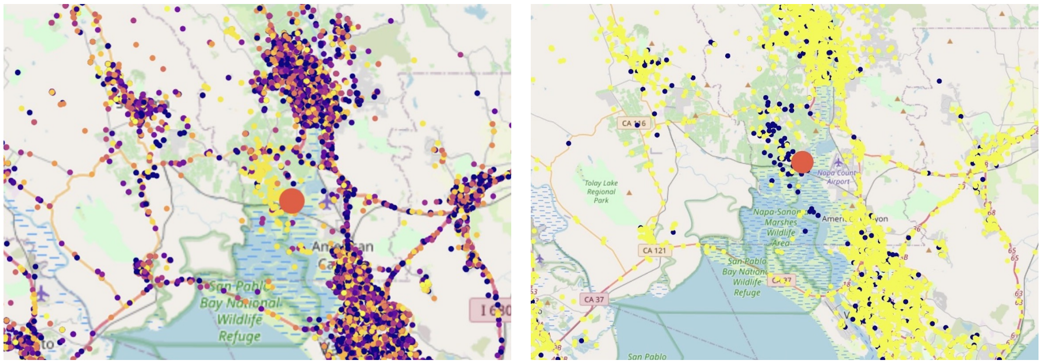

San Pablo Bay National Wildlife Refuge
CA 116
Tolay Lake Regional Park
Napa County Airport
CA 121
CA 37
Napa-Sonoma Marshes Wildlife Area
San Pablo Bay National Wildlife Refuge
San Pablo

Figure 3: Longitude and latitude of classified points for (a) training set (all 20 classes), and (b) test set (1 vs. rest).

The mean latitude and longitude datapoint belonging to the "earthquake" topic is displayed in Table 4. These results show that the epicentre of an earthquake can be located with high accuracy using a BoW and LDA model. In comparison to traditional machine learning models, LDA can experience a decrease in error from training to testing data, as additional words are added to the pre-existing corpus. Additional visualizations of this dataset are shown in Appendix I.

Predicted Epicentre	**Latitude (N)**	**Longitude(W)**
Training Data	38.256	-121.98
Testing Data	38.22	-122.26
Baseline (No LDA)	37.91	-121.96
Earthquake Epicentre	38.22	-122.31

Table 4: Mean latitude and longitude for LDA clustered data on training and testing data.

6 Conclusion

Knowing what information is available through social media aids emergency responders, humanitarian organizations, government organizations, and utility companies response to disasters allowing for effective resource allocation. In recent years, ML has been leveraged in conjunction with real time *microblogging*, such as Twitter, to aid with this task. Further research into this field has enormous potential for human benefit as millions of people, disproportionately in the developing world, gain internet access each year.

This project presented an approach for the prediction of an earthquake epicentre using probabilistic classification of Twitter data. We defined a set of pre-processing and feature engineering steps, and proposed a LDA model trained on text-based features to accomplish this task. The results confirmed that probabilistic classification of Twitter data can predict epicentres with high accuracy.

In discussion with Dr. Bernd Resch, we will continue to work on the prediction of an earthquake epicentre using natural language processing. Future directions include training other probabilistic models, such as variations on LDA, optimization of the pre-processing routines, and automated interpretation of the generated semantic topics.

References

[1] L. Thapa. Spatial-temporal analysis of social media data related to nepal earthquake 2015. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLI-B2:567–571, 2016.

[2] International Telecommunication Union. Measuring digital development: Facts and figures 2022 - internet use. `https://www.itu.int/itu-d/reports/statistics/2022/11/24/ff22-internet-use/`, Date accessed: 22.12.2022.

[3] Priyanka Rawat, Majed Haddad, and Eitan Altman. Towards efficient disaster management: 5g and device to device communication. In *2015 2nd International Conference on Information and Communication Technologies for Disaster Management (ICT-DM)*, pages 79–87, 2015.

[4] Kevin Makice. *Meeting the Twitter API*, page 133–141. O'Reilly, 2009.

[5] Efthymios Kouloumpis, Theresa Wilson, and Johanna Moore. Twitter sentiment analysis: The good the bad and the omg! *Proceedings of the International AAAI Conference on Web and Social Media*, 5(1):538–541, 2021.

[6] Hyo Jin Do, Chae-Gyun Lim, You Jin Kim, and Ho-Jin Choi. Analyzing emotions in twitter during a crisis: A case study of the 2015 middle east respiratory syndrome outbreak in korea. *2016 International Conference on Big Data and Smart Computing (BigComp)*, 2016.

[7] Alan Mishler, Erin Smith Crabb, Susannah Paletz, Brook Hefright, and Ewa Golonka. Using structural topic modeling to detect events and cluster twitter users in the ukrainian crisis. *Communications in Computer and Information Science*, page 639–644, 2015.

[8] Bernd Resch, Florian Usländer, and Clemens Havas. Combining machine-learning topic models and spatiotemporal analysis of social media data for disaster footprint and damage assessment. *Cartography and Geographic Information Science*, 45(4):362–376, 2018.

[9] Firoj Alam, Ferda Ofli, and Muhammad Imran. Descriptive and visual summaries of disaster events using artificial intelligence techniques: Case studies of hurricanes harvey, irma, and maria. *Behaviour amp; Information Technology*, 39(3):288–318, 2019.

[10] Sajjad Ahadzadeh and Mohammad Reza Malek. Earthquake damage assessment based on user generated data in social networks. *Sustainability*, 13(9), 2021.

[11] Sreenivasulu Madichetty and Sridevi M. A novel method for identifying the damage assessment tweets during disaster. *Future Generation Computer Systems*, 116:440–454, Mar 2021.

[12] Bumsoo Kim, Eric Cooks, and Seong-Kyu Kim. Exploring incivility and moral foundations toward asians in english-speaking tweets in hate crime-reporting cities during the covid-19 pandemic. *Internet Research*, 32(1):362–378, 2021.

[13] Kirill Kireyev. Applications of topics models to analysis of disaster-related twitter data. 2009.

[14] Benjamin Rachunok, Chao Fan, Ronald Lee, Roshanak Nateghi, and Ali Mostafavi. Is the data suitable? the comparison of keyword versus location filters in crisis informatics using twitter data. *International Journal of Information Management Data Insights*, 2(1):100063, Feb 2022.

[15] M Ikonomakis, S Kotsiantis, and V Tampakas. Text classification using machine learning techniques. *WSEAS Transactions on Computers*, 4(8):966–974, 2005.

[16] Alexander Pak and Patrick Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *International Conference on Language Resources and Evaluation*, 2010.

[17] David M Blei, Andrew Ng, and Michael Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, Mar 2003.

[18] Andreas Both Michael Röder and Alexander Hinneburg. Exploring the space of topic coherence measures. *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, 15:399–408, 2015.

[19] Jey Han Lau, David Newman, and Timothy Baldwin. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, page 288–296, Dec 2009.

7 Appendix I - Additional Figures

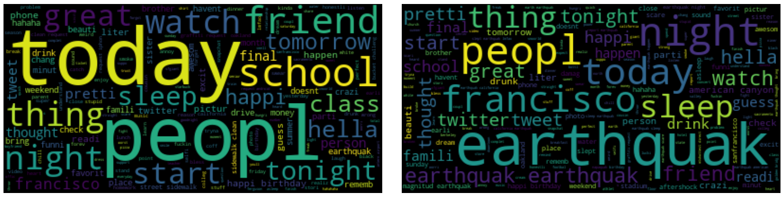


Figure 4: Word clouds created using tweets from (a) entire training set, and (b) August 24, 2014 only.

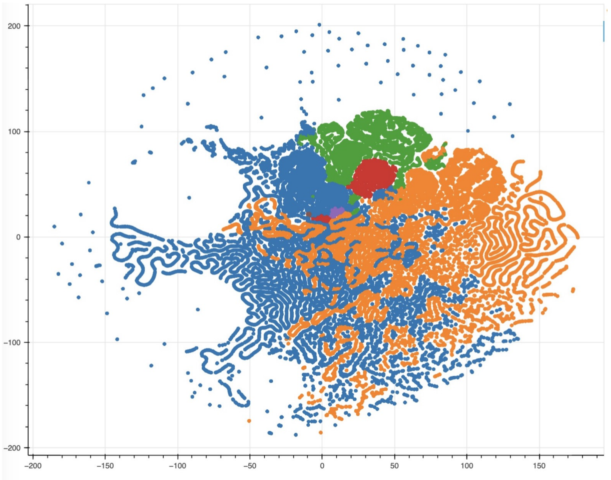
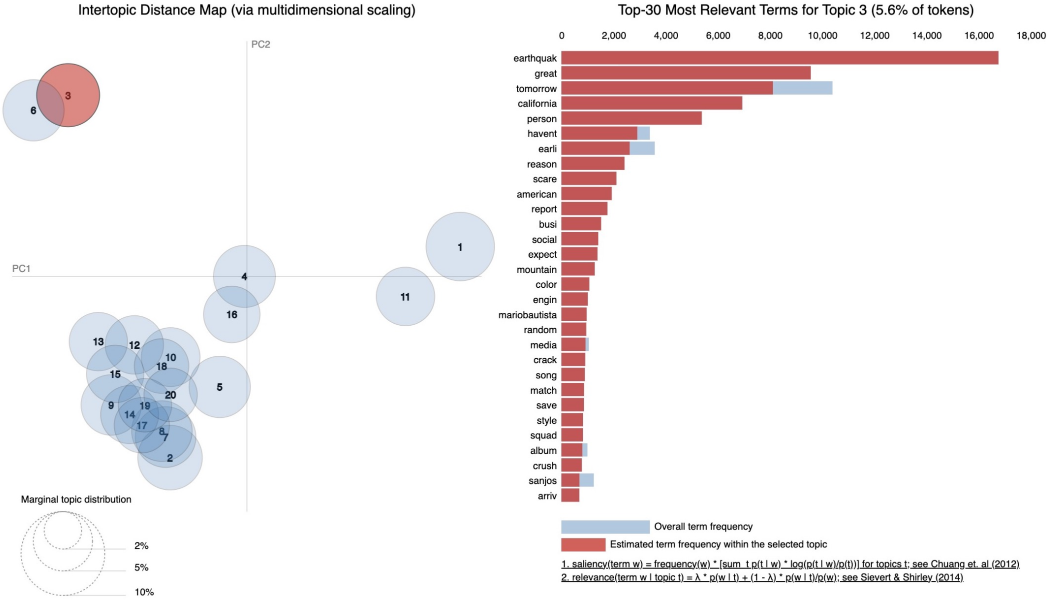


Figure 5: Results of TSNE clustering for tweets on the day of August 24, 2014.



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 6: Visualization of LDA on the test set with model trained using $N = 20$.