AI & SOCIAL MEDIA

based incident notification system

By Jiahua Gong
December 3, 2019

1 Introduction

Social media is not only a way to help people connect, but can also produce valuable data that makes people's life better.

The nature of social media that users generate the content, in a real-time manner, makes it a perfect source for incident detection, such as earthquake. However, there is challenge on analyzing and categorizing raw data.

This project aims to utilize AI to extract incident info from geo-tagged tweets and notify people nearby that may be affected to reduce losses.

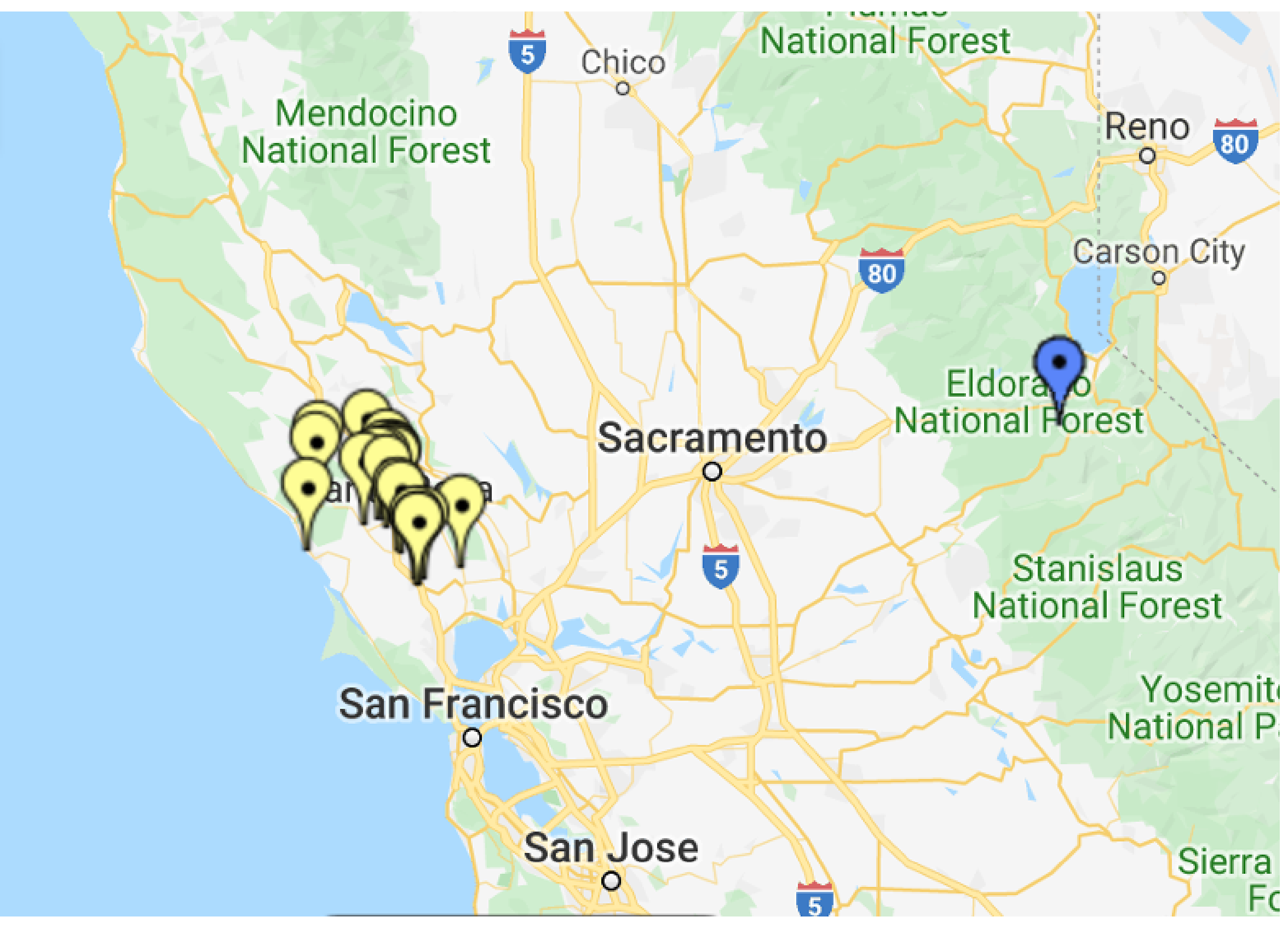


Figure 1. Real time incident map

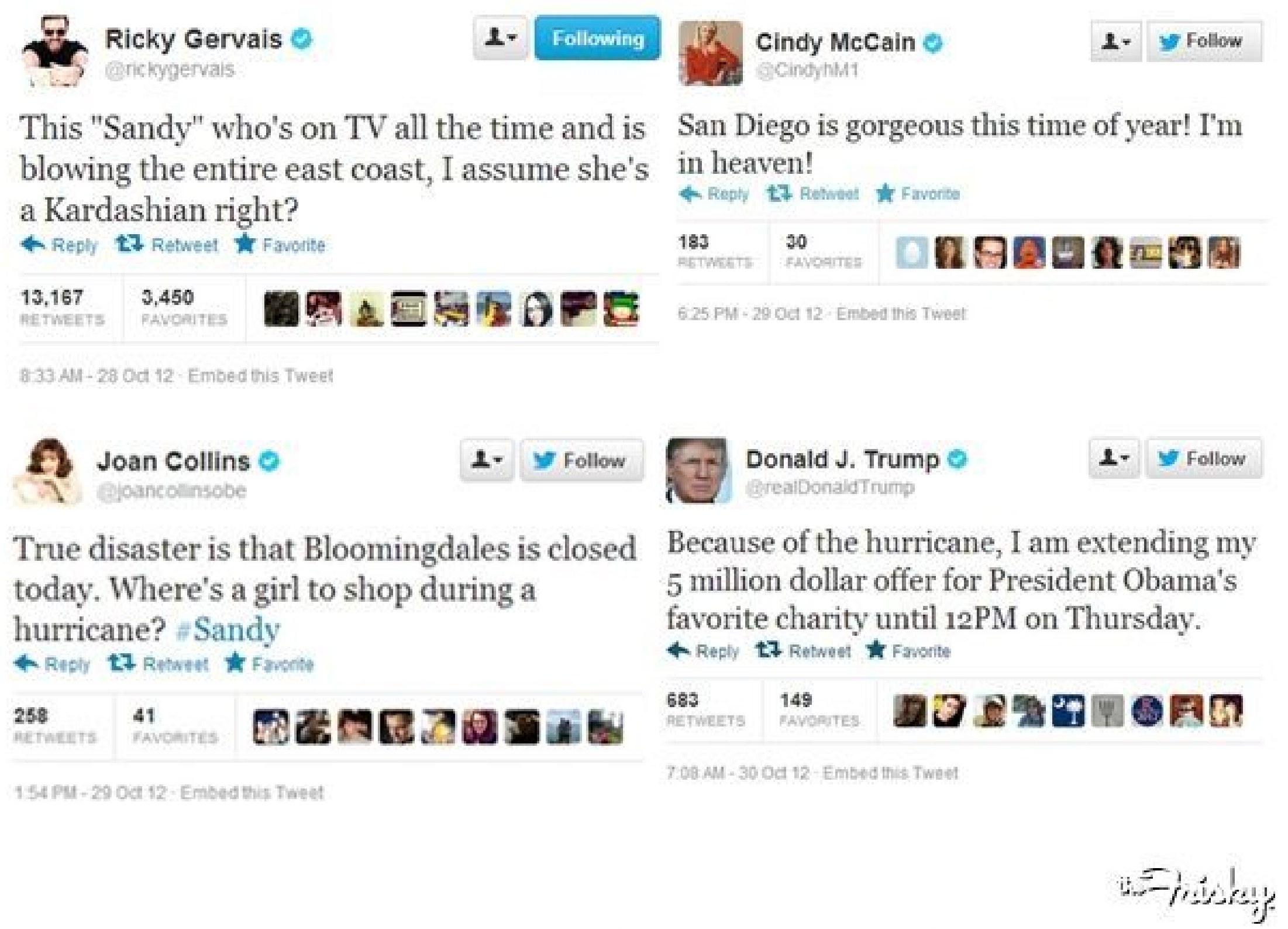
2 Method

Unsupervised learning and supervised learning are both applied to solve this problem.

First step is to train a classifier model using training dataset, i.e. tweets. The output is a list of weights for each word.

Second step is to run K-Mean algorithm to cluster the testing dataset by latitude and longitude.

Third step is to iterate each cluster and run classifier against each tweet to compute the relevance to an incident and overall percentage in that cluster to determine whether there is an incident.



Ricky Gervais @rickygervais
This "Sandy" who's on TV all the time and is blowing the entire east coast, I assume she's a Kardashian right?
13,167 RETWEETS 3,450 FAVORITES
8:33 AM - 28 Oct 12 · Embed this Tweet

Cindy McCain @CindyhM1
San Diego is gorgeous this time of year! I'm in heaven!
183 RETWEETS 30 FAVORITES
6:25 PM - 29 Oct 12 · Embed this Tweet

Joan Collins @joancollinsobe
True disaster is that Bloomingdales is closed today. Where's a girl to shop during a hurricane? #Sandy
258 RETWEETS 41 FAVORITES
1:54 PM - 29 Oct 12 · Embed this Tweet

Donald J. Trump @realDonaldTrump
Because of the hurricane, I am extending my 5 million dollar offer for President Obama's favorite charity until 12PM on Thursday.
683 RETWEETS 149 FAVORITES
7:08 AM - 30 Oct 12 · Embed this Tweet

Figure 2. Tweets during Hurrican Sandy

3 Results

1. Train classifier model

Input 104271 tweets that were geo-tagged and labeled.

For each tweet, extract features and compute the projected value and hinge loss.

Repeat 20 iterations with eta = 0.01, and then get a trained model which contains a list of weights for 102652 words.

2. K-mean cluster

Aggregate testing dataset which has 798 tweets based on longitude and latitude.

Start clustering from K=1, and then increase K by 1 for each re-cluster, until new centroid is close to an existing one.

Here when K = 3, centroids locate in New York City, Massachusetts, and Washington DC. And when K=4, new centroid locates again in New York City and clustering stops here.

3. Analyze testing dataset of tweets

Cluster 1(Elizabeth, NJ 07201):
a. Hurricane relevant: 535
b. Hurricane not relevant: 1
c. Hurricane possibility: 99.26%

Cluster 2(Garrett County, MD 21520):
a. Hurricane relevant: 106
b. Hurricane not relevant: 1
c. Hurricane possibility: 99.07%

Cluster 3:(Milford, MA 01757):
a. Hurricane relevant: 41
b. Hurricane not relevant: 2
c. Hurricane possibility: 95.35%

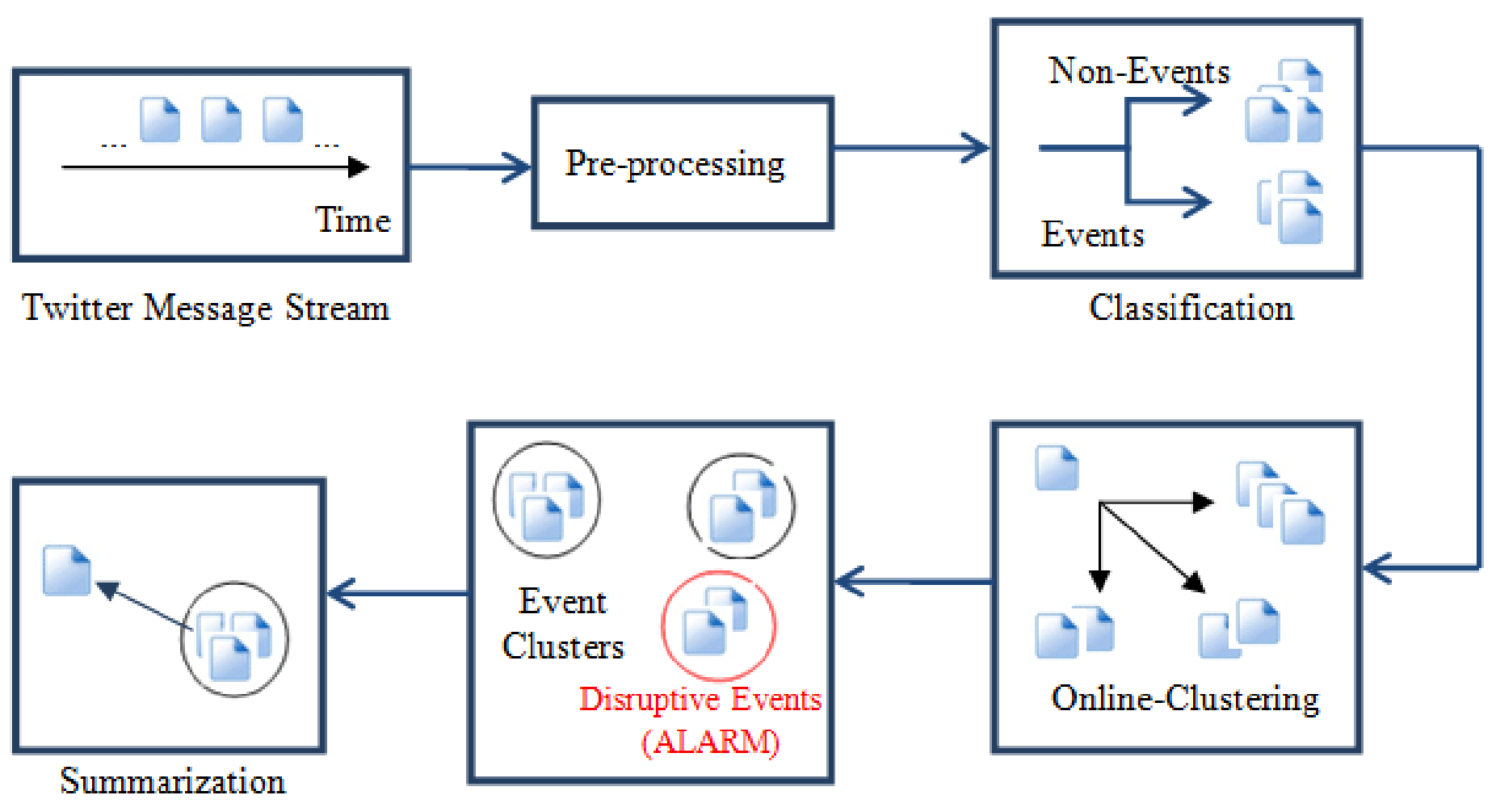


Figure 3. Workflow

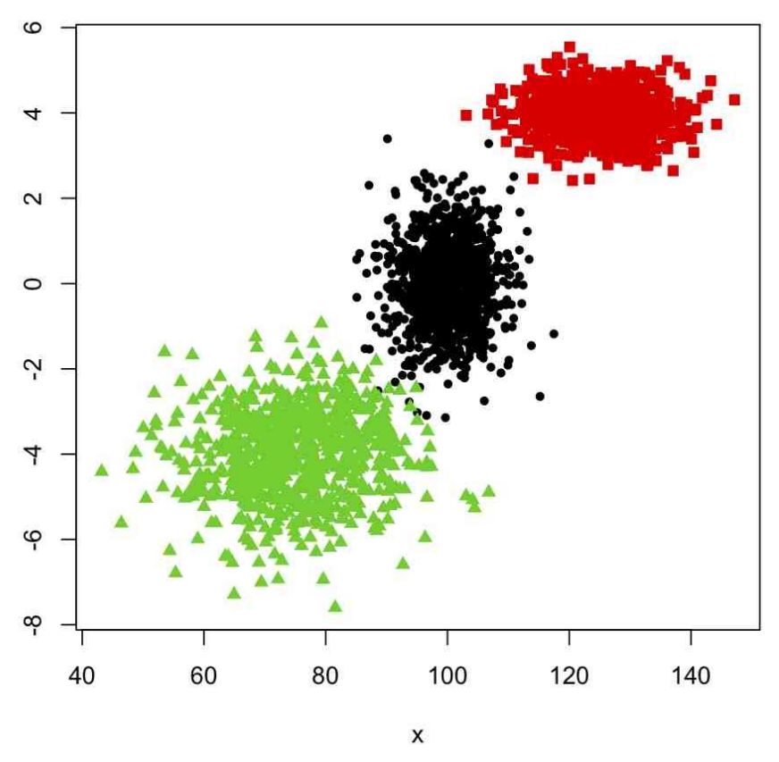
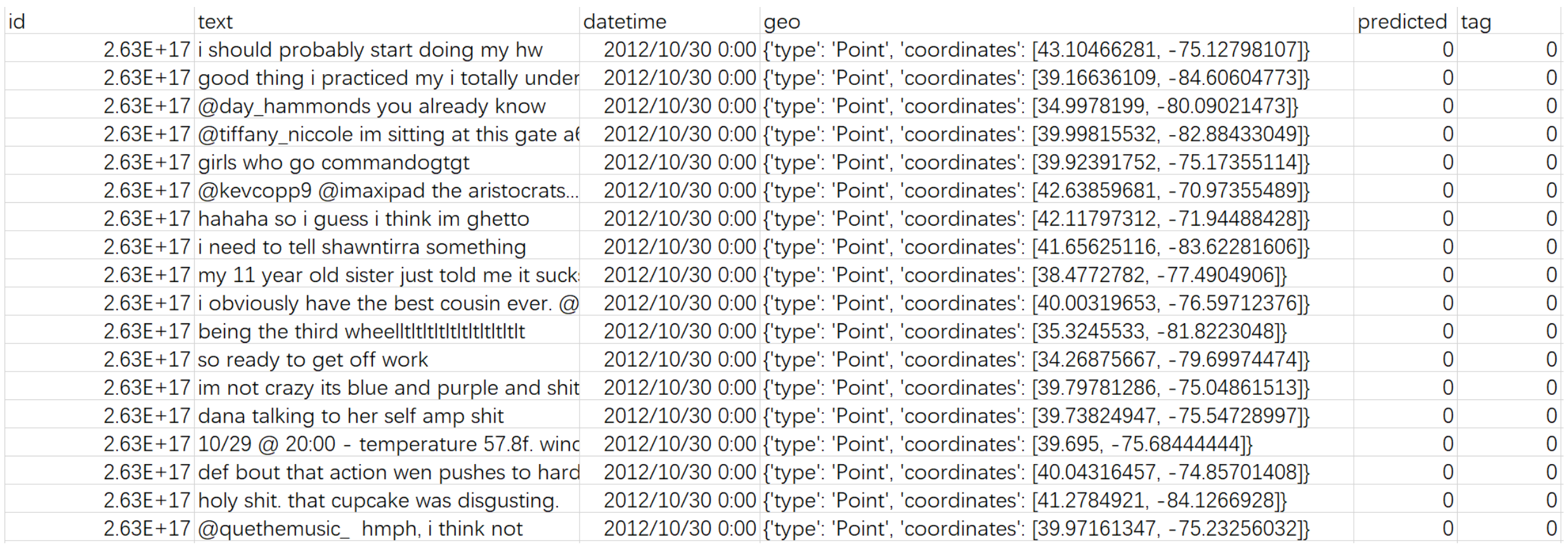
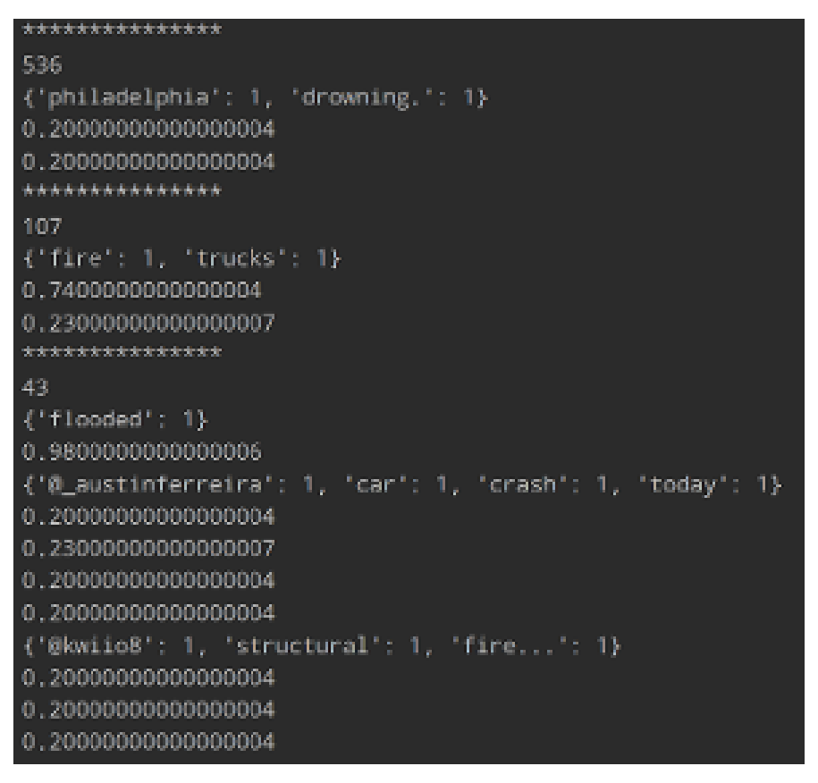


Figure 4. K-mean clustering



id	text	datetime	geo	predicted	tag
2.63E+17	i should probably start doing my hw	2012/10/30 0:00	{'type': 'Point', 'coordinates': [43.10466281, -75.12798107]}	0	0
2.63E+17	good thing i practiced my i totally under	2012/10/30 0:00	{'type': 'Point', 'coordinates': [39.16636109, -84.60604773]}	0	0
2.63E+17	@day_hammonds you already know	2012/10/30 0:00	{'type': 'Point', 'coordinates': [34.9978199, -80.09021473]}	0	0
2.63E+17	@tiffany_niccole im sitting at this gate at	2012/10/30 0:00	{'type': 'Point', 'coordinates': [39.99815532, -82.88433049]}	0	0
2.63E+17	girls who go commandogtgt	2012/10/30 0:00	{'type': 'Point', 'coordinates': [39.92391752, -75.17355114]}	0	0
2.63E+17	@kevcopp9 @imaxipad the aristocrats...	2012/10/30 0:00	{'type': 'Point', 'coordinates': [42.63859681, -70.97355489]}	0	0
2.63E+17	hahaha so i guess i think im ghetto	2012/10/30 0:00	{'type': 'Point', 'coordinates': [42.11797312, -71.94488428]}	0	0
2.63E+17	i need to tell shawntirra something	2012/10/30 0:00	{'type': 'Point', 'coordinates': [41.65625116, -83.62281606]}	0	0
2.63E+17	my 11 year old sister just told me it suck	2012/10/30 0:00	{'type': 'Point', 'coordinates': [38.4772782, -77.4904906]}	0	0
2.63E+17	i obviously have the best cousin ever. @	2012/10/30 0:00	{'type': 'Point', 'coordinates': [40.00319653, -76.59712376]}	0	0
2.63E+17	being the third wheelltltltltltltltltltltlt	2012/10/30 0:00	{'type': 'Point', 'coordinates': [35.3245533, -81.8223048]}	0	0
2.63E+17	so ready to get off work	2012/10/30 0:00	{'type': 'Point', 'coordinates': [34.26875667, -79.69974474]}	0	0
2.63E+17	im not crazy its blue and purple and shit	2012/10/30 0:00	{'type': 'Point', 'coordinates': [39.79781286, -75.04861513]}	0	0
2.63E+17	dana talking to her self amp shit	2012/10/30 0:00	{'type': 'Point', 'coordinates': [39.73824947, -75.54728997]}	0	0
2.63E+17	10/29 @ 20:00 - temperature 57.8f. winc	2012/10/30 0:00	{'type': 'Point', 'coordinates': [39.695, -75.68444444]}	0	0
2.63E+17	def bout that action wen pushes to hard	2012/10/30 0:00	{'type': 'Point', 'coordinates': [40.04316457, -74.85701408]}	0	0
2.63E+17	holy shit. that cupcake was disgusting.	2012/10/30 0:00	{'type': 'Point', 'coordinates': [41.2784921, -84.1266928]}	0	0
2.63E+17	@quethemusic_ hmph, i think not	2012/10/30 0:00	{'type': 'Point', 'coordinates': [39.97161347, -75.23256032]}	0	0

Figure 5. Dataset



```
***************
536
{'philadelphia': 1, 'drowning.': 1}
0.2000000000000004
0.2000000000000004
***************
107
{'fire': 1, 'trucks': 1}
0.7400000000000004
0.2300000000000007
***************
43
{'flooded': 1}
0.9800000000000006
{'@_austinferreira': 1, 'car': 1, 'crash': 1, 'today': 1}
0.2000000000000004
0.2300000000000007
0.2000000000000004
0.2000000000000004
{'@kwiio8': 1, 'structural': 1, 'fire...': 1}
0.2000000000000004
0.2000000000000004
0.2000000000000004
```

Figure 6. Result

4 Conclusion

The result is very promising that not only the confidence of incident, e.g. hurricane, is very high, but also it actually shows the path of this hurricane.

The model is built upon statistic frequency and relevance which proved to be more accurate than keyword frequency or indirect indicators such as tweet length, sentiment, or punctuation. Also, with over 100k training tweets for this supervised learning process, the model demonstrates a very high effectiveness.

Unsupervised learning, i.e. clustering, shows a clear path of this hurricane. However, given the small size of testing dataset which are 798 tweets, future work needs to be done to dynamically prioritize the clusters to skip un-categorized or non-significant ones which will also improve the system performance in general.

Overall, the project turns out to be a successful experiment to utilize AI and social media data to detect real time incident aggregated by geolocation.

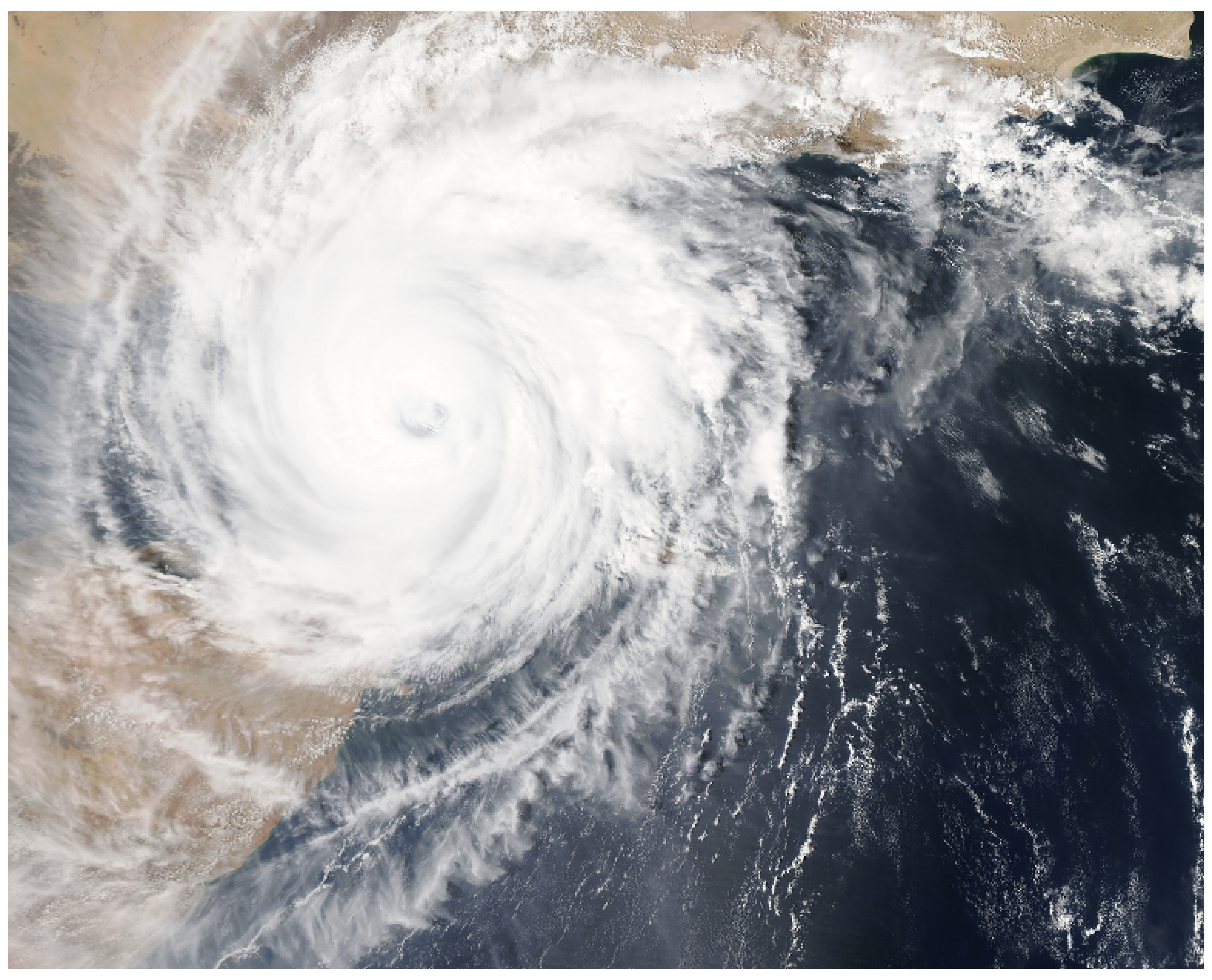


Figure 7. Hurricane

⚠ EMERGENCY ALERT ✕

Hurricane Warning this area. Check local media and authorities. -NWS

Settings

Figure 8. Hurricane Alert