Leveraging Social Media to Map Natural Disasters

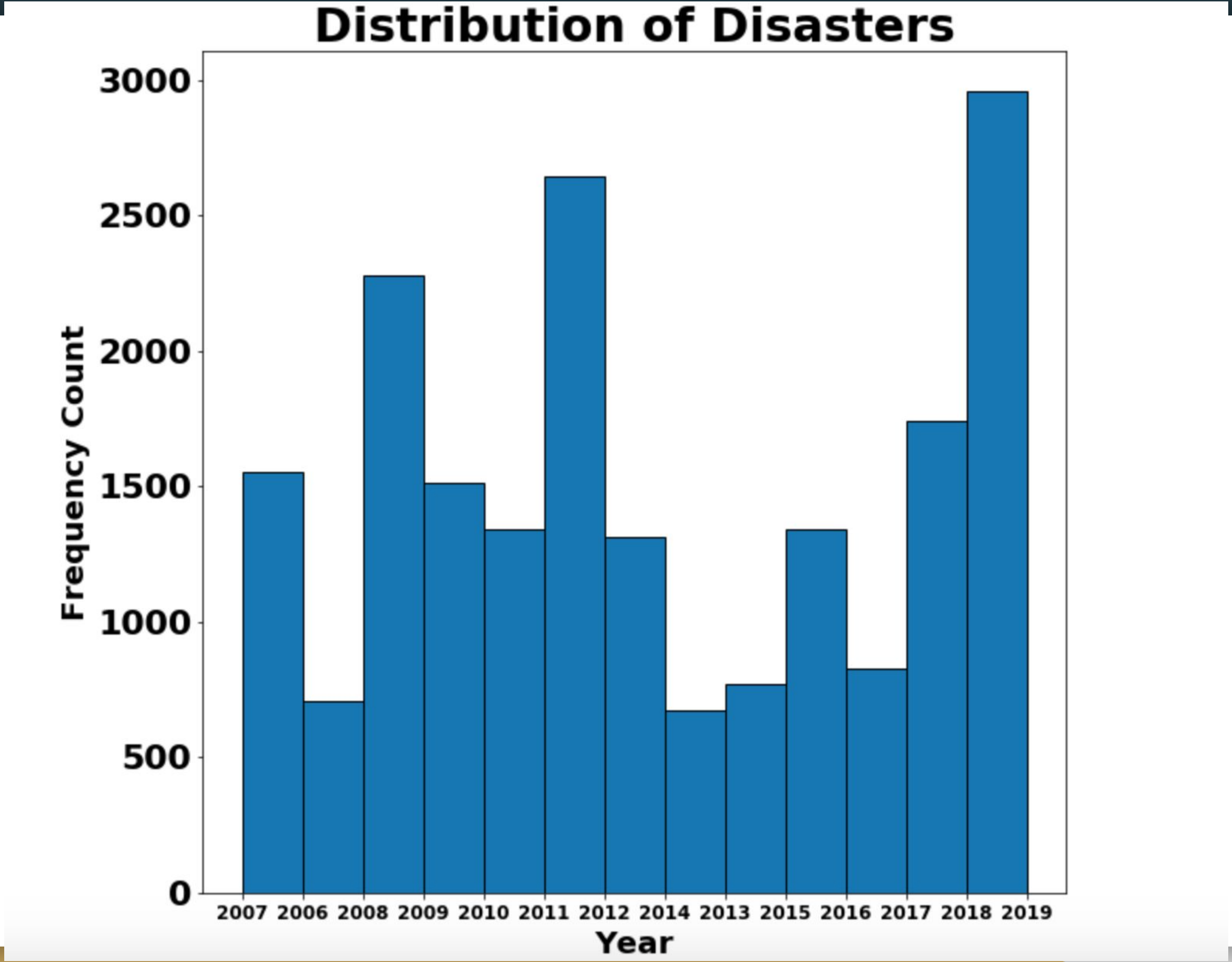
Presented by Grace Powell, Andrea Yoss, Dimitri Kisten

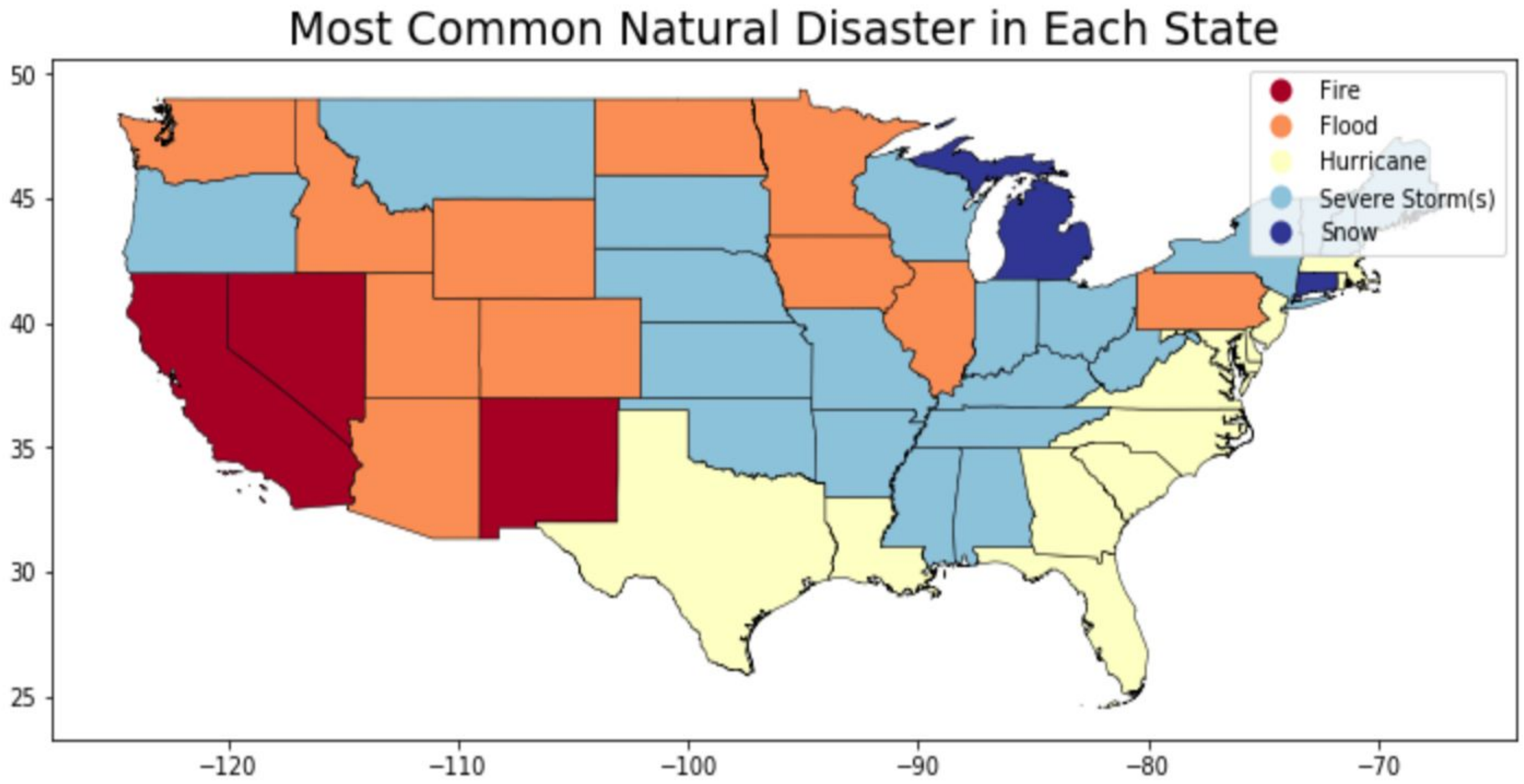
Problem Statement

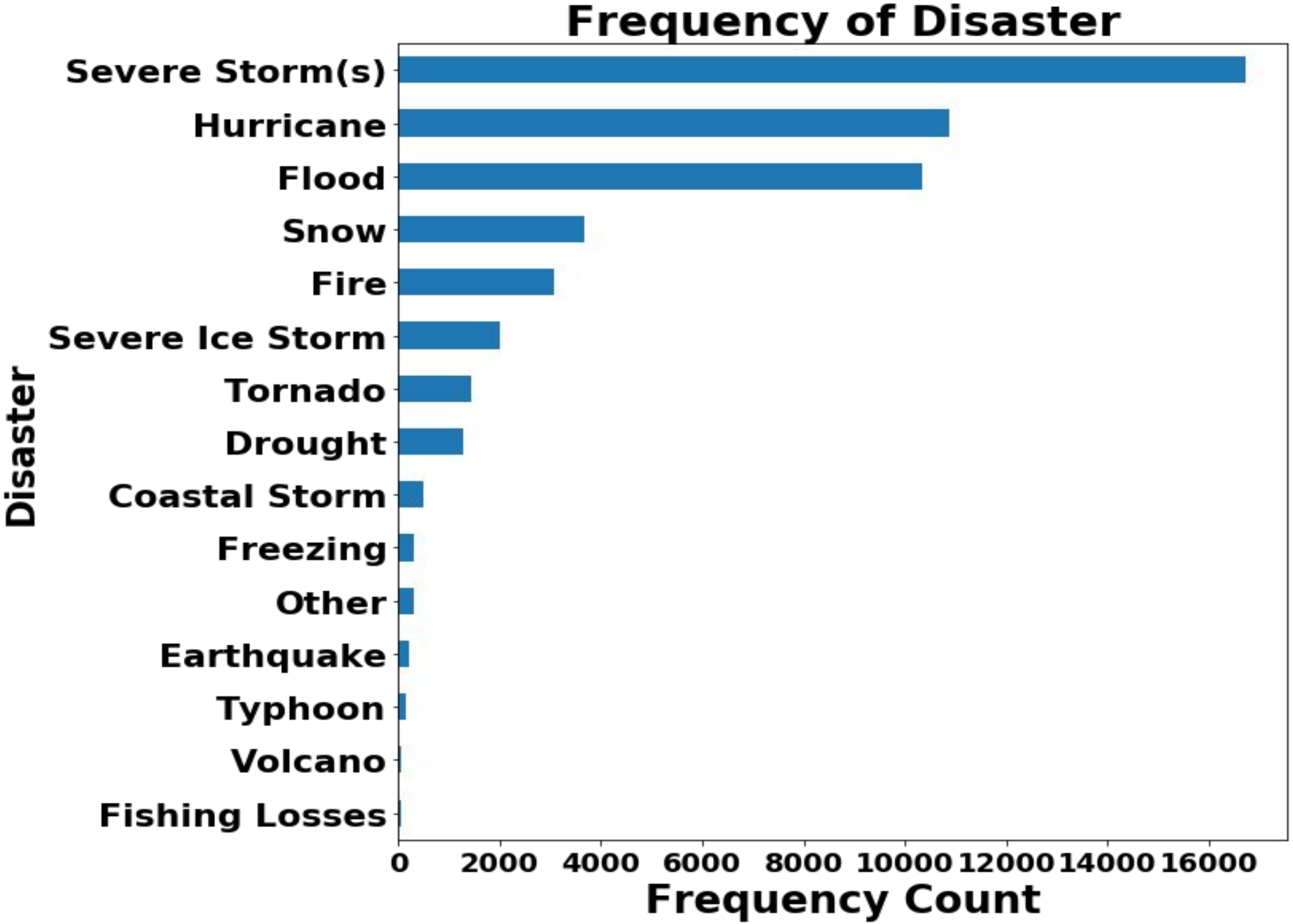
Social media, specifically Twitter, can be leveraged to accurately detect and map various types of natural disasters. We used Twitter combined with the Open FEMA dataset identify legitimate Floods using Natural Language Processing.

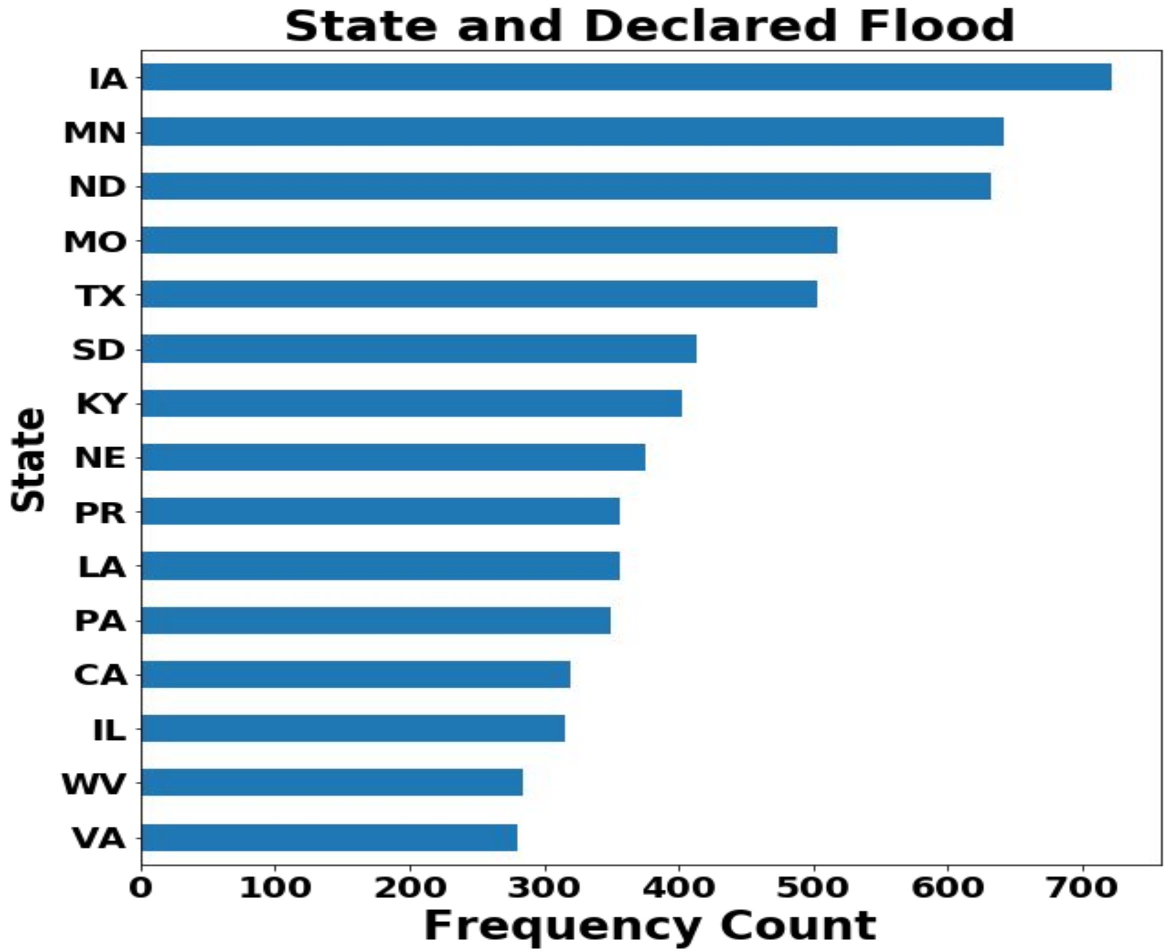
Why FEMA ?

- FEMA Disaster Declarations Summary is a dataset that stated all federally declared disasters.
- Dataset contained information on the location, disaster type, begin date and declaration date.
- Used FEMA data compare to the twitter dates scrapped that contained titles "Severe Flood" to verify whether the tweets were accurate


Distribution of Disasters
Frequency Count
3000
2500
2000
1500
1000
500
0
2007
2006
2008
2009
2010
2011
2012
2014
2013
2015
2016
2017
2018
2019
Year

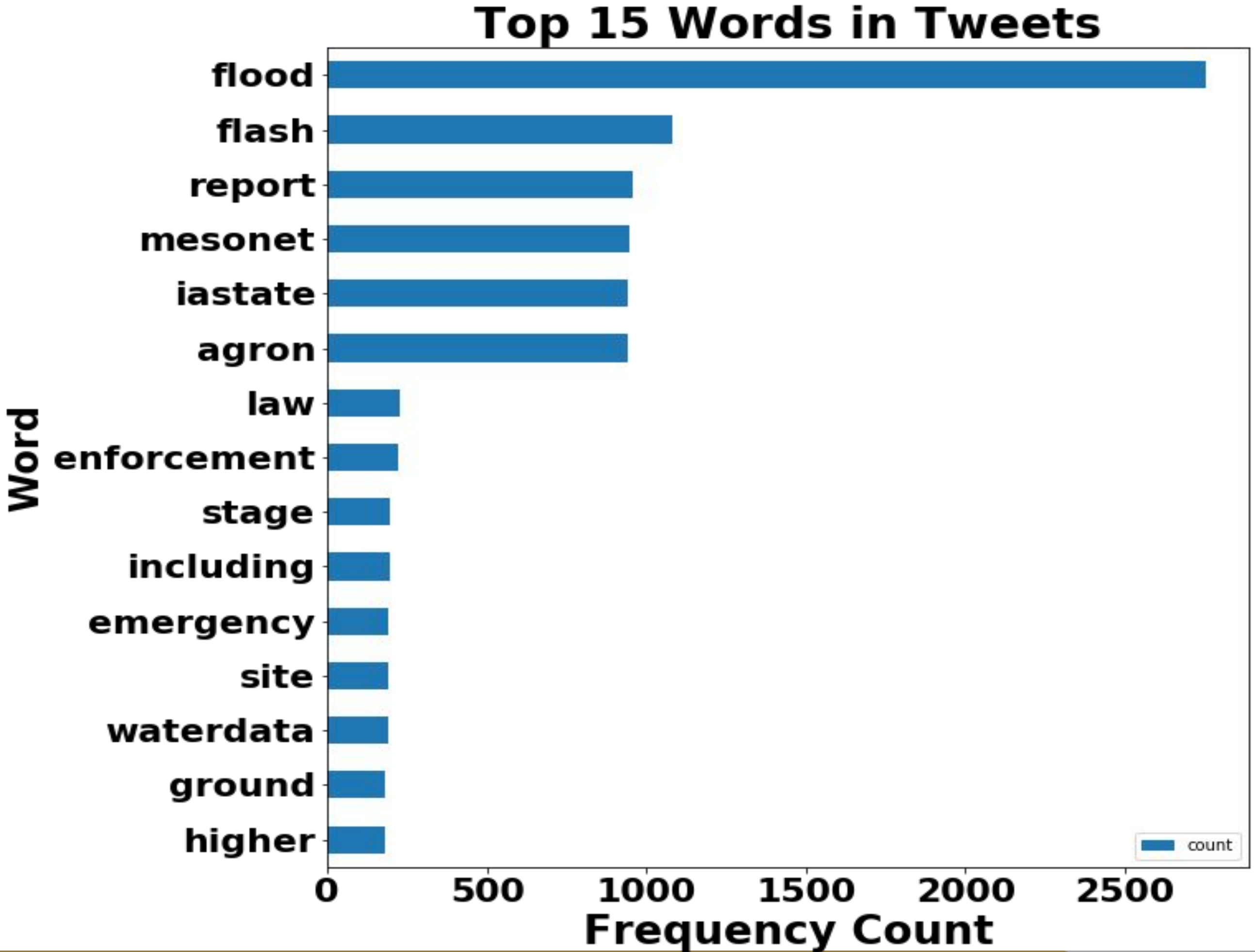

Most Common Natural Disaster in Each State
Fire
Flood
Hurricane
Severe Storm(s)
Snow
50
45
40
35
30
25
−120
−110
−100
−90
−80
−70

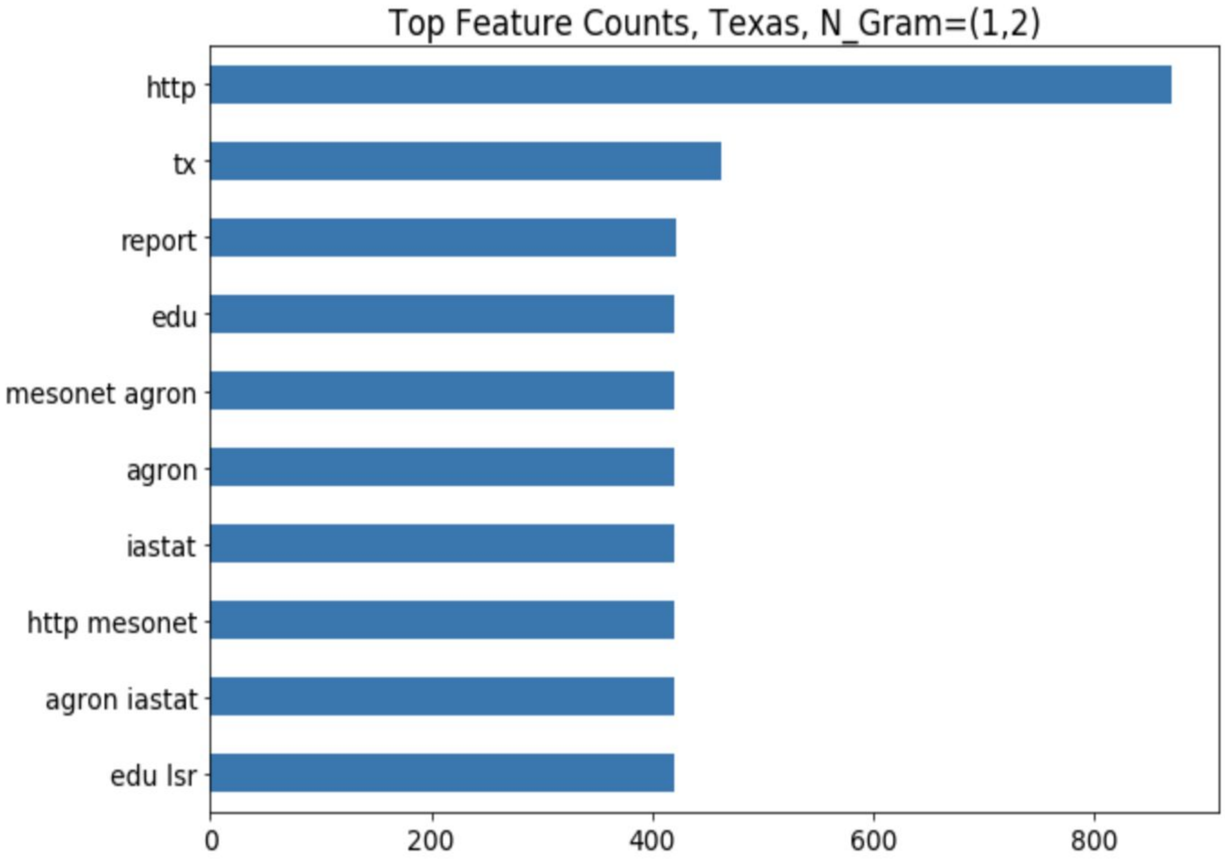

Frequency of Disaster
Severe Storm(s)
Hurricane
Flood
Snow
Fire
Severe Ice Storm
Tornado
Drought
Coastal Storm
Freezing
Other
Earthquake
Typhoon
Volcano
Fishing Losses
Disaster
0
2000
4000
6000
8000
10000
12000
14000
16000
Frequency Count

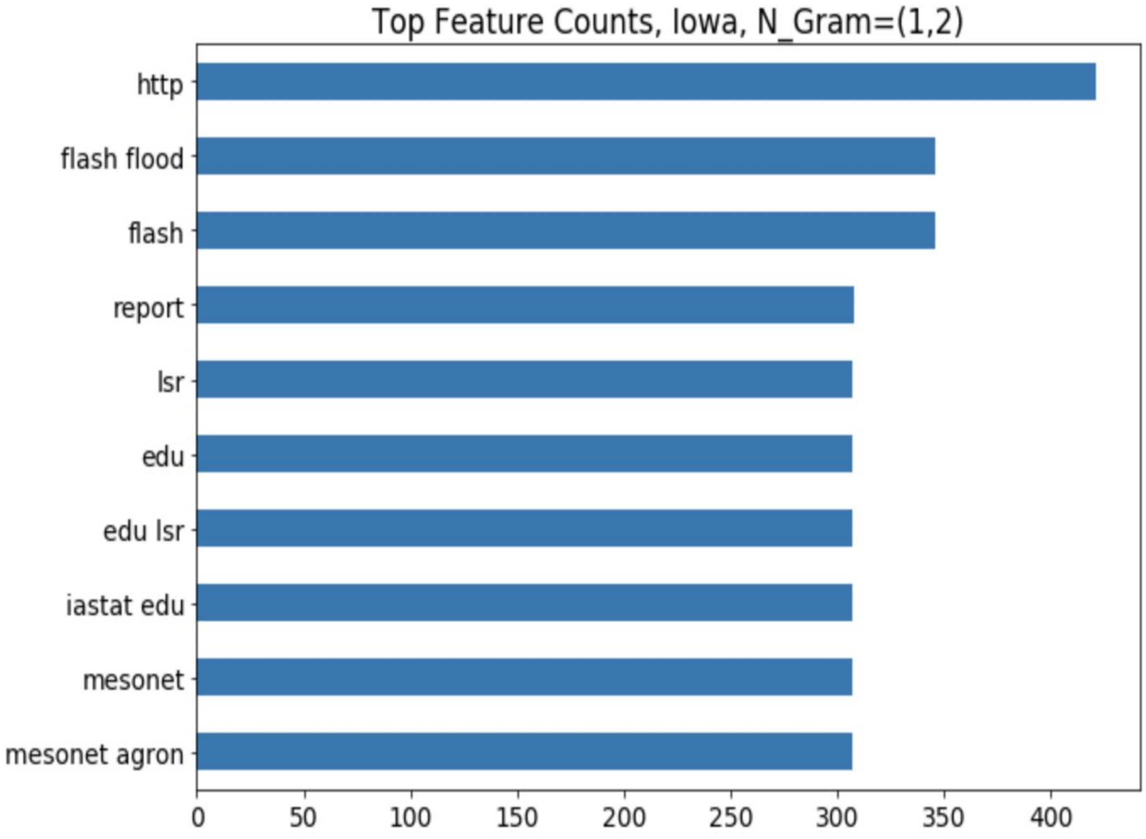

State and Declared Flood
State
IA
MN
ND
MO
TX
SD
KY
NE
PR
LA
PA
CA
IL
WV
VA
0
100
200
300
400
500
600
700
Frequency Count

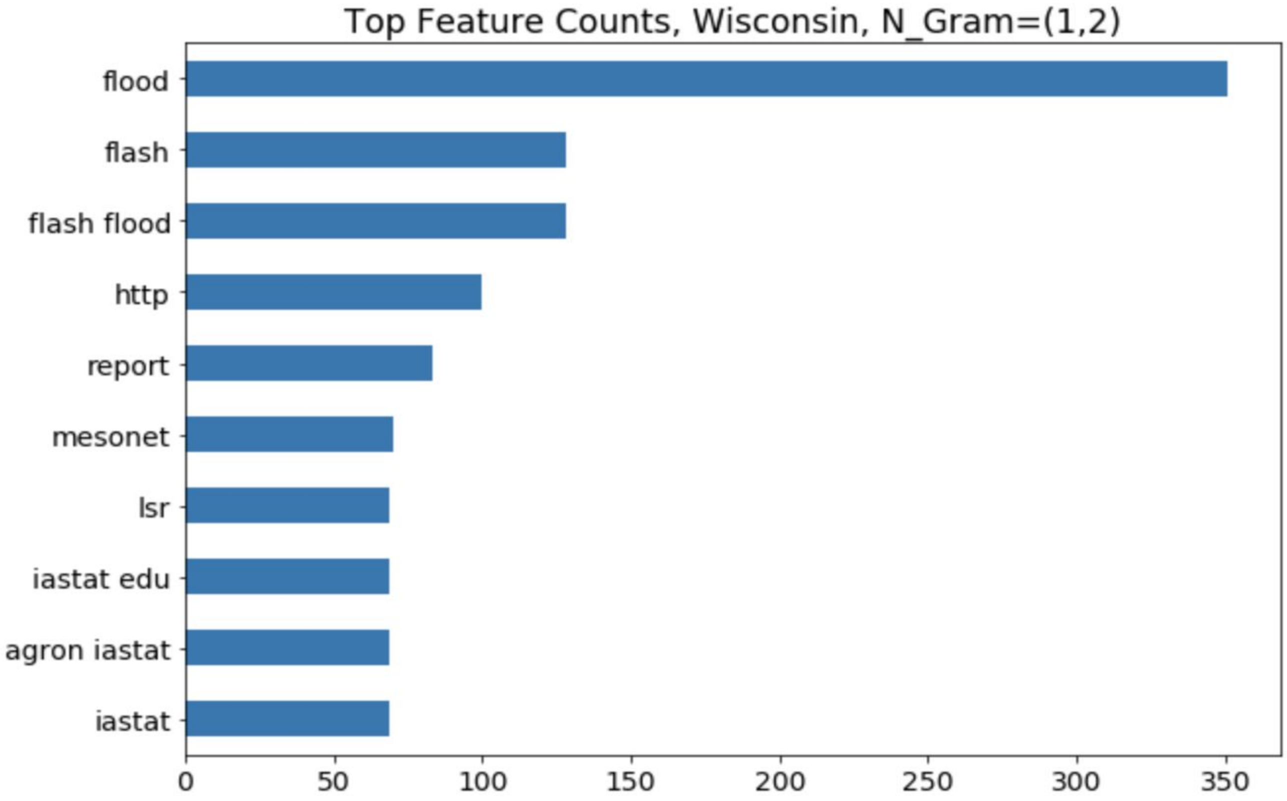
Data Acquisition and EDA Process

- GetOldTweets3 Scraper
    - Iowa, Wisconsin, Texas
    - <2015
    - Two distinct time periods
- Supervised Learning Problem
- NLP

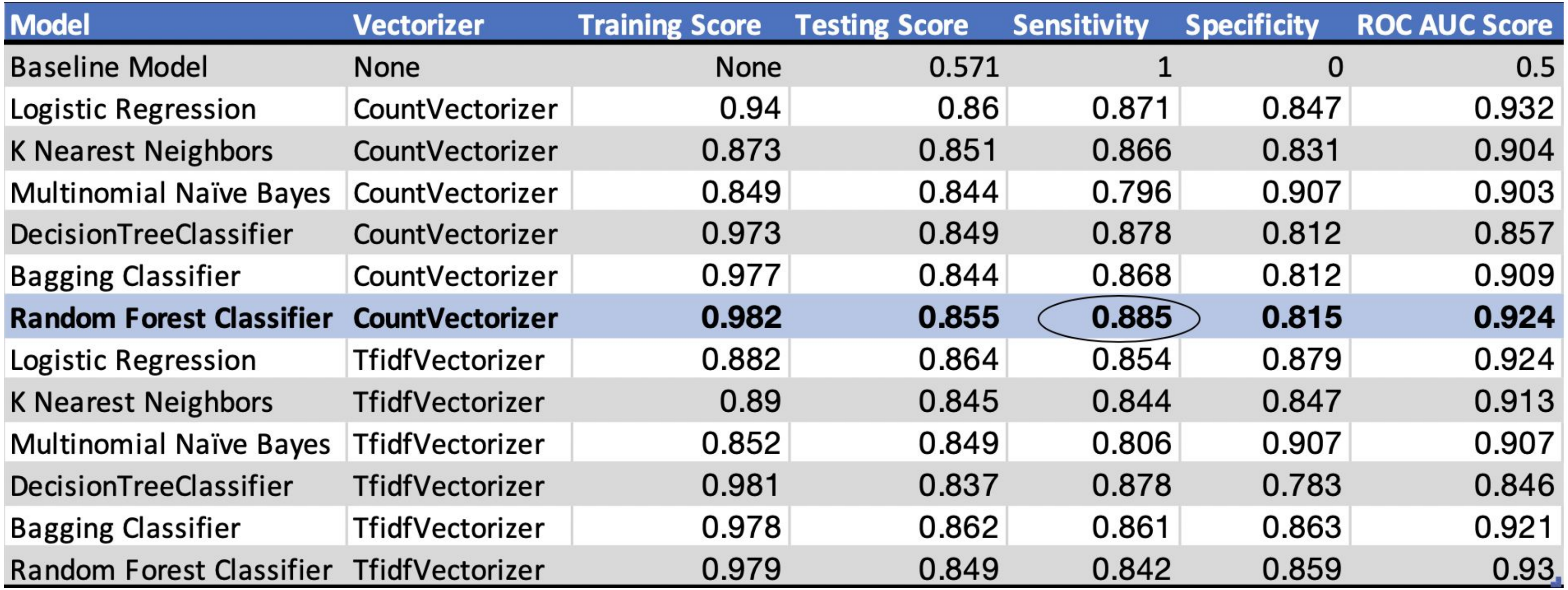

Top 15 Words in Tweets
flood
flash
report
mesonet
iastate
agron
law
enforcement
stage
including
emergency
site
waterdata
ground
higher
Word
count
0
500
1000
1500
2000
2500
Frequency Count


Top Feature Counts, Texas, N_Gram=(1,2)
http
tx
report
edu
mesonet agron
agron
iastat
http mesonet
agron iastat
edu lsr
0
200
400
600
800


Top Feature Counts, Iowa, N_Gram=(1,2)
http
flash flood
flash
report
lsr
edu
edu lsr
iastat edu
mesonet
mesonet agron
0
50
100
150
200
250
300
350
400

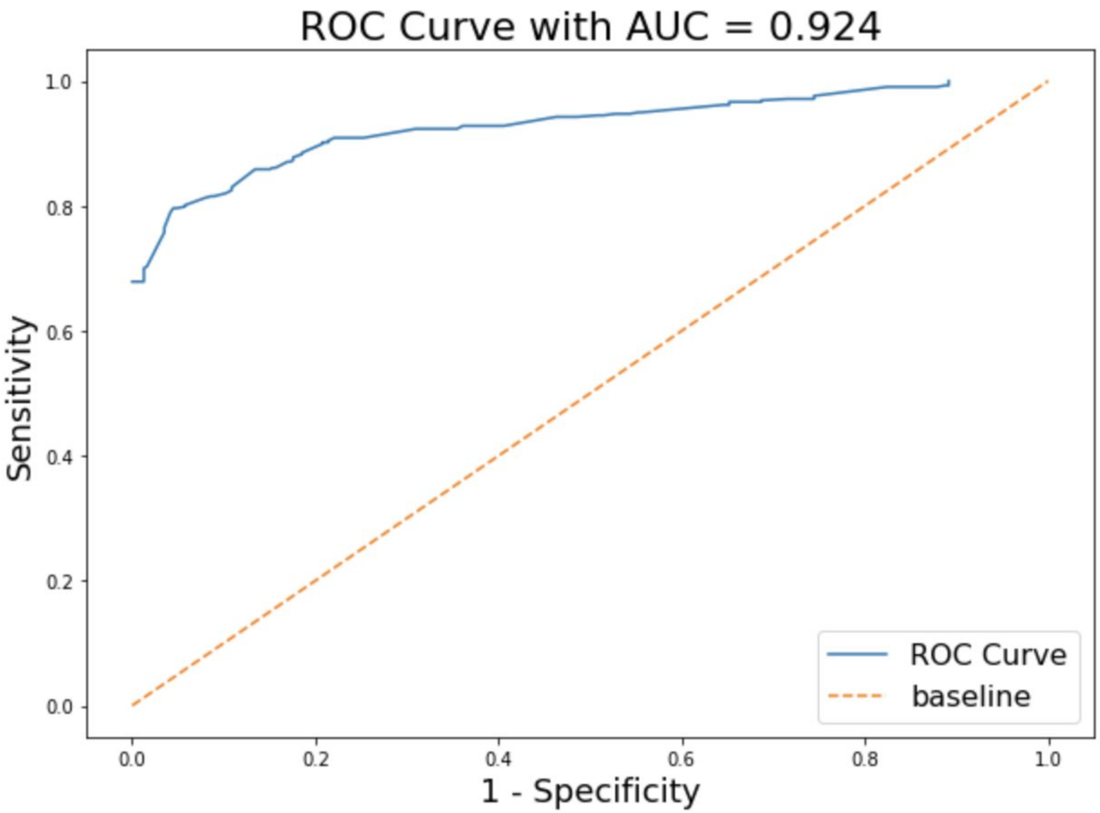

Top Feature Counts, Wisconsin, N_Gram=(1,2)
flood
flash
flash flood
http
report
mesonet
lsr
iastat edu
agron iastat
iastat
0
50
100
150
200
250
300
350

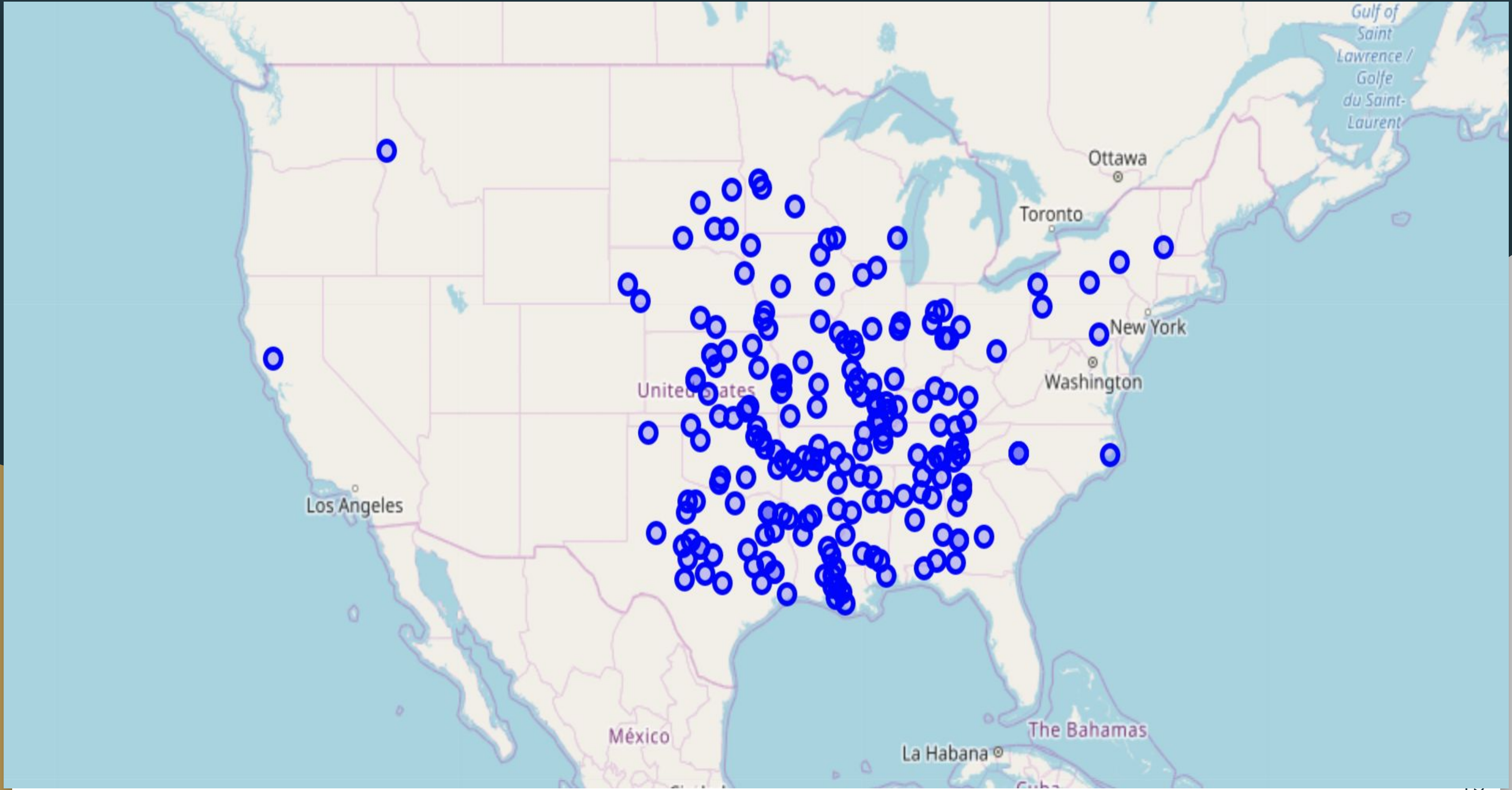
Model Selection

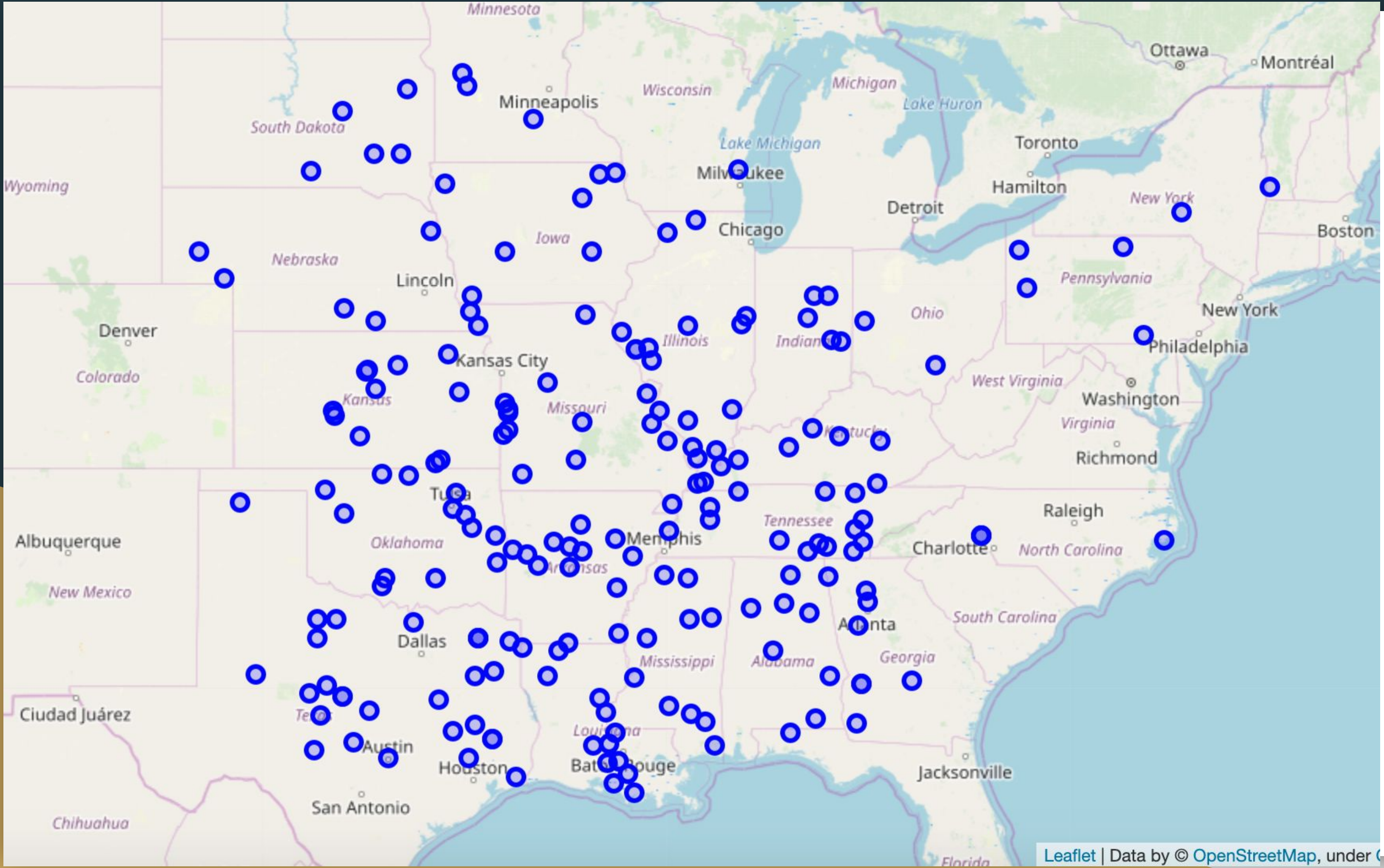


Model	Vectorizer	Training Score	Testing Score	Sensitivity	Specificity	ROC AUC Score
Baseline Model	None	None	0.571	1	0	0.5
Logistic Regression	CountVectorizer	0.94	0.86	0.871	0.847	0.932
K Nearest Neighbors	CountVectorizer	0.873	0.851	0.866	0.831	0.904
Multinomial Naïve Bayes	CountVectorizer	0.849	0.844	0.796	0.907	0.903
DecisionTreeClassifier	CountVectorizer	0.973	0.849	0.878	0.812	0.857
Bagging Classifier	CountVectorizer	0.977	0.844	0.868	0.812	0.909
Random Forest Classifier	**CountVectorizer**	**0.982**	**0.855**	**0.885**	**0.815**	**0.924**
Logistic Regression	TfidfVectorizer	0.882	0.864	0.854	0.879	0.924
K Nearest Neighbors	TfidfVectorizer	0.89	0.845	0.844	0.847	0.913
Multinomial Naïve Bayes	TfidfVectorizer	0.852	0.849	0.806	0.907	0.907
DecisionTreeClassifier	TfidfVectorizer	0.981	0.837	0.878	0.783	0.846
Bagging Classifier	TfidfVectorizer	0.978	0.862	0.861	0.863	0.921
Random Forest Classifier	TfidfVectorizer	0.979	0.849	0.842	0.859	0.93

ROC / AUC Curve




Gulf of Saint Lawrence / Golfe du Saint-Laurent
Ottawa
Toronto
New York
Washington
United States
Los Angeles
México
The Bahamas
La Habana
Cuba


Minnesota
Ottawa
Montréal
Minneapolis
Wisconsin
Michigan
Lake Huron
South Dakota
Lake Michigan
Toronto
Wyoming
Milwaukee
Hamilton
New York
Detroit
Boston
Iowa
Chicago
Nebraska
Pennsylvania
Lincoln
Ohio
New York
Denver
Illinois
Indiana
Philadelphia
Kansas City
Colorado
West Virginia
Washington
Kansas
Missouri
Virginia
Kentucky
Richmond
Tulsa
Raleigh
Tennessee
Albuquerque
Memphis
Oklahoma
Charlotte
North Carolina
Arkansas
New Mexico
South Carolina
Atlanta
Dallas
Mississippi
Alabama
Georgia
Ciudad Juárez
Texas
Louisiana
Austin
Houston
Baton Rouge
Jacksonville
San Antonio
Chihuahua
Florida
Leaflet | Data by © OpenStreetMap, under

Conclusion

- State by state models vs nationwide performance.
- False positives vs. False negative.

Limitations and Next Steps

- Floods are isolated to specific regions
- Twitter scraper limitations on location data
- Limited # of users who tweet about natural disasters.
- Limited availability of user data
- Potential Fixes:
  - Using available associate or ‘friend’ as a fix.
  - NLP on tweets/likes of a user