Mapping Natural Disasters Using Social Media

By Andrea Yoss, Grace Powell, & Dimitri Kisten

01

INTRODUCTION

Problem statement & FEMA Dataset EDA

02

Data Acquisition & EDA

Data acquisition process and preliminary findings.

03

Modelling & Mapping

Validating tweets and mapping our data

04

CONCLUSIONS

Findings, limitations, and next steps

01

Introduction

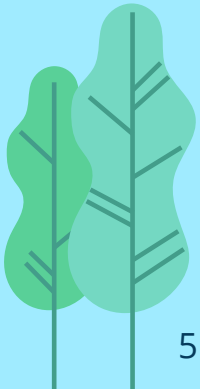
Problem statement and FEMA Dataset EDA

Problem Statement

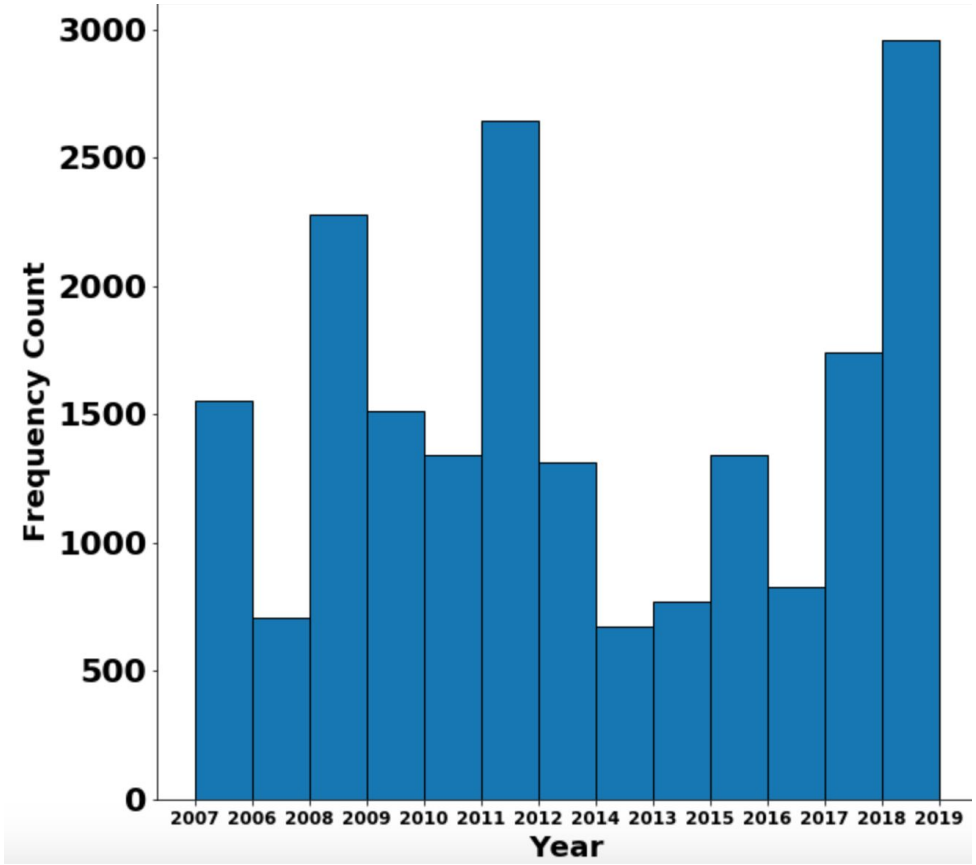
Social media, specifically Twitter, can be leveraged to accurately detect and map various types of natural disasters. We used Twitter combined with the Open FEMA dataset to identify legitimate floods from tweets using Natural Language Processing and then map those validated tweets.

Why the FEMA Dataset?

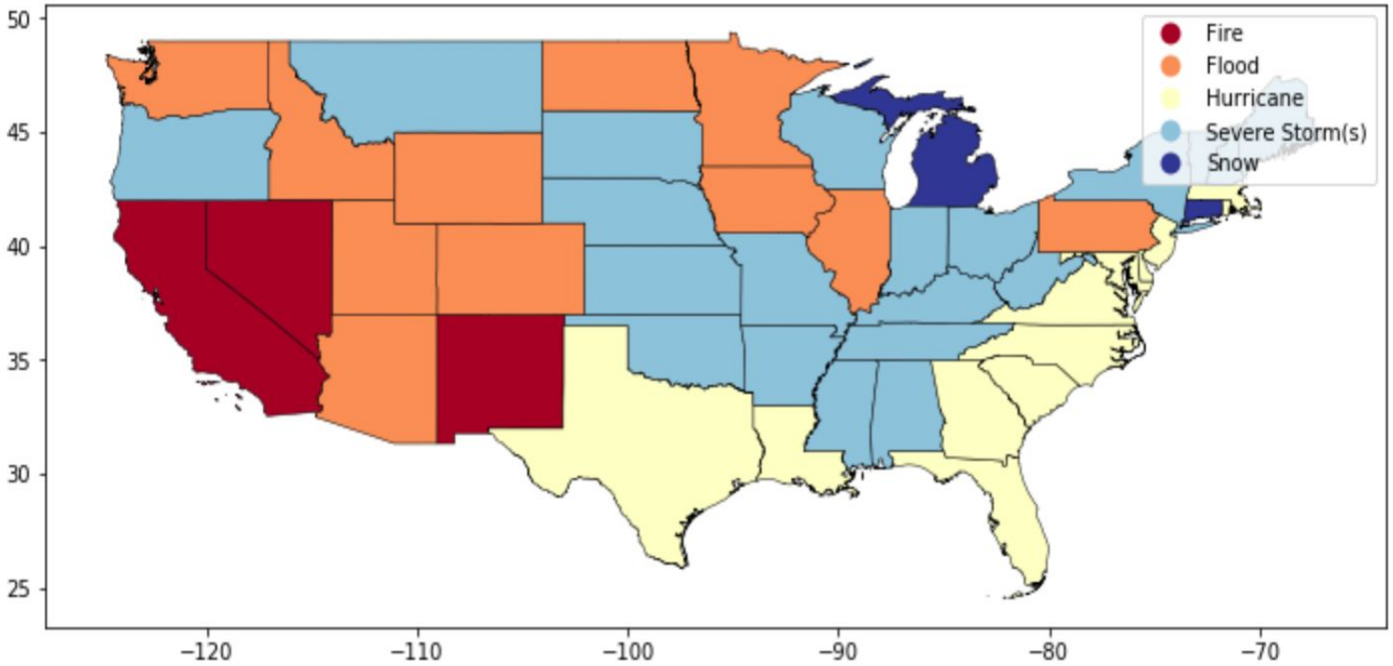
- FEMA Disaster Declarations Summary is a dataset that stated all federally declared disasters.
- Dataset contained information on the location, disaster type, begin date and declaration date.
- Used FEMA data compare to the twitter dates scrapped that contained titles "Severe Flood" to verify whether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



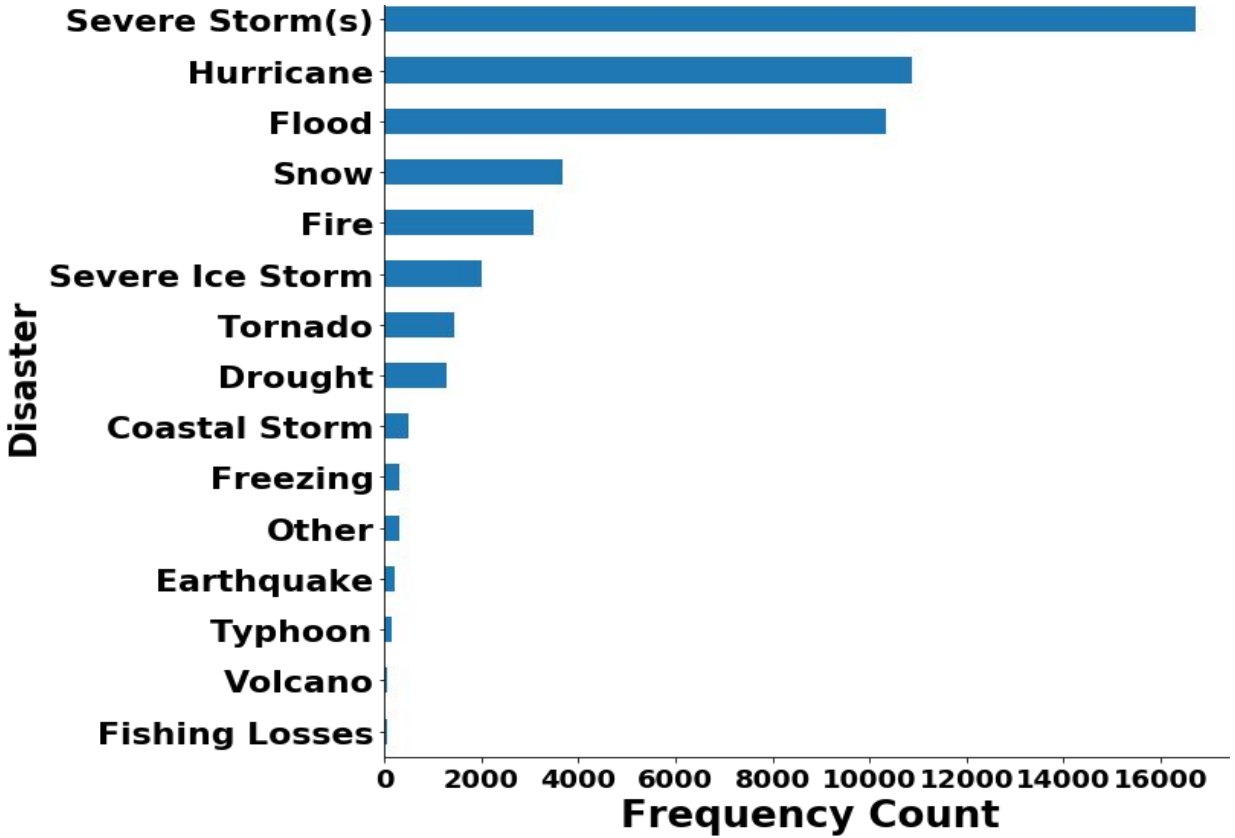
Total Disasters by Year



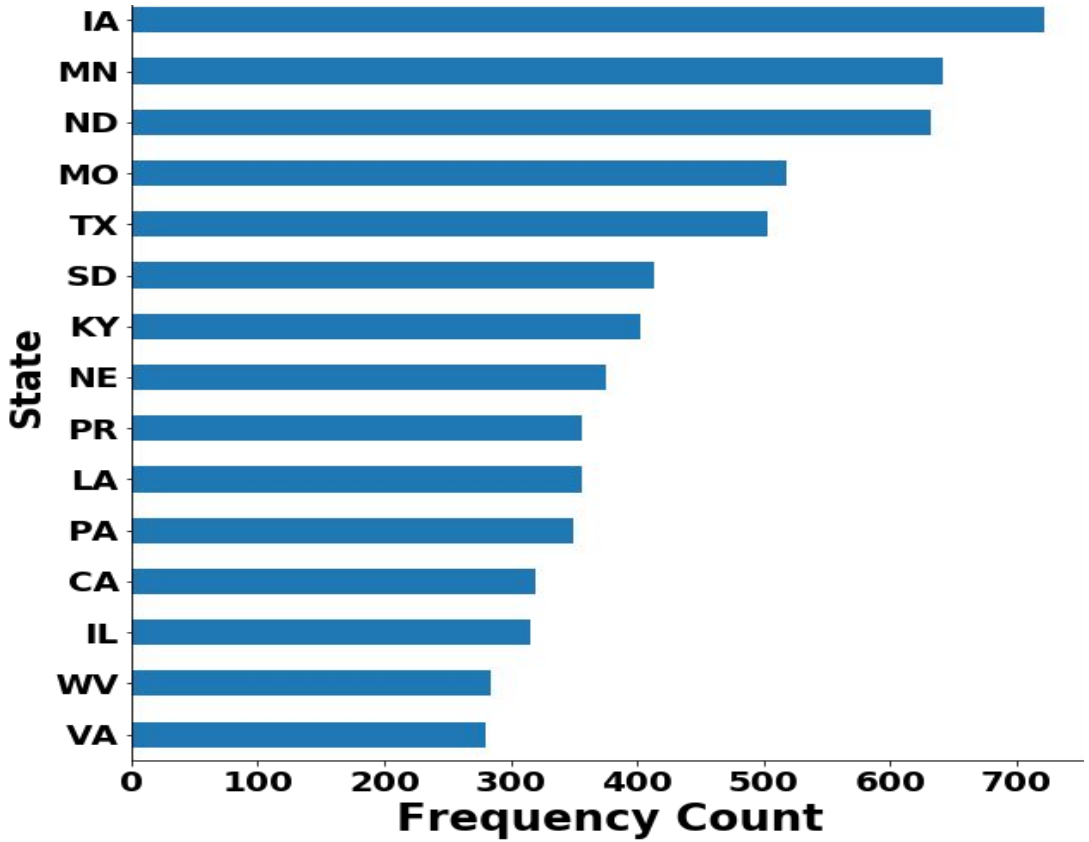
Most Common Natural Disaster by State



Disaster Frequency



Flood Frequency by State



2

...

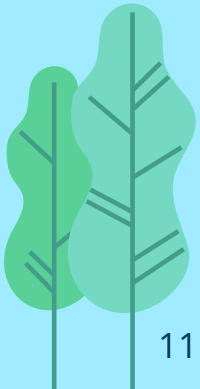
Data Acquisition & EDA

Gathering Tweets and preliminary findings

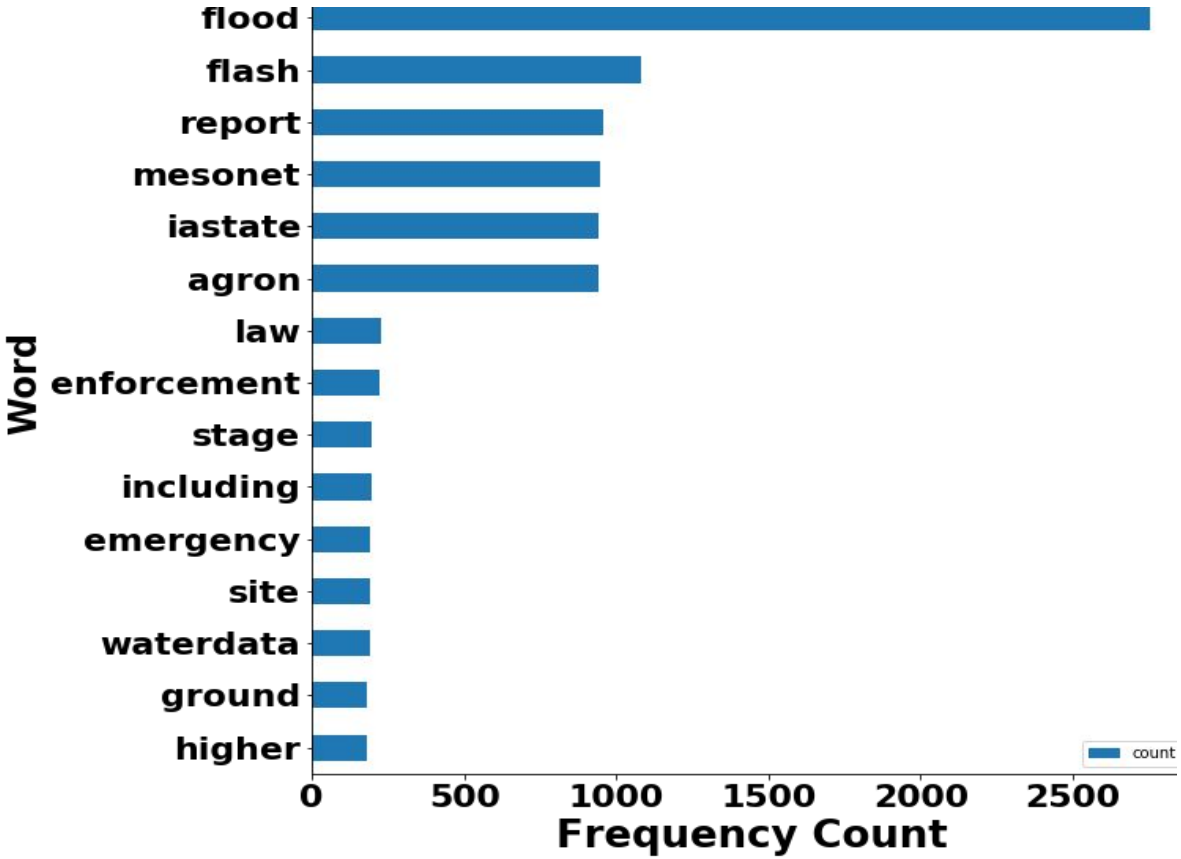
Data Acquisition & EDA Process

- GetOldTweets3 Scraper
  - Locations: Iowa, Wisconsin, Texas
  - Timeframe: > 2015
  - Two distinct time periods, during a flood and no flood.

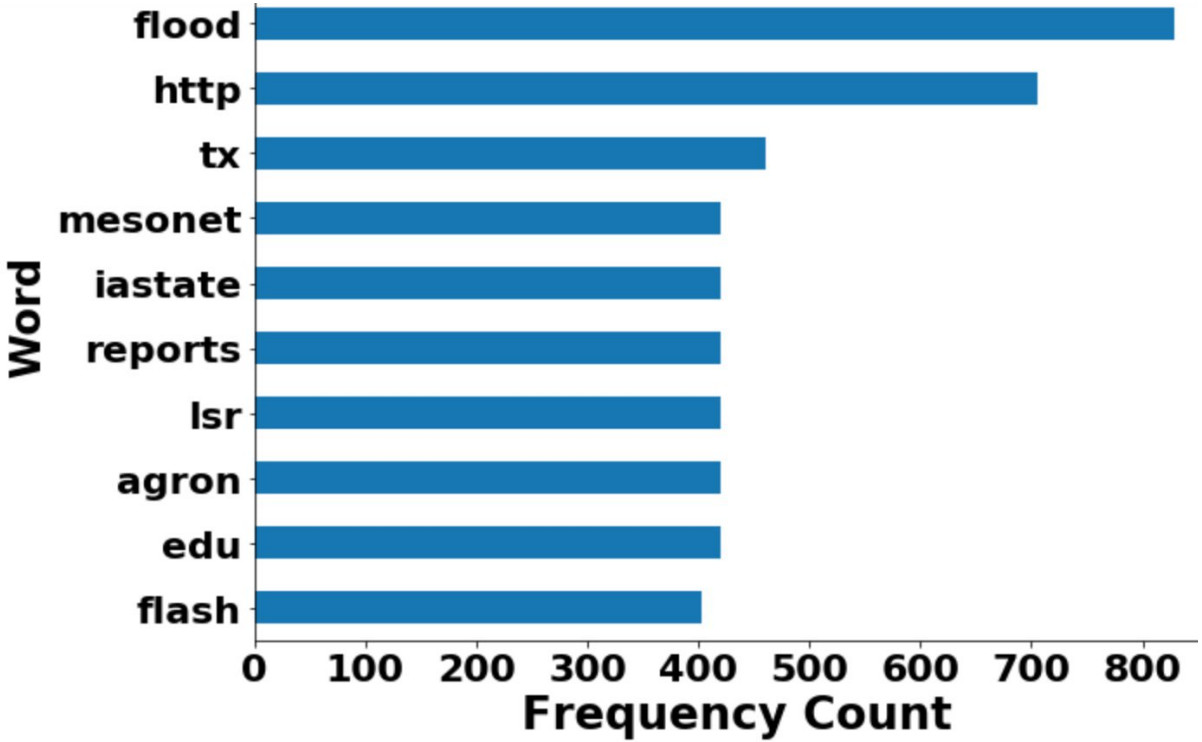
- Supervised Learning Problem
- Natural Language Processing (NLP)



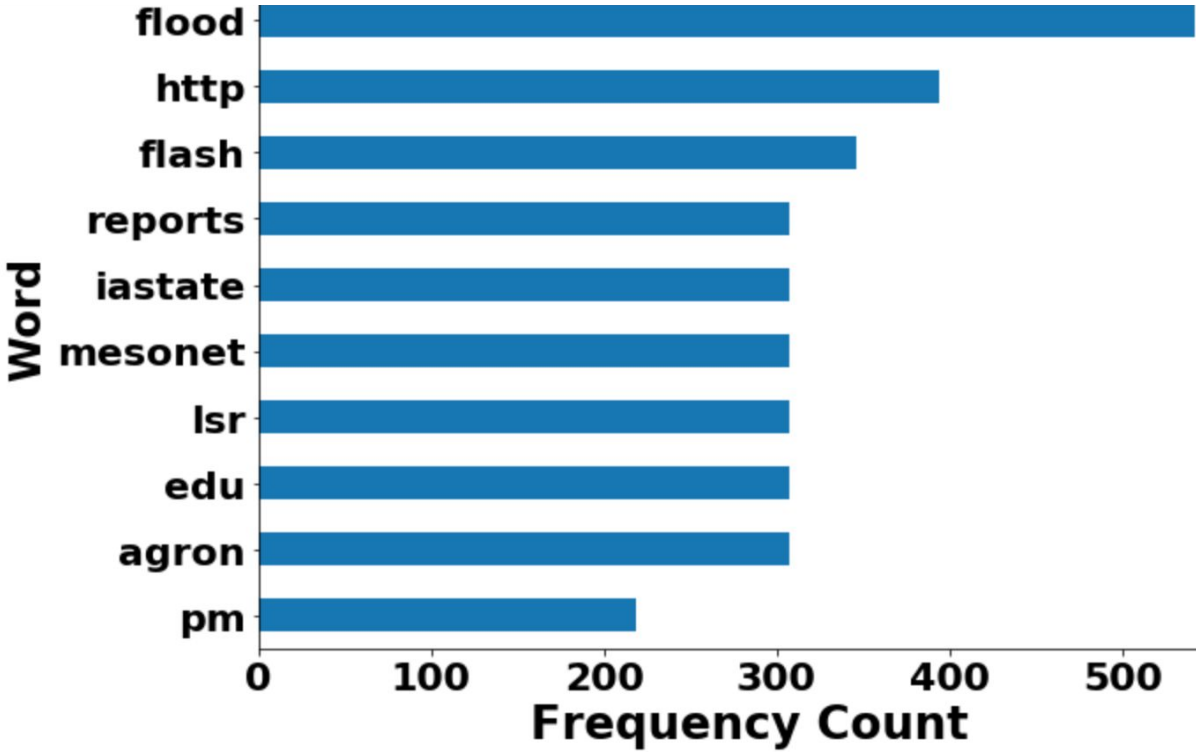
Most Frequent Words in Tweets



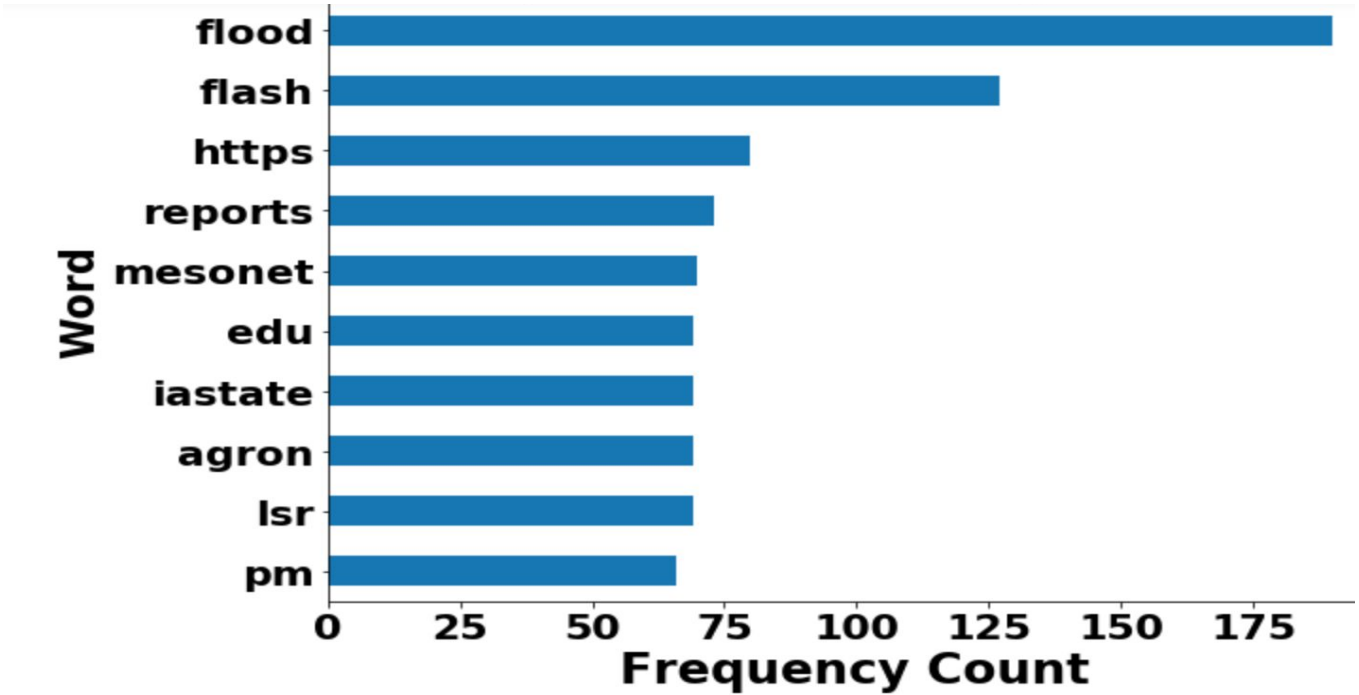
Most Frequent Words in Texas Tweets



Most Frequent Words in Iowa Tweets



Most Frequent Words in Wisconsin Tweets



03

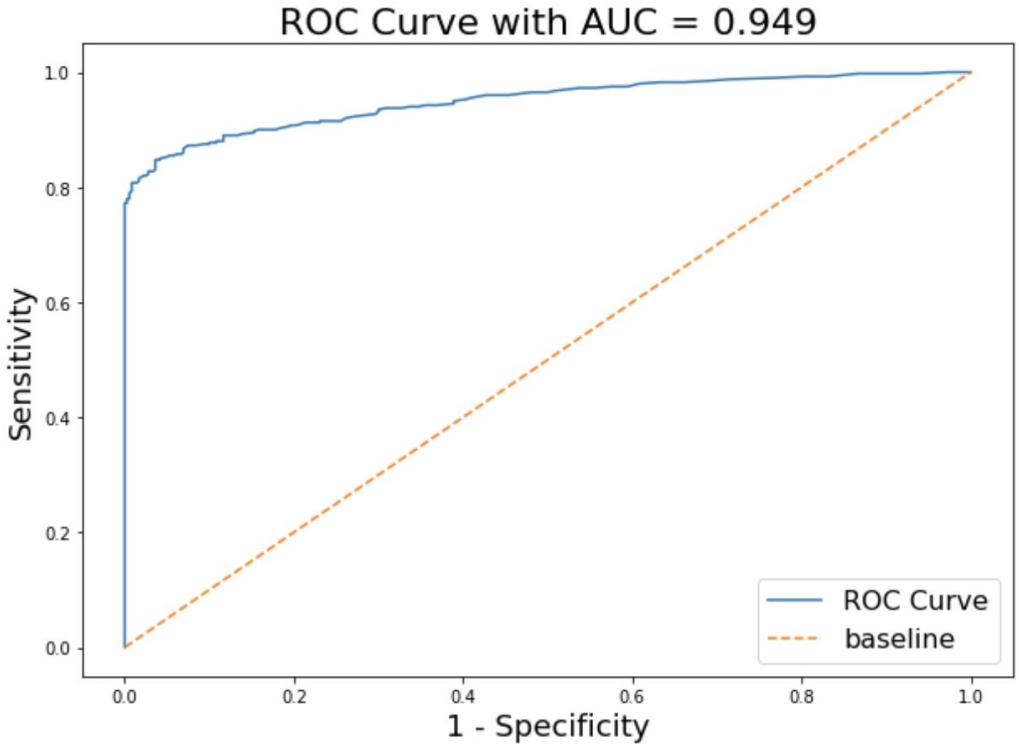
Modelling & Mapping

Model Evaluation: ROC / AUC Curve

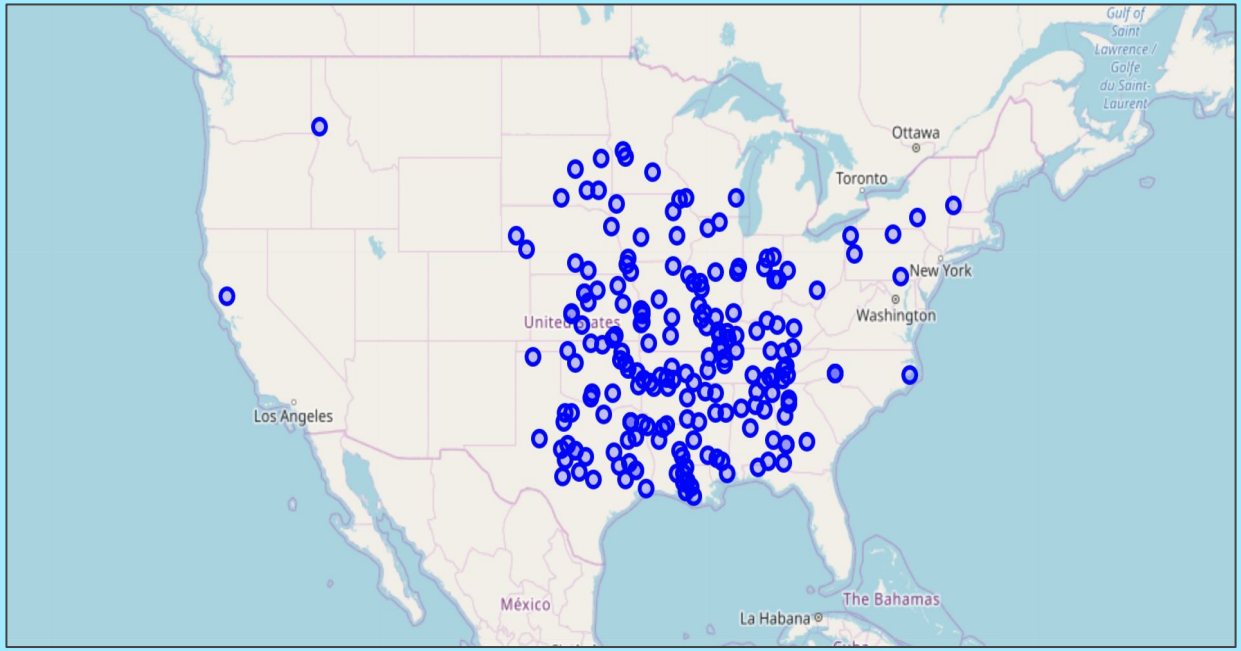
Model	Vectorizer	Training Score	Testing Score	Sensitivity	Specificity	False Positive R	Accuracy	Missclassification Rate	ROC AUC Score
Baseline Model	None	0.526		1					0.5
Logistic Regression	CountVectorizer	0.96	0.897	0.872	0.925	0.075	0.897	0.103	0.949
K Nearest Neighbors	CountVectorizer	0.907	0.896	0.84	0.958	0.042	0.896	0.104	0.942
Multinomial Naïve Bayes	CountVectorizer	0.882	0.886	0.818	0.961	0.039	0.886	0.114	0.936
DecisionTreeClassifier	CountVectorizer	0.994	0.867	0.888	0.844	0.156	0.867	0.133	0.863
Bagging Classifier	CountVectorizer	0.981	0.884	0.875	0.894	0.106	0.884	0.116	0.93
Random Forest Classifier	CountVectorizer	0.985	0.888	0.868	0.911	0.089	0.888	0.112	0.94
Logistic Regression	TfidfVectorizer	0.909	0.897	0.858	0.942	0.058	0.897	0.103	0.944
K Nearest Neighbors	TfidfVectorizer	0.905	0.87	0.84	0.903	0.097	0.87	0.13	0.917
Multinomial Naïve Bayes	TfidfVectorizer	0.887	0.891	0.828	0.961	0.039	0.891	0.109	0.941
DecisionTreeClassifier	TfidfVectorizer	0.953	0.87	0.848	0.894	0.106	0.87	0.13	0.879
Bagging Classifier	TfidfVectorizer	0.978	0.884	0.872	0.897	0.103	0.884	0.116	0.932
Random Forest Classifier	TfidfVectorizer	0.98	0.884	0.858	0.914	0.086	0.884	0.116	0.934

How well can our model validate whether or not a tweet with flood related text is actually referring to a flood?

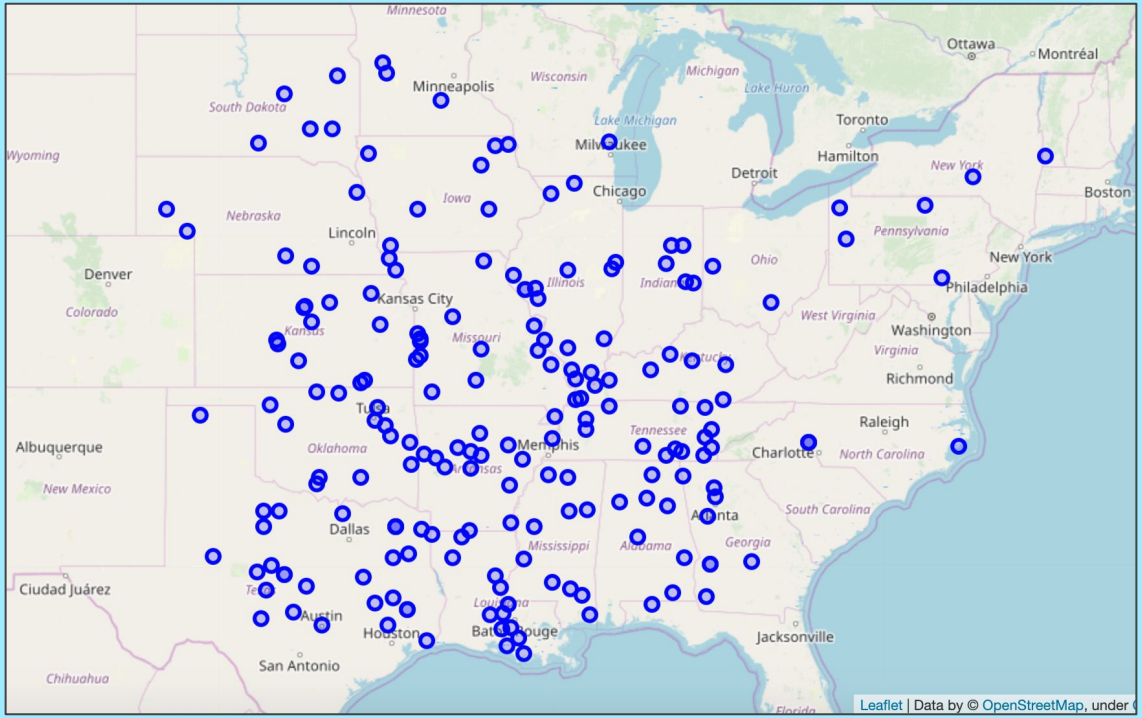
Model Evaluation: ROC / AUC Curve



Mapping Flood Tweets



Mapping Flood Tweets



4

...

Conclusions, Findings, & Limitations

Conclusions

- Models perform better on a state by state basis. NLP differs state to state.
- Must consider tradeoffs between too many false positives (saying there is a flood when there isn't) and too many false negatives (saying there is not a flood when there is).

Limitations & Next Steps

- Floods are isolated to specific regions therefore models in some areas perform better than others because of volume of data available.
- Twitter scraper and API have limitations on specificity of location data.
- Limited number of users who tweet about natural disasters.
- Limited dataset of users who offer location data.
- Potential fixes for lack of user location data:
  - Using available data from associates or ‘friend’ of a user.
  - Performing NLP on old tweets or liked tweets of a user to pinpoint location.

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, and infographics & images by **Freepik**.

Thanks!