



WILDFIRE PREDICTION USING TWITTER FEED

Presented By:

Alex Nemoto

Jovita Pinto

PROBLEM STATEMENT

- Social media, when used properly, can solve many problems during natural disasters. First responders can use social media to reach the most vulnerable in a prioritizing manner.
- Facebook, Instagram posts and Twitter tweets can be used in an efficient natural language processing model to guide relevant resources to tend to the severely affected.
- The goal of the project is to predict wildfires in a given area and map the locations using Twitter posts.

PROCESS



DATA COLLECTION

- Tweet data for modeling was gathered from two sources: CrisisLex.org (<https://crisislex.org/>), a website housing natural disaster datasets and Twitter scraping. In order for the model to be trained properly to detect whether a tweet is about a wildfire or not, along with tweets related to wildfire disasters, tweets unrelated to wildfire disasters are needed as well.
- Since CrisisLex.org did not have enough non-wildfire disaster related tweets, the keywords 'habit', 'sweet', 'snow', 'music', 'sports', and 'movies', were used to scrape Twitter to gather the required amount of the non-wildfire disaster related tweets to properly train the data.

PROCESS

DATA
COLLECTION

DATA
CLEANING &
PROCESSING

EXPLORATORY
DATA
ANALYSIS

MODELING
AND MAPPING

CONCLUSION



```
graph LR; A[DATA COLLECTION] --> B[DATA CLEANING & PROCESSING]; B --> C[EXPLORATORY DATA ANALYSIS]; C --> D[MODELING AND MAPPING]; D --> E[CONCLUSION];
```

DATA CLEANING AND PROCESSING

- Tweets were labeled 1 or 0 based on whether they belonged to either 'witness' or 'don't know' category.
- Tweet texts was then cleaned by removing urls, hyperlinks, punctuation, locations and was then converted to lowercase.
- Stopwords were removed from the cleaned data and it was then stemmed.

PROCESS

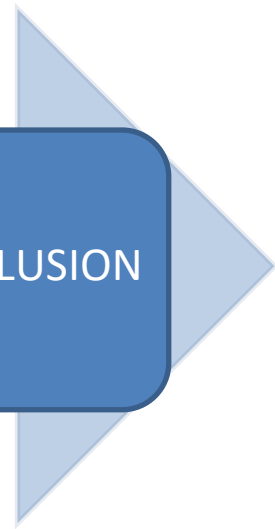
DATA
COLLECTION

DATA
CLEANING &
PROCESSING

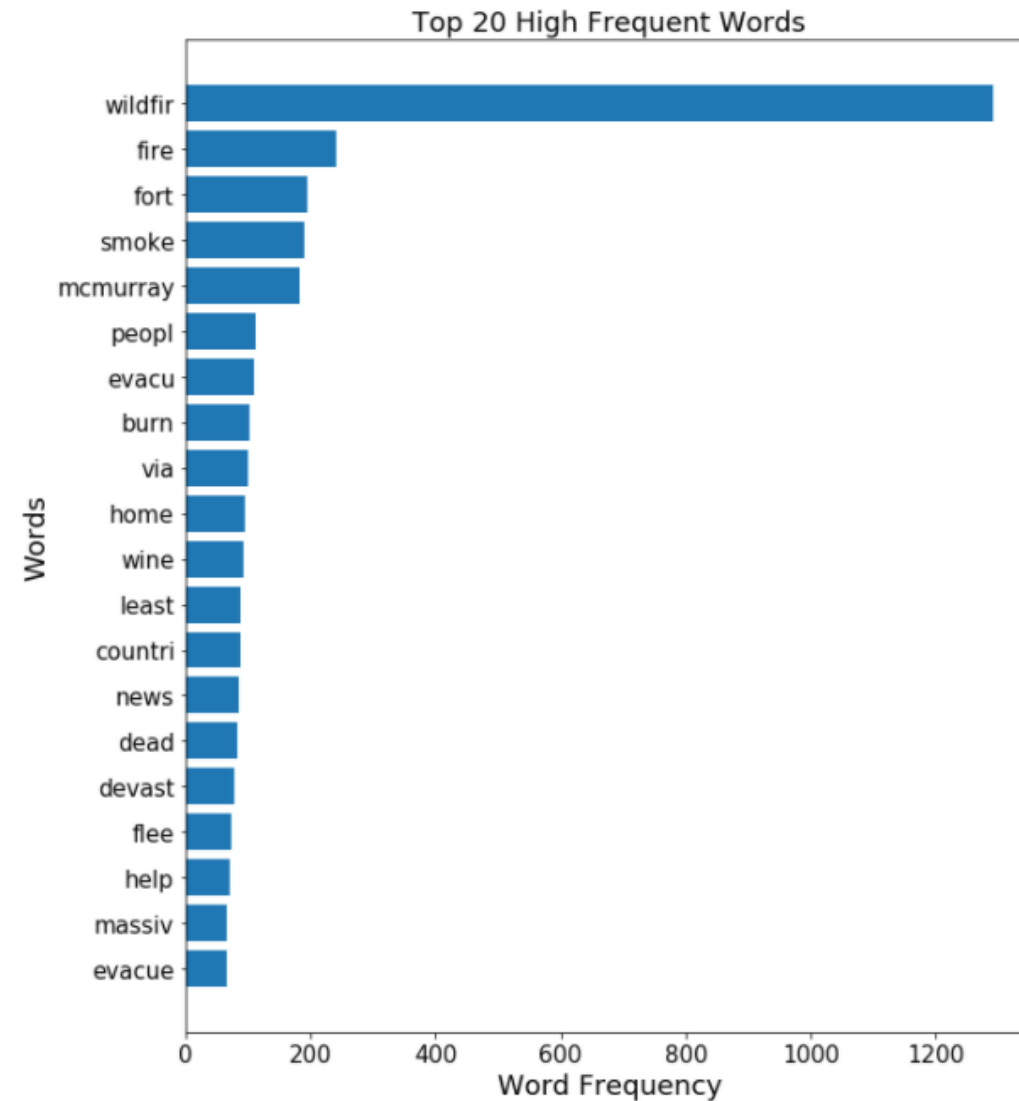
EXPLORATORY
DATA
ANALYSIS

MODELING
AND MAPPING

CONCLUSION



HIGH FREQUENCY WORDS



PROCESS

DATA
COLLECTION

DATA
CLEANING &
PROCESSING

EXPLORATORY
DATA
ANALYSIS

MODELLING
& MAPPING

CONCLUSION



```
graph LR; A[DATA COLLECTION] --> B[DATA CLEANING & PROCESSING]; B --> C[EXPLORATORY DATA ANALYSIS]; C --> D[MODELLING & MAPPING]; D --> E[CONCLUSION];
```

MODELLING

- Multiple models were developed as binary classifiers which labels tweets are 'witness' (1) or 'don't know' (0).
- Common pipeline parameters were kept constant through all models.

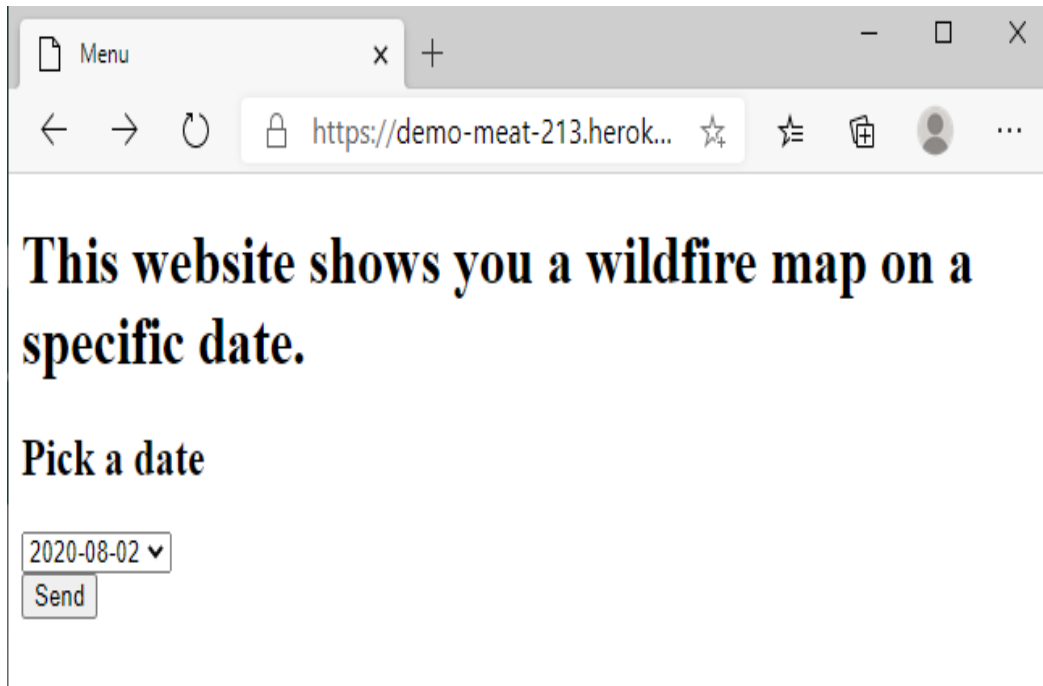
MODELLING

- A summary of accuracy scores from different models is shown in the image below:

Model Scores					
CountVectorizer	Training	Testing	TFIDFVectorizer	Training	Testing
KNN	0.999531	0.906011	KNN	0.984053	0.868852
SVC	0.998593	0.770492	SVC	0.999531	0.921311
Adaboost	0.982176	0.915847	Adaboost	0.95075	0.927869
Bagged Tree	0.934334	0.943169	Bagged Tree	0.937617	0.939891
Extra Tree	0.999531	0.948634	Extra Tree	0.998593	0.937705
Random Forest	0.999531	0.945355	Random Forest	0.998593	0.943169
Bernoulli NB	0.944653	0.856831	Bernoulli NB	0.958255	0.873224
Multinomial NB	0.977017	0.911475	Multinomial NB	0.982176	0.893989
Gaussian NB	0.94606	0.92459	Gaussian NB	0.94606	0.92459
LR	0.999531	0.931148	LR	0.999531	0.912568

- Bagged Tree model was selected to be deployed through the map web app.

WEB APPLICATION



URL: <https://demo-meat-213.herokuapp.com/>



PROCESS

DATA
COLLECTION

DATA
CLEANING &
PROCESSING

EXPLORATORY
DATA
ANALYSIS

MODELLING &
MAPPING

CONCLUSION



```
graph LR; A[DATA COLLECTION] --> B[DATA CLEANING & PROCESSING]; B --> C[EXPLORATORY DATA ANALYSIS]; C --> D[MODELLING & MAPPING]; D --> E[CONCLUSION];
```

CONCLUSION

Limitations

- Manually gathered geo-locational information
- Non-realtime data
- Arbitrarily set threshold at 50%
- Tweet context understanding model accuracy undetermined

Recommendation

- Automation of geo-locational information
- Real-time update
- Fine-tune wildfire threshold to increase the recall rate
- More testing to determine correct tweet understanding model accuracy

COMPARING OUR MODEL WITH BERT


```
[ ] tweet = [  
    'That wildfire feeling might have got a hold a month back, but it would not be allowed to do so again',  
    'In a press release Monday, Wildfire said the investments enable it to improve the accessibility and affordability of Wildfire',  
    'The slander spread like wildfire and was only checked when the drunk who invented it confessed in a magistrates court',  
    'The giggling spread like wildfire, and eventually forced the closing of some schools',  
    'The news had spread like wildfire'  
]
```


```
[112] predictions, _ = model_BERT.predict(tweet)  
      print(predictions)
```

100%  5/5 [00:00<00:00, 40.41it/s]

100%  1/1 [00:00<00:00, 12.59it/s]

[0 0 0 0 0]

 bag_treemodel.predict(vectorizer.transform(clean_tweet))

 array([1, 1, 0, 0, 0])

SOURCES

- CrisisLex.org (<https://crisislex.org/>)
- twitterscraper(<https://github.com/taspinar/twitterscraper>)
- Google Maps API(<https://developers.google.com/>)
- Heroku(<https://www.heroku.com/>)

ANY QUESTIONS?