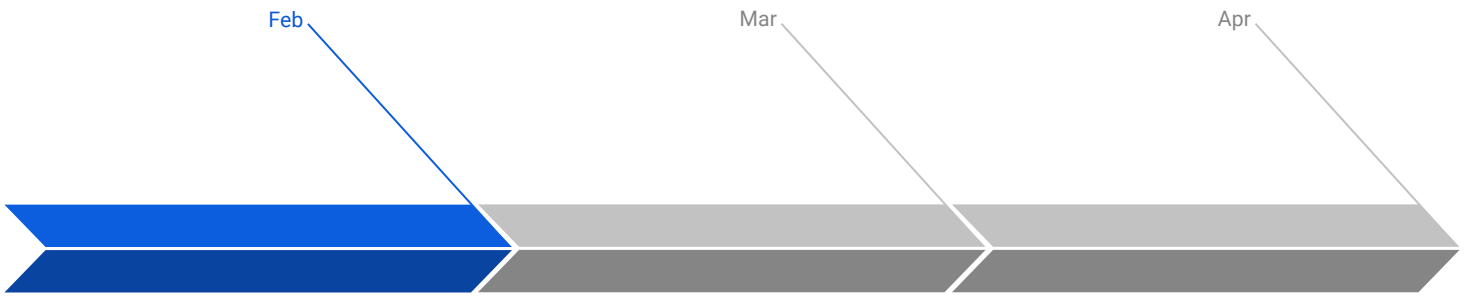
Disaster Tweets

Week 5 Check-in

Timeline



Baseline	Improved	Final
Working “simple” model	NLP / Attention	Interactive dashboard
POC Twitter scraping setup	BERT	Clustering mechanism

Business Lifecycle Progress

01	Business Understanding	• Prior Twitter maps (Live sentiment maps) • Researched real-world disasters/events
02	Data Acquisition	• Kaggle Competition • Twitter API - Student Access • Twitter selenium scraping
03	Modeling	• Data Cleaning • Baseline Models
04	Deployment	• Research into BERT Model
05	Performance Evaluation	• Several Classifiers Tested • Several metrics used to test effectiveness

Twitter Scraping

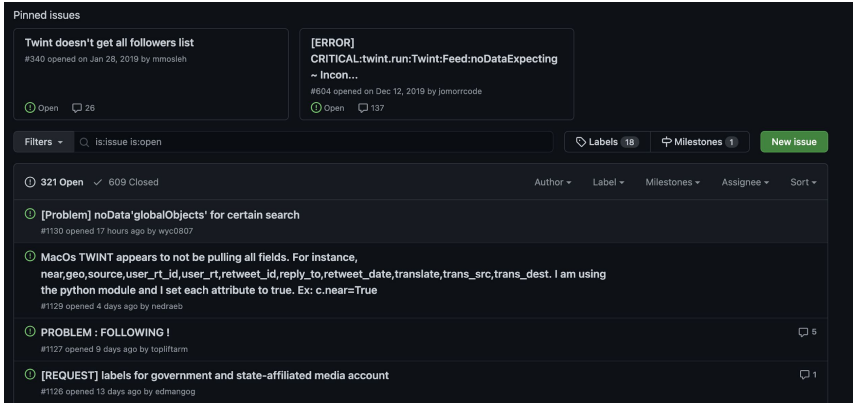
Scrapers broken

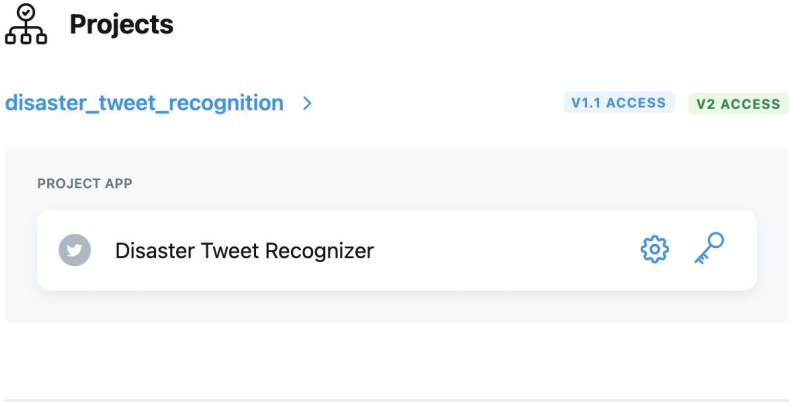
Legacy Web UI removed

Twitter API

Historical tweet access

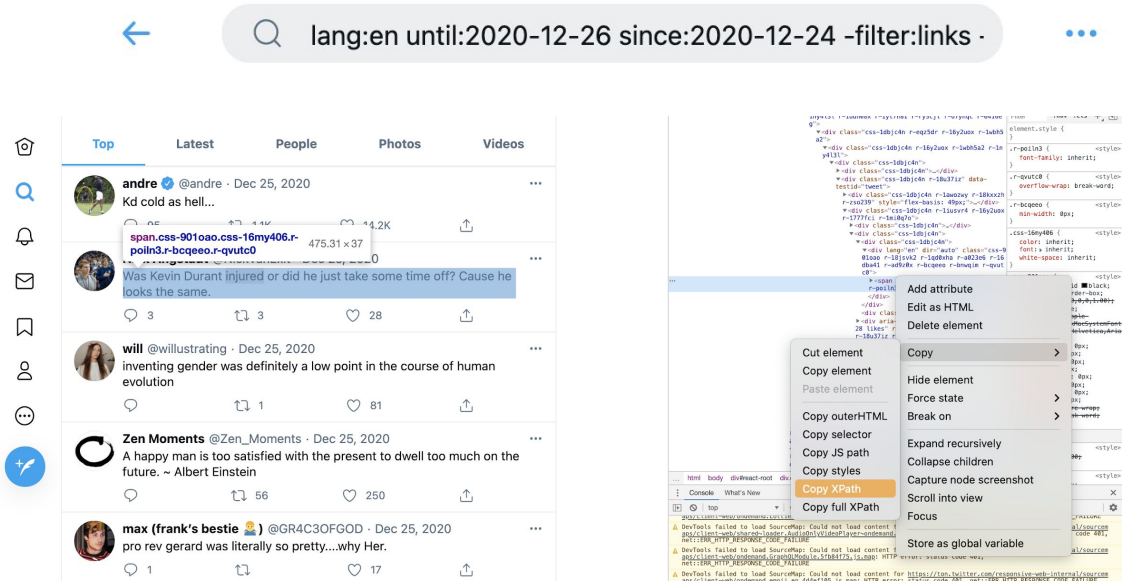
PhD student, research team, etc.





Twitter Scraping

- Selenium-based
  - Browser window
  - XPaths
- Limitations
  - Maximum number of tweets/search
  - Same tweets every time



Twitter Scraping

- Beirut: 100 tweets
- Nashville: 83 tweets
- Brunswick Co: 286 tweets

Scraping API

We scrape tweets on an "event" basis. This means that, given a real-world disaster, we can scrape tweets from around that time in the city itself and several other "control" cities.

```
nashville_bombing = ScrapeEvent(
    ['2020-12-24', '2020-12-26'],
    ['nashville', 'los angeles', 'miami', 'chicago', 'philadelphia'],
    [1,            0,            0,       0,         0            ],
    scrape_job_iterations=20
)
```

Output Format

To get the output as a pandas dataframe, call `nashville_bombing.toPandas()`

The table will look like

key	date	contents	city	city_in_disaster
0	2020-12-25T15:41:14.000Z	Tweet (presumably) about the event	nashville	1
1	2020-12-25T21:11:54.000Z	Probably not a disaster tweet	los angeles	0

Performance Evaluation

	Classifier	AUC	Accuracy	F1 Score
0	Logistic Regression	0.777215	0.782563	0.746634
3	SVM	0.779095	0.796218	0.734247
5	Perceptron	0.754251	0.757878	0.723123
1	Random Forest	0.756242	0.778361	0.696839
2	AdaBoost	0.733616	0.745798	0.686528
4	Gradient Boosting Classifier	0.720549	0.750525	0.630350

Next Steps: Looking toward Phase 2...

- Have researched NLP models and specifically BERT in Phase 1
  - Continue to build knowledge base as project moves forward
- In process of configuring environment
  - Getting appropriate tools/libraries set-up
  - Start experimenting with different models