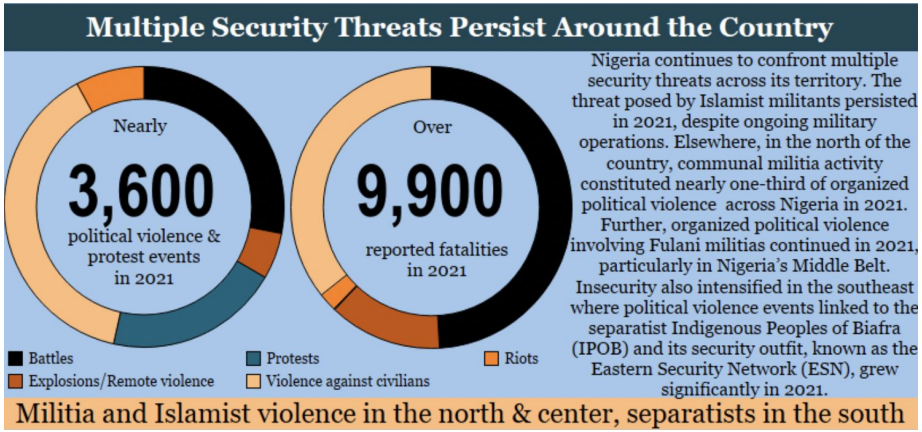
Nigeria News Detective

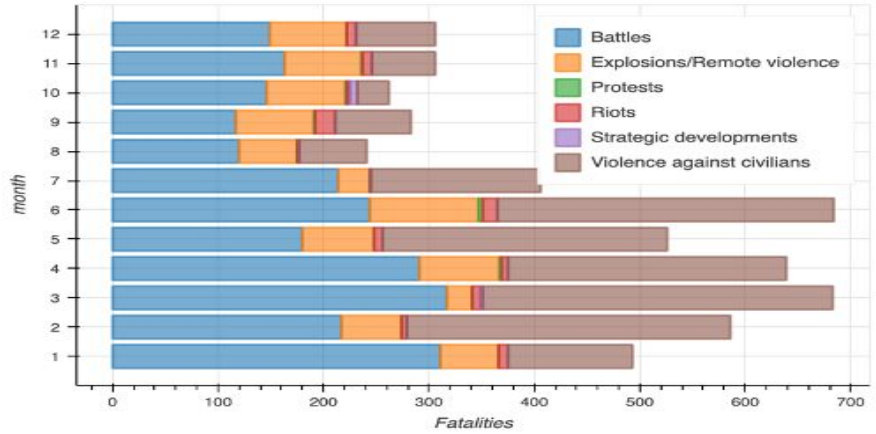
By Bhargav J Shetgaonkar

bs337@duke.edu

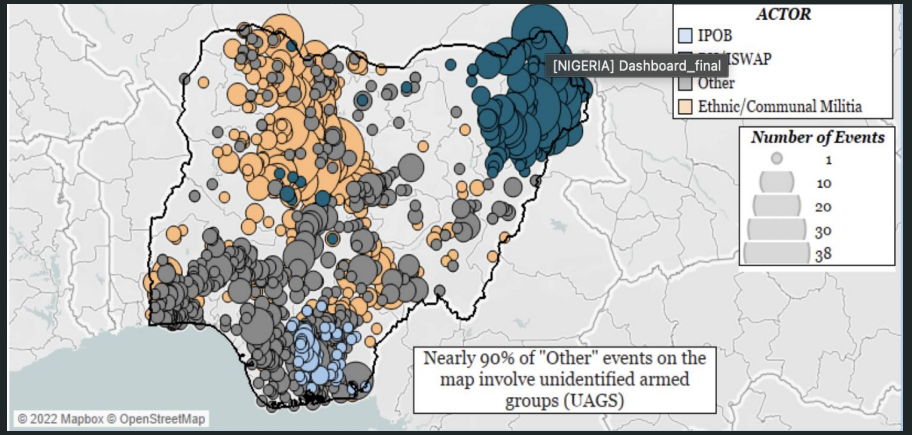
https://github.com/BombayBrownBoy09/Nigeria-News-Detective

Violence in Nigeria 2019 - present





ACLED and the government



These numbers are **alarming!** In order to protect citizens we must monitor & analyze ACLED data

1. Prominent actors include **Boko haram**, **ISWA,** and similar militia groups
2. Civilian safety is compromised due to **lack of government action**

How can we help?

Identify entities and Inform Citizens accurately

What purpose does ACLED serve?

1. Research tool for media, policy analysts, etc.
2. Information to the public
3. Dashboards for investigation on forecasting events, fatalities, assess risk and plan resources

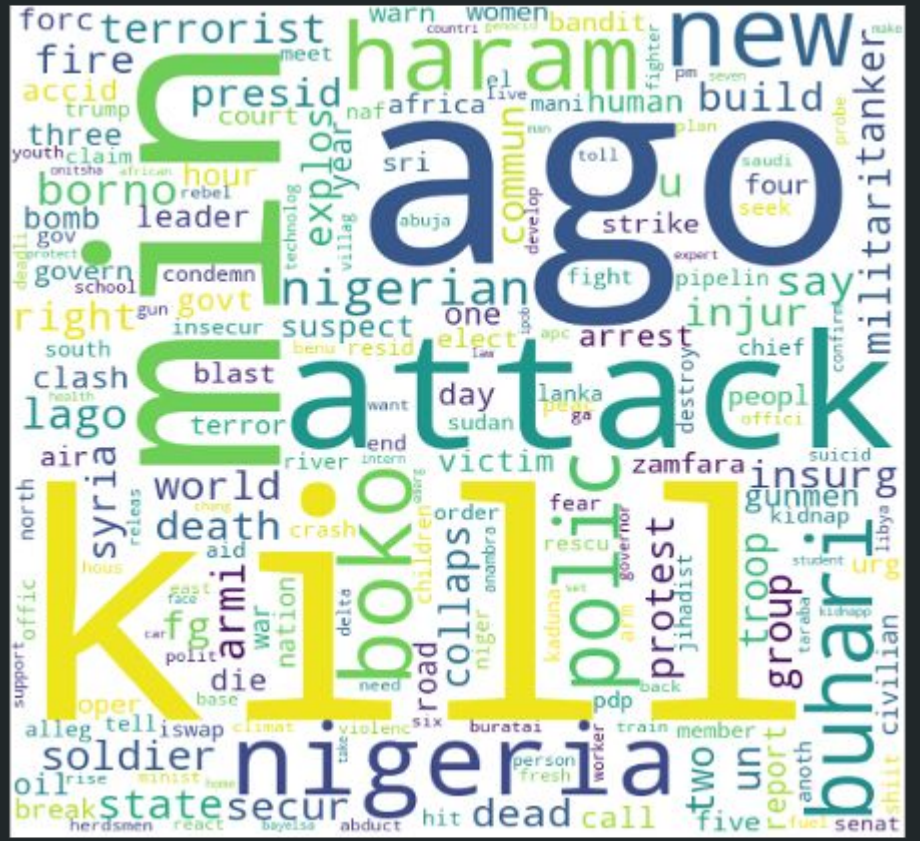
Data Sourcing

Dates used

January 2019 to April 2022

```
Index(['data_id', 'iso', 'event_id_cnty', 'event_id_no_cnty', 'event_date',
       'year', 'time_precision', 'event_type', 'sub_event_type', 'actor1',
       'assoc_actor_1', 'inter1', 'actor2', 'assoc_actor_2', 'inter2',
       'interaction', 'region', 'country', 'admin1', 'admin2', 'admin3',
       'location', 'latitude', 'longitude', 'geo_precision', 'source',
       'source_scale', 'notes', 'fatalities', 'timestamp', 'iso3'],
      dtype='object')
```

This may be downloaded via the official ACLED site using your unique access key



Word cloud representing the notes column

Problem Statement

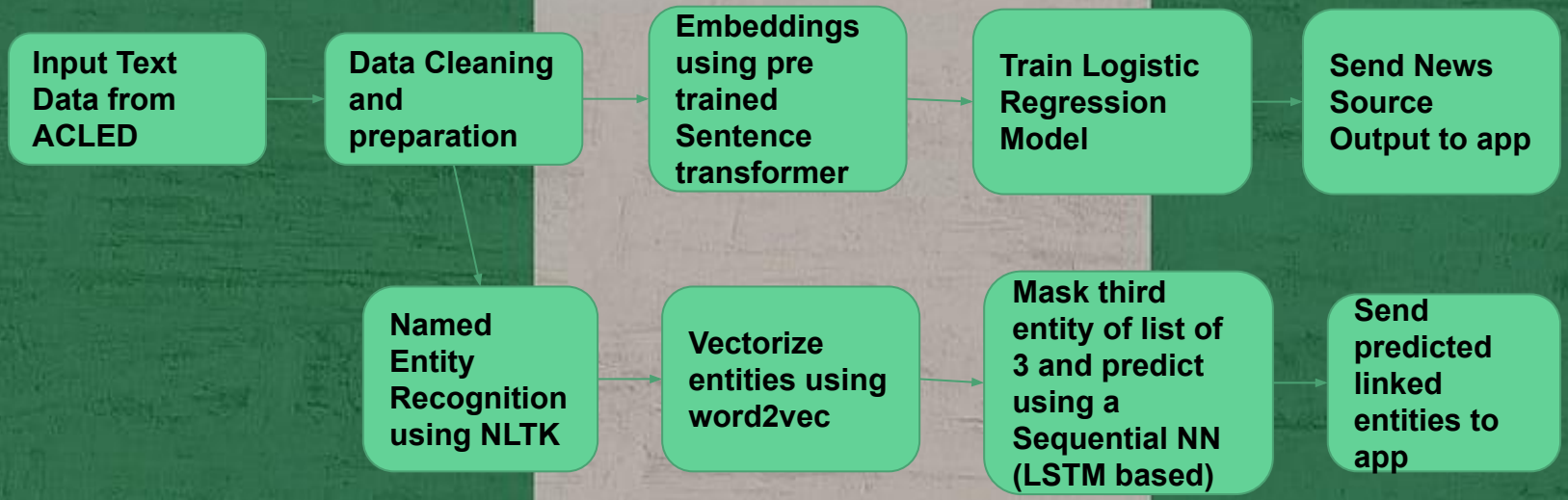
Using publicly available **ACLED** - The Armed Conflict Location & Event Data to get **news source** and **linked entities** of news articles for government research

# of events - approx. **10,000**

Average entities per event **3**

unique news sources **1915**

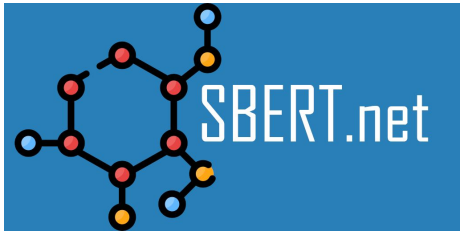
Modeling Pipeline



Performance of Models

- Sentence Transformer based news source classification - 80/20 training validation split: Train/Val accuracy - **21% / 16% to classify 1950 types of sources**. Other approaches included **pytorch based** (12% train acc) and **bag of words model** (11.2 % train acc) that performed significantly lower due to the large vocabulary size that results in sparse matrices
- Entity prediction model (NN with 2 dense and 2 LSTM layers) had a training / validation split of 75% / 15% and train/validation accuracy of **8%**

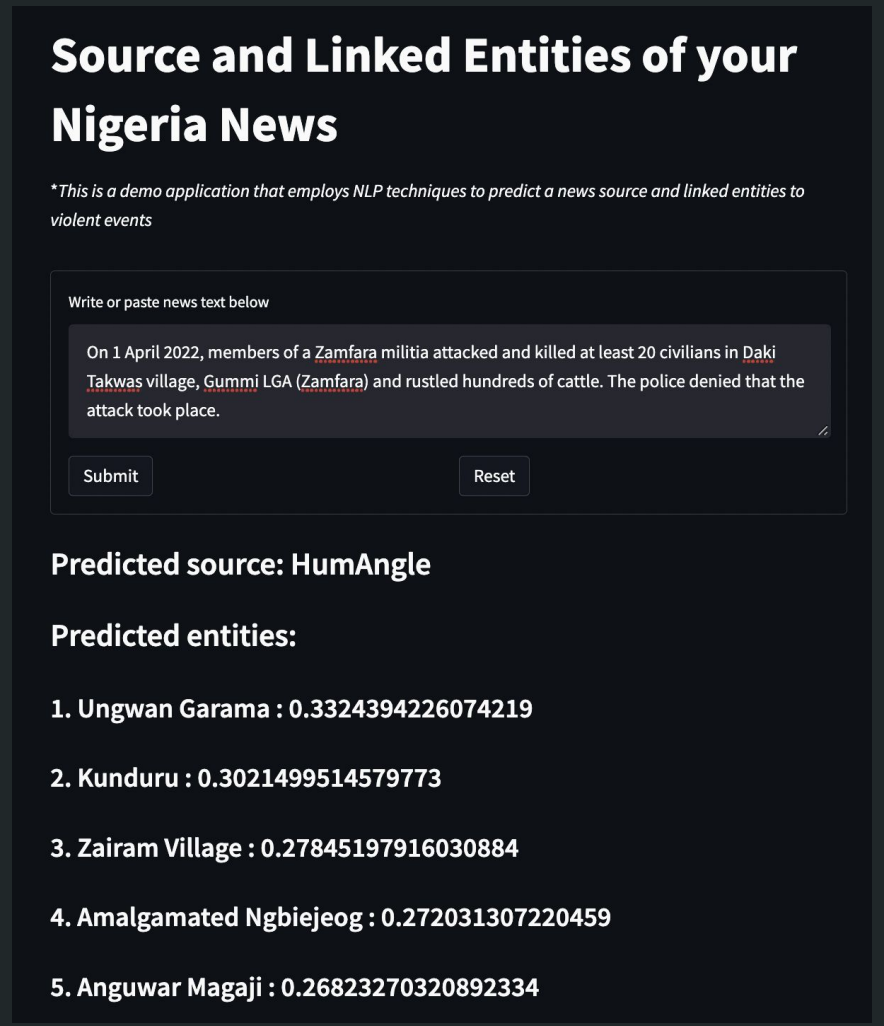
```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 lstm (LSTM)                 (None, 2, 100)            80400

 lstm_1 (LSTM)               (None, 50)                30200

 dense (Dense)               (None, 50)                2550

 dense_1 (Dense)             (None, 100)               5100

=================================================================
Total params: 118,250
Trainable params: 118,250
Non-trainable params: 0
_________________________________________________________________
```



Application Demo

Note that these entities are specific to Nigeria and can be verified via Google Search

https://share.streamlit.io/bombaybrownboy09/nigeria-news-detective/main/app.py



Conclusion

We have used NLP techniques such as text pre processing, named entity recognition, text classification and LSTM based models to build a research tool for media personnel and investigators in Nigeria

What next?

- *Forecasting feature* to predict events
- Improve model performance by *deeper models* and automated updation of data
- *Build UX* for Government Investigators as main admin

Questions?

“Nigeria [in 1990] was all rumour, an unbelievable amount of rumour – largely about crime and almost mythical manifestations of evil.”

- REM KOOLHAAS