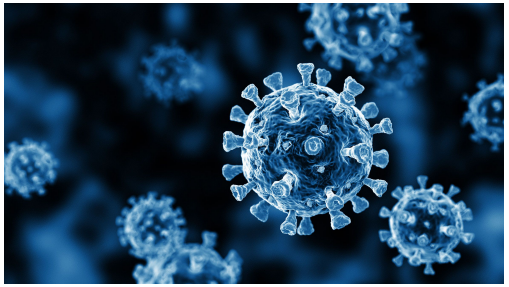
Geological Extraction for Twitter

- Varad Pimpalkhute, affiliated with Indian Institute of Information Technology, Nagpur.
- Project Advisor: Dr. Arjun Magge.
- Supervisor: Dr. Graciela Gonzalez-Hernandez.

Why do we need geolocation of a social media user?

- Tracking Infectious Diseases.
- Gathering analysis on how well an advertisement is received in a particular region.
- Keeping tabs on spread of diseases such as COVID-19.



Task : Given an input tweet and user screen name, identify the user's current location of residence.

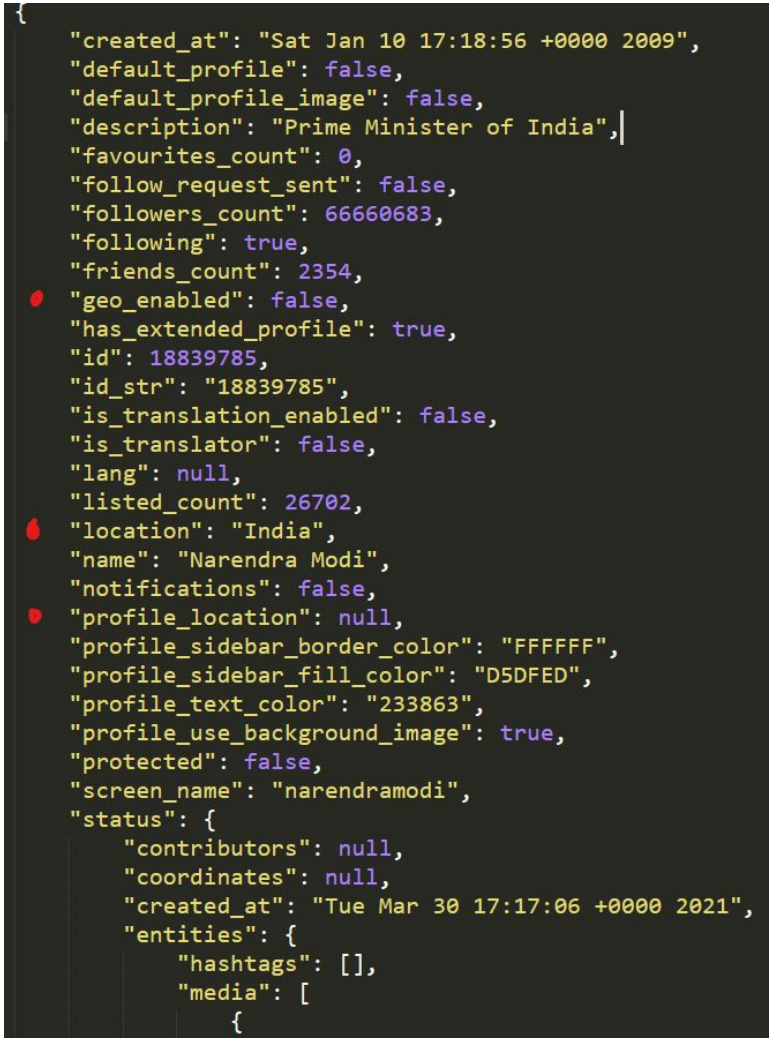
Let’s Take an Example

User Screen Name :

Input User Tweet :

A few assumptions and challenges

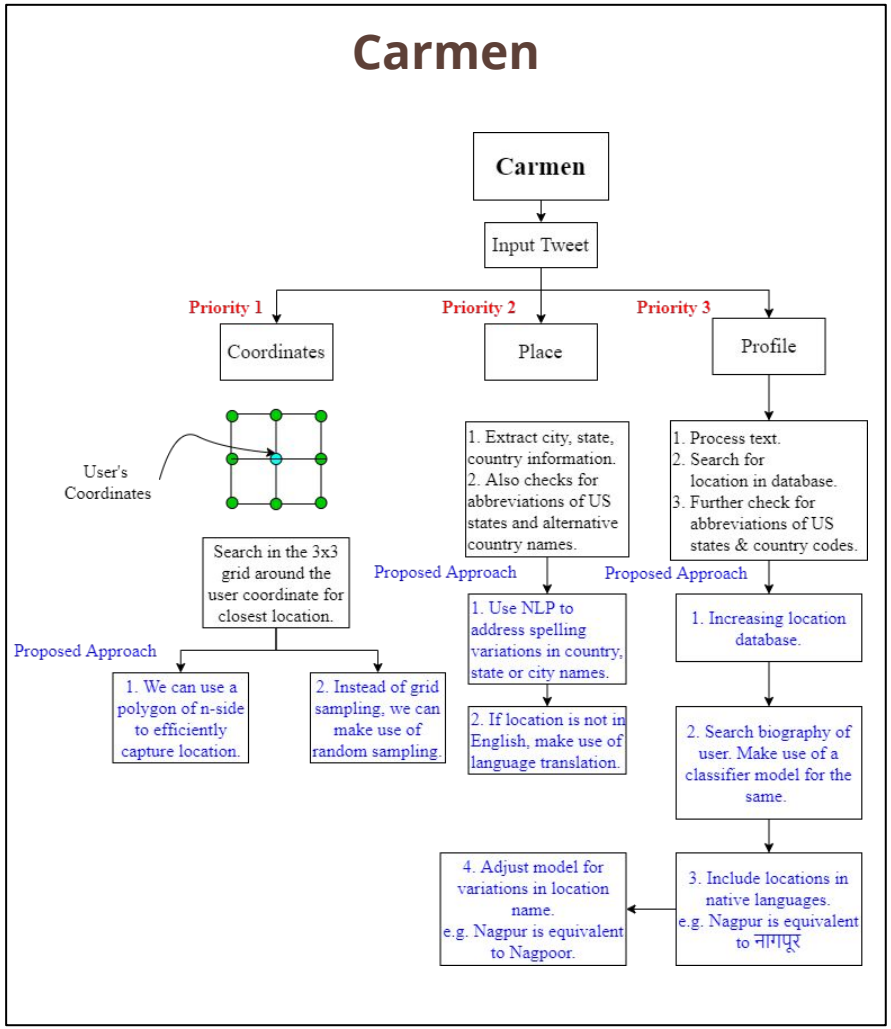
- Old datasets are publicly available for use.
- We assume that a given user has not changed his place of residence.
- If an account is deleted, extracting metadata becomes arduous.
- We assume that user feeded information is true.
- Tweet with metadata amounts to only 1% of all the tweets available.



USER METADATA

Some of the important location related tags are:

- ★ Geo tag : (latitude, longitude)
- ★ Place tag : Location based on coordinates
- ★ Profile tag : Location form present in profile.

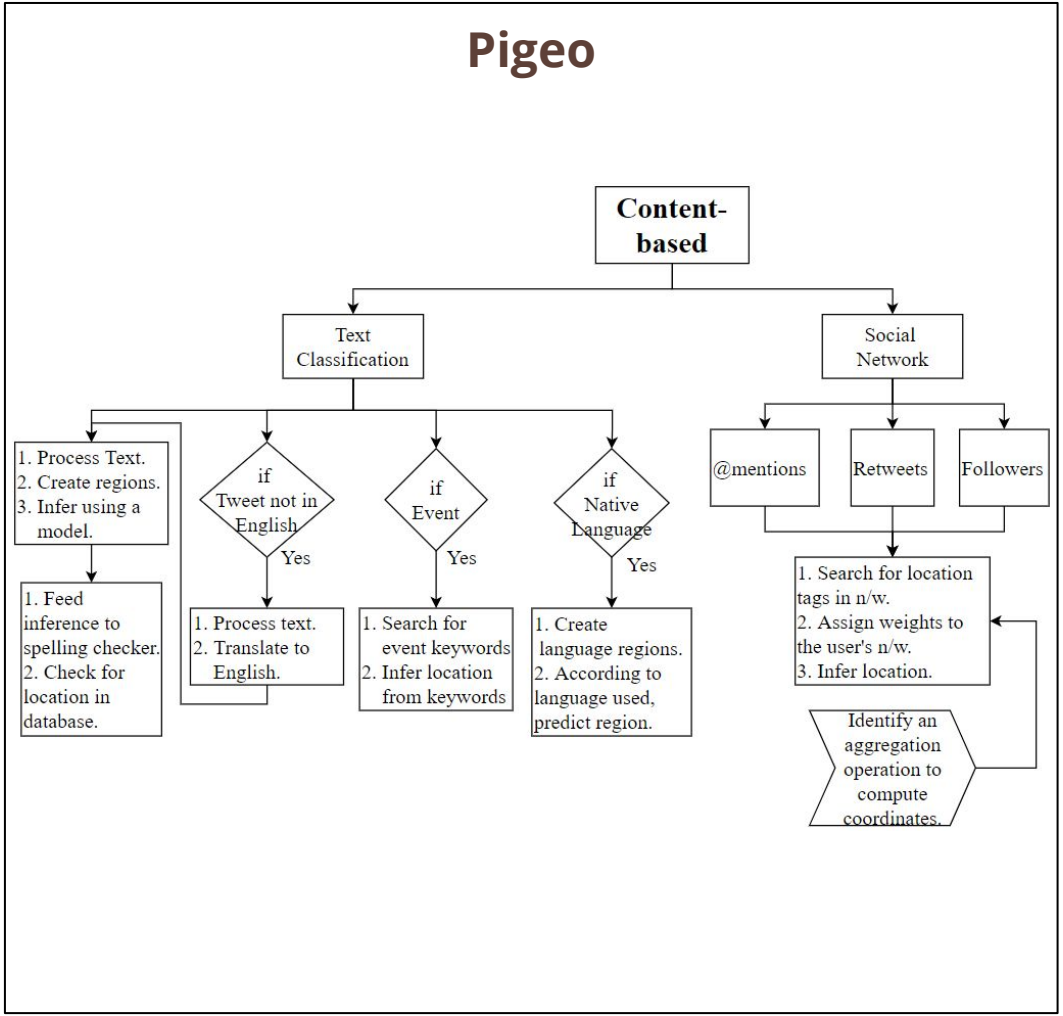


Exemplar methods in literature

+ Simple and fast
+ Tends to cover all meta tags
- Small location database
- Doesn't account for content-based information.
- Prone to errors due to spelling variations.

Exemplar methods in literature

+ Accounts for content-based information.
+ Makes use of social network of a Twitter user.
- Doesn’t account for meta information.
- Accuracy is low.
- Doesn’t structure the location hierarchically.



Goal

1. Improve on Carmen and Pigeo. Finally, integrate both the methodologies in one model.
2. Generating a dataset for country-wise classification by scrapping user data from Twitter using tools like Tweepy.
3. Addressing the issue of classification on a hierarchical level.

Methodology

Content-based Approach

- Generating country-wise dataset for classification.
- Building different models for text classification using FastText, Transformers, CNNs, etc.
- Fine tuning models.
- Addressing the issue of hierarchical classification.

Metadata-based Approach

- Generating a location database.
- Introducing population heuristics for better prediction of location.
- Computing the radius of each location using GADM and geopandas.

➢ Generating country-wise dataset for classification.

```
# UK: 12 regions
uk_states = ['Scotland', 'Wales', 'North East', ..., 'West Midlands', 'East Midlands', 'Greater London']

# main script
user_list, tweets = [], []
for state in states:
    user_state_list = []
    # Searching for users living in the state by making use of 4 different keywords and updating the
    # state dictionary accordingly
    user_state_list = keyword_search(state, 1, user_state_list)
    user_state_list = keyword_search(state, 2, user_state_list)
    user_state_list = keyword_search(state, 3, user_state_list)
    user_state_list = keyword_search(state, 4, user_state_list)
    # Update the main dictionary
    user_list.extend(user_state_list) # or we can do something like this:
                                      # user_dict[state] = user_state_dict
    # Extracting Tweets from the user screen name.
    for user in user_state_list:
        tweet = user_tweets(state, user)
        if tweet is not None:
            tweets.extend(tweet)
    print('Completed Scrapping', state)
```

➢ Generating country-wise dataset for classification.

```
Completed Scrapping Scotland
Completed Scrapping Northern Ireland
Completed Scrapping Wales
Completed Scrapping North East
Completed Scrapping North West
Completed Scrapping Yorkshire and the Humber
Completed Scrapping West Midlands
Completed Scrapping East Midlands
Completed Scrapping South West
User has protected tweets, skipping ...
Completed Scrapping South East
Completed Scrapping East of England
Completed Scrapping Greater London
Total Tweets Scrapped: 122152
Total user IDs: 1272

(geoloc) C:\Users\Varad\OneDrive\Desktop>python create_dataset.py
```

- Building different models for text classification using
  - FastText,
  - Transformers,
  - CNNs, etc.
- Generating a location database.
- Computing the radius of each location using GADM and geopandas.

Future work

- ★ Improve on the classification.
- ★ Develop a hierarchical classifier.
- ★ Make use of social network of a Twitter user.
- ★ Account for information in native languages.

References

[1] X. Zheng, J. Han and A. Sun, "A Survey of Location Prediction on Twitter," in IEEE Transactions on Knowledge and Data Engineering, vol. 30, no. 9, pp. 1652-1671, 1 Sept. 2018, doi: 10.1109/TKDE.2018.2807840.

[2] Han, T. (2012). Geolocation Prediction in Social Media Data by Finding Location Indicative Words. In Proceedings of COLING 2012 (pp. 1045–1062). The COLING 2012 Organizing Committee.

[3] Roller, J. (2012). Supervised Text-based Geolocation Using Language Models on an Adaptive Grid. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (pp. 1500–1510). Association for Computational Linguistics.

[4] Roesslein, J. (2020). Tweepy: Twitter for Python!. URL: https://github.com/tweepy/tweepy.

[5] Mark Dredze, Michael J Paul, Shane Bergsma, & Hieu Tran (2013). Carmen: A Twitter Geolocation System with Applications to Public Health. In AAAI Workshop on Expanding the Boundaries of Health Informatics Using AI (HIAI).

[6] Rahimi, A., Cohn, T., & Baldwin, T. (2016). pigeo: A Python Geotagging Tool. In Proceedings of ACL-2016 System Demonstrations (pp. 127–132). Association for Computational Linguistics.