भारतीय प्रौद्योगिकी संस्थान धारवाड
Indian Institute of Technology Dharwad

Report

Data collection and preprocessing

Ashwin Waghmare
Department of Computer Science
210010060

Supervisor:
Prof. Anand Kojengbam

2024

Acknowledgements

I would like to express my sincere gratitude to Professor Anand Konjengbam for his invaluable guidance and mentorship throughout the duration of this RnD project. They provided me with the necessary resources, academic insight, and continuous encouragement, which significantly contributed to the quality and depth of this project.

I am also thankful for the opportunity to learn from Dr Konjengbam, not only in the context of this project but also in my overall academic journey.

This project would not have been possible without the collective efforts and support of these individuals and institutions. I am truly fortunate to have had the opportunity and to be a part of this academic community.

Thank you.
Ashwin Waghmare
IIT Dharwad
2024

Contents

0.1 Abstract

This report presents a comprehensive collection of data collection tools covering major media agencies. The primary objective of these tools is to collect data specifically in light of the storminess amidst ethnic groups of Manipur. [4]On 3 May 2023, ethnic violence erupted among the Meitei people, a majority that lives in the Imphal Valley, and the Kuki-Zo tribal community from the surrounding hills. According to government figures, as of 15 September, 175 people have been killed in the violence. 1,108 others were injured while 32 are missing. 4,786 houses were burnt and 386 religious structures including temples and churches were vandalized. The violence left more than 70,000 people displaced from their homes.

The resulting dataset contains preprocessed articles and comments written in context of the aforementioned scenario. This dataset intends to undergo sentiment analysis to gain an insight about the conflicting ethnic groups of Manipur.

0.2 Introduction

Data collection is the process of gathering and measuring information on targeted variables in an established system, which then enables one to answer relevant questions and evaluate outcomes. The various tools implemented in the report are listed below:

- **Beautiful soup**
  Beautiful Soup (bs4 [1]) is a Python library for pulling data out of HTML and XML files. Websites of various media agencies based in Manipur are scraped for data.
- **Twikit**
  Twikit [2] is a simple API wrapper that interacts with Twitter's (now X) internal API. Using this API, tweets for a set of particular hashtags are extracted.
- **YouTube data API**
  Using YouTube's data API [3] comments from specific channels and videos are scrapped.

0.3 Epapers

0.3.1 Brief

Newspapers are the primary source of information that shed light on current events. The opinions of the older generation are significantly impacted by these media. Data from popular newspaper agencies including Sangai Express, Imphal Times, Hueiyenlanpao and Frontier Manipur is scraped and preprocessed into a dataset.

0.3.2 Methodology

Using automated python scripts, beautiful soup is used to parse the HTML/XML code of the websites. Each website having different coding patterns and terminologies, is carefully inspected and specific data is scraped and stored as a raw csv file. These raw csv files are intended to undergo basic preprocessing, wherin individual entries would be analysed based on their relevancy to the motive of this report, that is Manipur's violence.

- **Sangai Express, Frontier Manipur, Imphal Times**
  The website of Sangai Express ("`https://www.thesangaiexpress.com/`"), The Frontier Manipur ("`https://thefrontiermanipur.com/`"), Imphal Times ("`https://www.imphaltimes.com/`") are individaully scanned for articles using a simple webcrawler in separate python notebooks. The articles are stored in a csv file as "article name, article text". These articles need to be manually segregated for relevancy, using the title of the article.

- **Hueiyenlanpao**
  Epapers from a particular start date till a particular end date are downloaded and stored in a folder, using an automated python script. These digital versions of the daily newspaper Hueiyenlanpao are basically jpeg versions of the newspaper stored in pdf format. Google's OCR model, pytesseract is used to optically recognise the text and is stored in an individual sentence format. Newspapers from desired dates can be picked manually for further analysis.

0.3.3 Results

Extracted data from respective websites are stored in separate csv files.

Sr. No	Media name	Number of articles scraped	Total csv file size
1	Sangai Express	644	1.6Mb
2	The Frontier Manipur	222	1.1Mb
3	Imphal Times	10	68.5Kb

Table 1: Statistics of extracted data from Media

Epapers from Hueiyenlanpao are stored in pdf format in separate directory.

0.4 Twitter (X)

0.4.1 Brief

Twitter is a free social networking site where users broadcast short posts known as tweets. People use Twitter to get the latest updates and promotions from brands, communicate with friends, and follow business leaders, politicians and celebrities. They also use it to stay current on news and events. Twitter is very popular among millennials and young adults thus making a good source to analyse this age group of people.

0.4.2 Methodology

An automated python script is used to extract tweets from twitter. Using the python package twikit [2], specific tweets can be extracted for free. Latest tweets with the following hashtags are extracted:

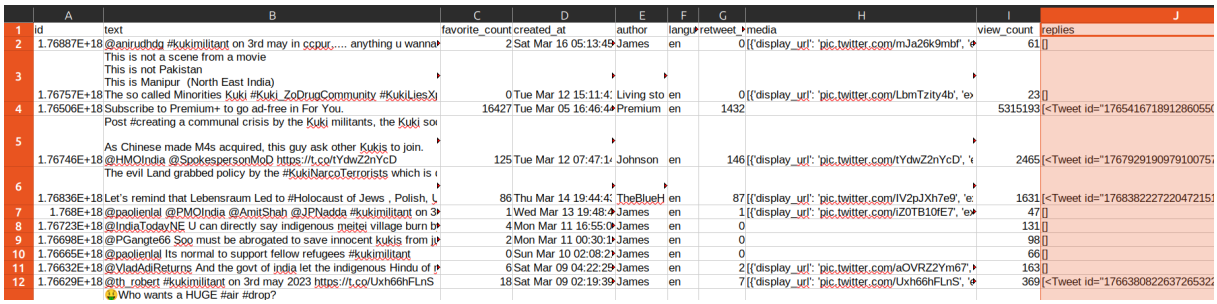
```
#KukiMilitant #SaveManipur #Savemeiteis #Ethniccleansing #
kukimilitants #kuki #meitei #visitManipur #Kuki_ZoEngineeredViolence
#SaveMoreh #SaveManipurSaveIndia #MorehBurning #ManipurFights Back
#KukiMilitantvioleteSoO #AbrogateSoO #poppy #illegalImmigration
#terror #Sanamahi #7MonthsOfNoInternet#StopTheViolence
#MyManipur #Stand4Manipur #ManipurUnderAttack #StopGenocideofMeiteis
#lies #genocide #KukiLiesXposed #KukiAtrocites #KukiZoEngineerdManipurConflict
#meiteiyouths #KukiWarCrimes #PoppyCultivators #kukinarcoterrorist
#Narcokukiterrorist #SavekukiZoTribals #ManipurCrisis #KukiRapist
#AssamRifles#lamkatalk #stoppoppycultivators #kukizo #kukizolivesmatter
#ManipurPolic #FailedGovernment #manipurisburning #MeiteiMilitant
```

For each tweet the following data has been extracted: id (unique identifier of the tweet), text (full text of the tweet), favorite_count (count of favorites or likes for the tweet), created_at (created_at converted to datetime), username (the username of the user that created this tweet), lang (language of the tweet), retweet_count (number of times this tweet was reused by different users), media (list of media entities associated with the tweet, mostly contains urls of the images attached), view_count (number of users who have viewed this tweet), replies (count of replies to this tweet).

The drawback to this method is that, only 500 tweets can be requested for every 15 minutes. The python script sleeps for 900 seconds after it has extracted 500 tweets. All tweets have been extracted from a personal account.

0.4.3 Results

The data extracted from the tweets are stored in a single csv file with separate columns for each attribute mentioned above. Roughly 50 tweets per hashtag (total 47 hashtags) are extracted.



	A	B	C	D	E	F	G	H	I	J
1	id	text	favorite_count	created_at	author	langu	retweet_	media	view_count	replies
2	1.76887E+18	@anirudhdg #kukimilitant on 3rd may in ccpur,.... anything u wanna	2	Sat Mar 16 05:13:49	James	en	0	{'display_url': 'pic.twitter.com/mJa26k9mbf', 'e	61	[]
3	1.76757E+18	This is not a scene from a movie This is not Pakistan This is Manipur (North East India) The so called Minorities Kuki #Kuki_ZoDrugCommunity #KukiLiesXj	0	Tue Mar 12 15:11:4	Living sto	en	0	{'display_url': 'pic.twitter.com/LbmTzity4b', 'e	23	[]
4	1.76506E+18	Subscribe to Premium+ to go ad-free in For You.	16427	Tue Mar 05 16:46:44	Premium	en	1432		5315193	[<Tweet id="176541671891286055
5	1.76746E+18	Post #creating a communal crisis by the Kuki militants, the Kuki so As Chinese made M4s acquired, this guy ask other Kukis to join. @HMOIndia @SpokespersonMoD https://t.co/tYdwZ2nYcD	125	Tue Mar 12 07:47:1	Johnson	en	146	{'display_url': 'pic.twitter.com/tYdwZ2nYcD', '	2465	[<Tweet id="176792919097910075
6	1.76836E+18	The evil Land grabbed policy by the #KukiNarcoTerrorists which is Let's remind that Lebensraum Led to #Holocaust of Jews , Polish, L	86	Thu Mar 14 19:44:4	TheBlueH	en	87	[{'display_url': 'pic.twitter.com/IV2pJXh7e9', 'e	1631	[<Tweet id="176838222720472151
7	1.768E+18	@paolienlal @PMOIndia @AmitShah @JPNadda #kukimilitant on 3	1	Wed Mar 13 19:48:4	James	en	1	{'display_url': 'pic.twitter.com/iZ0TB10fE7', 'ex	47	[]
8	1.76723E+18	@IndiaTodayNE U can directly say indigenous meitei village burn b	4	Mon Mar 11 16:55:0	James	en	0		131	[]
9	1.76698E+18	@PGangte66 Soo must be abrogated to save innocent kukis from ji	2	Mon Mar 11 00:30:1	James	en	0		98	[]
10	1.76665E+18	@paolienlal Its normal to support fellow refugees #kukimilitant	0	Sun Mar 10 02:08:2	James	en	0		66	[]
11	1.76632E+18	@VladAdiReturns And the govt of India let the indigenous Hindu of	6	Sat Mar 09 04:22:29	James	en	2	{'display_url': 'pic.twitter.com/aOVRZ2Ym67', '	163	[]
12	1.76629E+18	@th_robert #kukimilitant on 3rd may 2023 https://t.co/Uxh66hFLnS Who wants a HUGE #air #drop?	18	Sat Mar 09 02:19:39	James	en	7	{'display_url': 'pic.twitter.com/Uxh66hFLnS', 'e	369	[<Tweet id="176638082263726532

Figure 1: Screenshot of manipur_violence_tweets.csv

0.5 YouTube

0.5.1 Brief

YouTube is a video sharing service where users can watch, like, share, comment and upload their own videos. The video service can be accessed on PCs, laptops, tablets and via mobile phones. YouTube is a free to use service and a can be a great space for teens to discover things they like. For most young people YouTube is the only social media platform that they are exposed to. Comments made by these viewers on specific videos related to the motive of this report can be scraped to analyse the sentiments young generation.

0.5.2 Methodology

YouTube does not allow third party access to their platform. However, they do have an API that allows users to extract a set amount of data, that is 10000 requests per day. Below is a stepwise implementation of YouTube Data API.

1. Set Up a Project and Enable the YouTube Data API
   - Go to the Google Cloud Console (`https://console.cloud.google.com/`).
   - Create a new project or select an existing one.
   - In the navigation pane, click on "APIs & Services" then "Library".
   - Search for "YouTube Data API" and enable it for your project.
2. Create API Credentials
   - In the Google Cloud Console, navigate to "APIs & Services" then "Credentials".
   - Click on "Create credentials" and select "API key".
3. Install Necessary Libraries
   - A library `google-api-python-client` for Python should be installed using pip
   - `pip install --upgrade google-api-python-client`
4. Write Code to Access the API
   - Use the API key you generated to authenticate your requests to the YouTube Data API.
   - If the "video ID" of the YouTube video from which you want to retrieve comments is known then below is the pseudocode to get comments from that video.

0.5.3 Results

A collection of videos specifically highlighting the Manipur's conflict are found manually. Roughly 20 comments from each video is scraped.

0.6 Futureworks

- **Nltk Library**
  The Nltk library will be employed for further preprocessing of the raw dataset, including tasks such as tokenization, stemming, and lemmatization to prepare the data for advanced analysis.

- **Sentiment Analysis**
  The generated dataset will be utilized for in-depth sentiment analysis, aiming to uncover nuanced insights into the sentiments related to the ethnic conflict in Manipur. This analysis will provide a deeper understanding of the emotional and psychological dimensions of the conflict, paving the way for informed interventions and policy decisions.

Bibliography

[1] https://beautiful-soup-4.readthedocs.io/en/latest/

[2] https://pypi.org/project/twikit/1.0.3/

[3] https://developers.google.com/youtube/v3

[4] https://en.wikipedia.org/wiki/2023%E2%80%932024_Manipur_violence

[5] https://pypi.org/project/pytesseract/

[6] https://github.com/tesseract-ocr/tesseract