Geo-semantic surveillance and clustering of crime locations and social risk hotspots using print media reports.

Toyib Ogunremi, Olubayo Adekanmbi, Anthony Soronnadi, David Akanji
Data Scientists Network Foundation
{Toyib, olubayo, Anthony, David}@datasciencenigeria.ai

Abstract

Kidnapping is a significant social risk in Nigeria which often lack adequate intervention due to the unavailability of local crime data, underreporting of cases due to fear of retaliation from suspected perpetrators or involvement of security operatives. In response, we have developed a data-driven solution by generating a reliable dataset of crime locations and entities in Nigeria. Our approach involves geoparsing newspaper reported crime locations and entities using NLP techniques and Google geocoder, as well as clustering and geospatial analysis of identified social risk hotspots. We have designed an algorithm that geoparse locations in unstructured raw text. Our research aims to provide insights and solutions for combating the menace posed by kidnappers to Nigeria.

1. Introduction

Geo-semantics is an interdisciplinary field that combines Natural Language Processing (NLP) which enables computers to understand human language and Geospatial techniques which analyze data with respect to space. The process of converting unstructured text into structured geospatial information is Geoparsing which involves Toponym recognition using NER techniques from NLP and Toponym resolution using Geocoding tools which resolves the extracted locations to the global space. This study aims to apply Named Entity Recognition (NER) and geospatial analysis to generate a detailed dataset of crime incidents in Nigeria, contributing to crime analysis and prevention initiatives in developing countries.

2. Related works

In 2010, MorphoSyntactic Parser, was proposed to provide input to WikiCrimes, a web-based platform for recording aggregated crime information in Portuguese text. Asharef et al (2012) proposed a rule-based Arabic NER approach for crime-related entity extraction using morphological information, predefined and general crime indicator lists, and Arabic named entity corpus. Arulanandam et al (2014) used Conditional Random Field (CRF) to classify theft-related sentences and extract crime locations from news articles in New Zealand. Shabat and Omar (2015) employed classification algorithms for crime NER and type identification tasks. Goncalo C. et al (2019) trained an NER module to identify entities such as person, organization, location, and date from Portuguese police reports and online news about crime. Umair et al. (2020) extracted 900 crime data from 8years news archive of Pakistan using NLP and performed hotspot based spatial analysis to predict the behavior of criminal networks using two different classifiers namely K-Nearest Neighbor (KNN) and Random Forest algorithm.

While these studies have contributed significantly to crime analysis and mitigation in places all over the world, there is a call for replication in Nigeria as languages varies, this study therefore aims to use reliable

and reputable print media to wet the floor in the Nigerian context for future investigation into the automatic crime surveillance techniques for a better and efficient crime interventions.

3. Methodology

This study involved the Natural language processing and exploratory data analysis (Geo-spatial EDA) of ten years' worth of kidnapping news articles from the Punch newspaper webpage, one of the top daily news outlets in Nigeria as explained below:

3.1 Data Acquisition:

The research methodology involved the collection of data from the Punch online website, Web scraping techniques was utilized to extract the hypertext markup language (HTML) content from the Punch newspaper webpage. The search query term 'kidnap' was used to gather kidnap news articles. The data was extracted using the 'request library,' a standard python module for sending http requests, and the 'BeautifulSoup library' for parsing the returned HTML structure. The extracted information, including the publishing date, news headline, and full content, was stored in a pandas dataframe for analysis.

3.2 Filtering:

The selection process involved filtering out articles that were not directly related to kidnap incidents, such as those covering rescue operations or general discussions of kidnappings. To achieve this, we used the spacy 'en_core_web_trf' a transformer model to perform Part of Speech (POS) tagging and Dependency parsing on all retrieved headlines. This allowed us to identify the root verb of each headline and lemmatize it to remove inflections. We then checked whether the resulting word was synonymous with "kidnap", using synonyms such as "abduct" and "whisk-away".

a. Data Preprocessing: Entity Recognition and Information Retrieval
   To identify crime entities mentioned in news articles. Firstly, we have segmented each article into sentences and analyzed the syntactic relationship between elements of each sentence. To identify the victims, we have considered the words with a direct object relationship with the verb "kidnap" and its synonyms in each sentence. Furthermore, we have extracted all the elements present in the subtree to account for compound objects. The same approach has been employed to extract the kidnappers using nominal subject relationship with the "kidnap" verb and its synonyms in the sentence.
b. Geoparsing
   i. Toponym recognition: After segmenting the sentences and filtering for those related to kidnappings, we extracted all locational phrases introduced by a locational preposition. This step was necessary to avoid mistakenly including locations mentioned in unrelated sentences. We then removed all stop words in every entity recognized and filtered for capitalized entities that met the criteria for a proper noun. Finally, we combined all split locations as a single string.
   ii. Toponym resolution: We used the Google geocoder through the geopy library to geocode the locations identified, this works well with our locations in that, it was able to resolve locations despite not being arranged properly, it was also able to locate street level location to obscure places and mis-spelt location names.

3.3 Exploratory Data Analysis:

The publishing date of each article was assumed to be the date of the kidnapping event because most news articles are published online within a few hours of the event. The generated dataframe was then converted into a geodataframe to perform geospatial analysis. The Nigerian state and local government administrative boundaries[1] were spatially joined with the kidnap geodataframe using the 'within' operation parameter of the geopandas sjoin method to capture the boundary of crime locations.

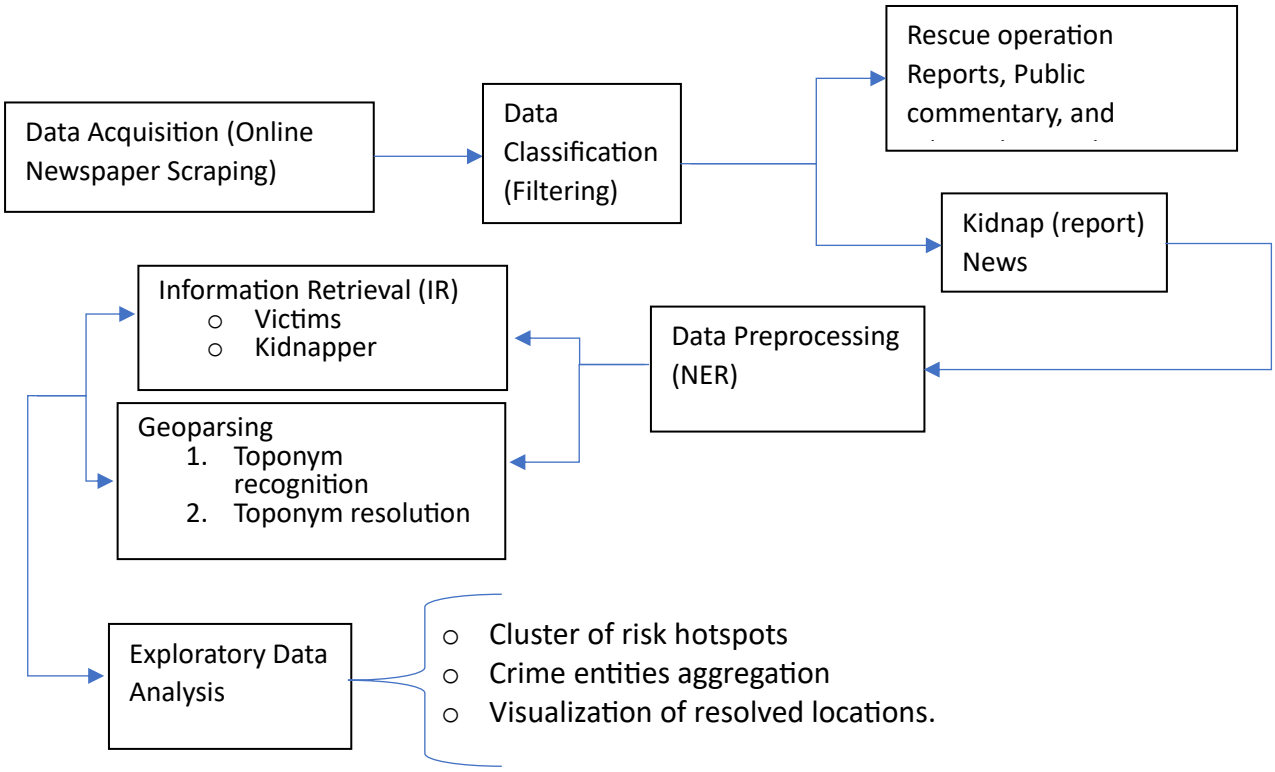

Figure 1: Diagram showing the proposed workflow.

Table 1: Sample of Dataset generated

publish_date	headline	victims	kidnapper	address
5/9/2023	Bandits kidnap 40 Kaduna worshippers, nine emirâ€™s children	40 Kaduna worshippers, nine emir â€™s children	Bandits	Chikun, Kaduna, Nigeria
4/7/2023	Gunmen kidnap Nasarawa ex-deputy governor, Gye-Wado	Nasarawa ex-deputy governor, Gye-Wado	Gunmen	960134, Wamba, Nasarawa, Nigeria
3/12/2023	Gunmen kidnap nine in Abuja estate	nine	Gunmen	Grow Home Estate, Bwari Area council, 901101, FCT, Nigeria

[1] Local government Areas (LGA) and States Boundary curated from the Humanitarian Data Exchange (HDX), last updated by The World Bank on Jan 19, 2023 from https://datacatalog.worldbank.org/search/dataset/0039368

4. Results and Conclusions

Our analysis revealed the spatio-temporal analysis of kidnapping cases in Nigeria, from the clustering of crime locations, with Kaduna State emerging as the most affected region. We observed that Kaduna's high kidnapping rates could be attributed to factors such as overpopulation and high cost of living, as noted in other high-risk states like Lagos, Rivers, and Federal Capital Territory FCT.

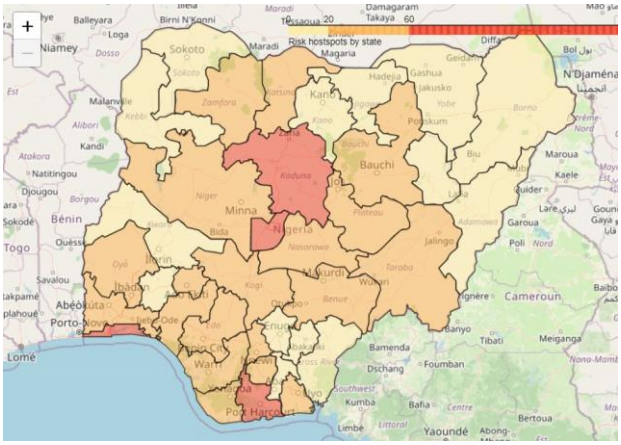

Figure 2: Choropleth of Nigerian states' social risk hotspots.

Examining the trends of kidnapping incidents also highlight the negative effect COVID-19 pandemic had on the socio-economic conditions of Nigerian states. The pandemic-induced hardships, including business disruptions, income loss, and limited access to resources, potentially contributed to an increase in kidnapping incidents as individuals sought alternative means of survival during this period.

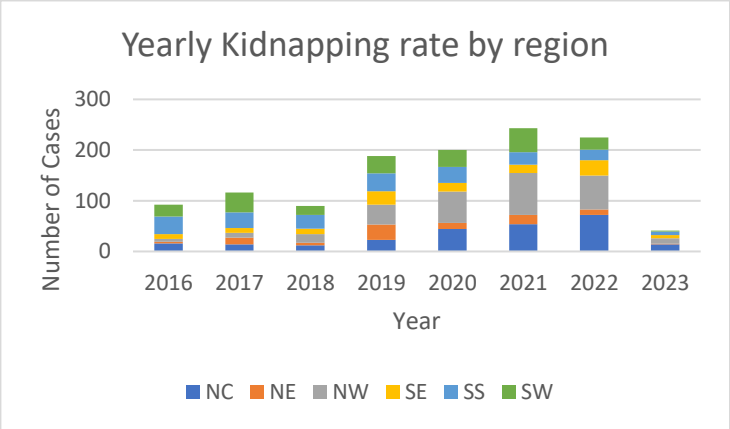

Figure 3: Yearly kidnap rate by zones

Future work will focus on improving feature extraction from newspaper articles beyond the syntactic entities to circumstances of crime events and other crime types like theft, rape, riot, murder etc. while also extending this surveillance technique to other major African languages like Hausa, Igbo and Yoruba which can finally be augmented with a speech to text pipeline to handle audio reported crimes.

References

[1] Al-Shoukry, S., & Omar, N. (2015). Arabic Named Entity Recognition for Crime Documents Using Classifiers Combination. International Review on Computers and Software. https://doi.org/10.15866/irecos.v10i6.6767

[2] Arulanandam, Rexy & Savarimuthu, Bastin Tony Roy & Purvis, Maryam. (2014). Extracting crime information from online newspaper articles.

[3] Asharef, M., Omar, N., & Albared, M. (2012). Arabic named entity recognition in crime documents. Journal of Theoretical and Applied Information Technology, 44(1), 1–6.

[4] Ikenwa, C. (2023). 10 Most Expensive Cities in Nigeria to Live in (Cost of Living 2023). Nigerian Infopedia. Retrieved from https://nigerianinfopedia.com/10-expensive-cities-in-nigeria-to-live-in/

[5] Oyelere, J. (2023, March 14). Scraping News Articles on Kidnapping using Python & BeautifulSoup. [Medium Blog Post]. Retrieved from https://medium.com/@joyelere/scraping-news-articles-on-kidnapping-using-python-beautifulsoup-e5970adec709.

[6] Pinheiro, V., Furtado, V., Pequeno, T., & Nogueira, D. (2010). Natural language processing based on semantic inferentialism for extracting crime information from text. 2010 IEEE International Conference on Intelligence and Security Informatics. https://doi.org/10.1109/isi.2010.5484783

[7] Xuke Hu, Zhiyong Zhou, Hao Li, Yingjie Hu, Fuqiang Gu, Jens Kersten, Hongchao Fan, and Friederike Klan. 2022. Location reference recognition from texts: A survey and comparison. 1, 1 (July 2022), 35 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn