University of Zurich[UZH]

Extracting Structured Flood Information for Nigeria from News Media in GDELT

GEO 511 Master's Thesis

Author
Eric Tharmalingam
17-711-235

Supervised by
Dr. Thomas Phillips (phillitp@colorado.edu)

Faculty representative
Prof. Dr. Ross Purves

30.04.2024
Department of Geography, University of Zurich

Extracting Structured Flood Information for Nigeria from News Media in GDELT

GEO 511 Master's Thesis

Author
Eric Tharmalingam
17-711-235

Supervised by
Dr. Thomas Phillips
Prof. Dr. Ross Purves

Faculty representative
Prof. Dr. Ross Purves

30.04.2024
Department of Geography, University of Zurich

Acknowledgements

I would like to express my gratitude to the following individuals, whose support made my journey during this thesis significantly less difficult:

I am grateful to Dr. Thomas Phillips for guiding me through this journey and for his close mentorship throughout the entire project. Discussing a matter with you has always helped me to view a problem from a different perspective. Further, I would like to extend my gratitude to Prof. Dr. Ross Purves for guiding me in the scientific process and providing your expertise during our meetings.

A huge thank you goes to you, Wensi Zhang, for being there for me throughout my entire thesis. I truly value the way you critically analysed my statements and always providing constructive feedback. Thank you for always lending an ear to any problems I encountered while programming, which helped to resolve some issues faster than I initially thought. Finally, I am very grateful for your meticulous proofreading.

I would also like to thank Linus Rüegg and Tim Waldburger for manually annotating Nigerian place names in news articles. Yet, I am even more grateful for our discussions at work, which sparked the initial idea to automate the entire annotation process.

A big shoutout to all the people generously making data and tutorials available on the internet at no cost, making this thesis possible in the first place.

Finally, I sincerely thank my family and friends for their endless support and encouragement throughout this journey.

Abstract

Floods in African countries are becoming more common and Nigeria, the most populous country in Africa, is no exception. In 2022, Nigeria experienced its worst flooding in a decade, impacting many of its states. These floods, likely exacerbated by climate change altered weather patterns and anthropogenic activities, pose significant risks to Nigerian communities and ecosystems, potentially leading to economic damage and the displacement of populations. The exposure of the population to floods can be approximated using flood-related data from past events. However, this requires access to reliable data and information sources. This thesis presents a NLP-based pipeline for extracting flood-related information from news media in the GDELT database, particularly for data scarce regions such as Nigeria. A custom trained text classification model is utilised to identify newspaper articles relevant to floods from GDELT. Subsequently, a NER model, tailored specifically for Nigeria, is employed to identify place names within these articles. This model is trained thanks to a novel approach developed to fully automate the generation of annotated training data. Ultimately, a rule-based approach is applied to extract quantitative information from news articles. The text classification model attained an F1-score of 80%, while the NER model achieved an accuracy of 86%. Initial spatial analyses for the studied period revealed, that place names primarily from southern Nigerian states, including Anambra, Bayelsa, Delta, Imo and Rivers, were frequently mentioned in flood-related news articles. On a test dataset, the extracted quantitative information achieved a cosine similarity of 78%.

Contents

1 Introduction **1**
- 1.1 Motivation . . . 1
- 1.2 Research Gaps . . . 4

2 Background **6**
- 2.1 Flood . . . 7
  - 2.1.1 Fluvial Flood . . . 7
  - 2.1.2 Pluvial Flood . . . 8
  - 2.1.3 Coastal Flood (Storm Surge) . . . 10
- 2.2 Impact Semantics . . . 10
- 2.3 Disaster Management Cycle . . . 12
- 2.4 Newspapers as a Data Source . . . 14
  - 2.4.1 Reporting Style of Newspapers about Flood Disasters . . . 15
  - 2.4.2 Flood-Related Information in Newspapers . . . 16
  - 2.4.3 Structured Information Extraction from Newspapers . . . 16
- 2.5 Information Science . . . 18
  - 2.5.1 IR and IE . . . 18
  - 2.5.2 GIR . . . 19
  - 2.5.3 NLP . . . 23

3 Data **26**
- 3.1 GDELT - GKG . . . 26
- 3.2 Preprocessing of GDELT GKG Data . . . 27

4 Methods **31**
- 4.1 Text Classification . . . 32
  - 4.1.1 Labelled Training Data Generation for Text Classification 33
  - 4.1.2 Text Classification Model . . . 35
  - 4.1.3 Text Classification Model Evaluation . . . 36
- 4.2 Place Name Recognition and Resolution . . . 38
  - 4.2.1 Automated Nigerian Place Name Annotation for NER Training Data Generation . . . 39
  - 4.2.2 NER Model . . . 42
  - 4.2.3 NER Model Evaluation . . . 44

4.2.4 Toponym Resolution of Extracted Place Names 46
4.2.5 Point Pattern Analysis 46
4.3 News Article Publication Date 47
4.4 Numeric Attribute-Value Pairs 48
4.4.1 Development of AV Extraction Rules 49
4.4.2 Testing and Comparison of Extracted AV Pairs 54

5 Results **57**
5.1 Text Classification 57
5.1.1 Text Classification Model Training Process and Performance Analysis 57
5.1.2 Applicability of the Text Classification Model to GDELT GKG News Articles 60
5.2 Place Name Recognition and Resolution 62
5.2.1 NER Model Training Process and Performance Analysis . 63
5.2.2 Applicability of the NER Model to GDELT GKG News Articles 67
5.3 News Article Publication Date 76
5.4 Numeric Attribute-Value Pairs 76

6 Discussion **80**
6.1 Text Classification 80
6.1.1 Text Classification of News Articles in the Validation and Test Datasets 81
6.1.2 Text Classification of GDELT News Articles 81
6.1.3 Enhancing the Text Classification Model: Limitations and Recommendations 83
6.2 Place Name Recognition and Resolution 84
6.2.1 Toponym Recognition 84
6.2.2 Toponym Resolution 90
6.2.3 Spatial Clustering of Analysed Point Pattern 92
6.3 News Article Publication Date 94
6.4 Numeric Attribute-Value Pairs 95
6.4.1 Comparison with GPT-4 AV Pairs 95
6.4.2 Comparison with GDELT AV Pairs 95

7 Conclusion **98**

A GPT Conversation **115**

B AV Rules (Text) **117**

List of Abbreviations

ADAM	Adaptive Moment Estimation
AI	Artificial Intelligence
API	Application Programming Interface
AV	Attribute-Value
BERT	Bidirectional Encoder Representations from Transformer
DB	Database
DBSCAN	Density Based Spatial Clustering of Applications with Noise
FCT	Federal Capital Territory
FN	False Negative
FP	False Positive
GDELT	Global Database of Events, Language and Tone
GIR	Geographic Information Retrieval
GKG	Global Knowledge Graph
GPT	Generative Pretrained Transformer
HTML	Hypertext Markup Language
IE	Information Extraction
IFRC	International Federation of Red Cross
IR	Information Retrieval
LGA	Local Government Area
ML	Machine Learning
MUC	Message Understanding Conferences

NEMA	National Emergency Management Agency
NER	Named Entitiy Recognition
NLP	Natural Language Processing
OSM	Open Street Map
RQ	Research Question
POS	Part-Of-Speech
RNN	Recurrent Neural Network
TN	True Negative
TP	True Positive
UD	Universal Dependencies
URL	Uniform Resource Locator

List of Figures

1.1 Flood affected Nigerian states between 1970 and 2011 based on Muili and Ikotun (2013), data: ©OpenStreetMap. 2

2.1 Thematic embedding of this thesis. Numbers in brackets indicate the section, in which the respective thematic foundation is discussed. 7
2.2 Schematic representation of fluvial flood processes by Zurich (2023). 8
2.3 Schematic representation of pluvial flood processes by Zurich (2023). 9
2.4 Schematic representation of coastal flood processes by Zurich (2023). 10
2.5 Components of risk based on Crichton (2002). 12
2.6 Disaster management cycle based on Todd (2011). 13
2.7 spaCy NLP pipeline for NER and text classification (class) based on spaCy (2024). 24
2.8 spaCy visualisation of POS tags and dependency arcs (spaCy, 2024). 24

3.1 Overview of all carried out filtering and preprocessing steps. . . . 29
3.2 HTML structure of news article on DailyPost. The title of news articles is contained under the `<h1>` tag. 30

4.1 Overview of the methodological concepts to be introduced in this chapter. Numbers in brackets indicate the sections, in which the respective methods are discussed. 31
4.2 Overview of labelled data generation, text classification model training and model evaluation. Numbers in brackets indicate the sections, in which the respective parts of the workflow are discussed. 32
4.3 HTML structure of news articles on DailyPost. The title of news articles is contained under the `<h1>` tag. 34
4.4 spaCy configuration file settings for an accuracy optimised'textcat _multilabel' pipeline. 36

4.5 Confusion Matrix showing the relationship between predicted and true news article classifications. Green boxes show the news articles correctly predicted either as flood relevant (True Positive, TP) or not flood relevant (True Negative, TN). Red boxes show the incorrectly predicted news articles (False Negative, FN) and (False Positive, FP). 37
4.6 Overview of automated annotated data generation, NER model training, model evaluation and point pattern analysis. Numbers in brackets indicate the sections, in which the respective parts of the workflow are discussed. 38
4.7 Overview of automated annotated data generation. Nigerian place names from OSM and GeoNames were used to be included in synthetically generated news articles. A string matching algorithm automatically annotated the Nigerian place names in the news articles, which were then ready to be used for NER model training. 39
4.8 Query used to download Nigerian place names along with latitude and longitude values from OSM. 40
4.9 Structure of automatically annotated news articles, a JSON object with two elements: a string representing the news article and a dictionary containing information about the entities mentioned in the string. The entities are annotated with their starting and ending index positions within the news article, along with their entity type 'GEO' representing Nigerian place names. 42
4.10 spaCy configuration file settings for a transformer based NER pipeline. 44
4.11 Confusion Matrix showing the relationship between predicted and true Nigerian place names. Green boxes show the correctly predicted tokens, whereas red boxes show the incorrectly predicted tokens. 45
4.12 Visualisation of the potential disadvantages of the token level evaluation method. Here, York gets counted as correctly predicted. However, York on its own is a different place name than New York. 45
4.13 Visualisation of spaCy's POS tagging and dependency parsing used to develop rules to automatically extract numeric AV pairs from news articles. 49
4.14 Overview of the developed algorithm to extract numeric AV pairs in news articles. 50
4.15 Cosine similarity comparison setup between AV pairs provided by GPT-4 and extracted by rule-based method from the same synthetic news articles. Here shown with an example (rectangular boxes), where compound metric is larger than 30. Therefore, assumption was made that correct AV pairs within an article were compared with each other. 56

5.1 Training progress of the text classification pipeline classifying news articles into flood and not flood-relevant news articles. The blue line indicates the loss value over epoch iteration steps. . . . 58
5.2 Confusion matrix of the trained text classification model tested on the validation dataset containing 1530 labelled news articles. . 59
5.3 Confusion matrix of the trained text classification model tested on the test dataset containing 1532 labelled news articles. 59
5.4 Confusion matrix of the trained text classification model tested on 632 news articles in GDELT GKG. 61
5.5 Percentage of flood-relevant topics covered in the news articles from the GDELT test dataset. A: TP news articles. B: FN news articles. 62
5.6 A: Loss values of both NER pipeline components after tuning the hyperparameter Maxout Pieces from two to three. B: Loss values of both NER pipeline components after tuning the hyperparameter Hidden Layers from 64 to 40 with a set Maxout Pieces value of three. 64
5.7 Impact of tuning the hidden layer hyperparameter value on the loss value of the NER pipeline components. The loss changes drastically between the hidden layer values of 35 and 45. 65
5.8 Confusion matrix of the trained NER model tested on the validation dataset containing 278 news articles. The numbers represent the classified tokens of all articles within the validation dataset. . 66
5.9 Confusion matrix of the trained NER model tested on the test dataset containing 278 news articles. The numbers represent the classified tokens of all articles within the test dataset. 66
5.10 Confusion matrix of the trained NER model tested on the manually annotated test dataset containing 100 flood-relevant news articles contained within the GDELT GKG dataset. The numbers represent the classified tokens of all articles within the test dataset. 68
5.11 Nigerian place names extracted by GDELT and the trained NER model classified according to Nominatim's search ranks. 70
5.12 Percentage of manually analysed candidates, for which Nominatim could not provide a search rank. 70
5.13 Map indicating the spatial error for Nigerian place names geocoded by Nominatim and a manual disambiguation process. 71
5.14 Distribution of spatial error for Nigerian place names geocoded by Nominatim and a manual disambiguation process. 72
5.15 Place names mentioned in flood-related news articles clustered in hexagons representing the count of geocoded place names inside each hexagon. 73
5.16 Three by three grid used to count the quadrat statistics for the underlying point pattern. 74
5.17 Visualisation of Ripley's G (A) for the point pattern in B. 75
5.18 Spatial clusters for the analysed point pattern using DBSCAN. . 75

5.19 Percentage of thematic categories for numeric AV pairs extracted from 50 news articles. 79

6.1 Trained NER model accurately detecting Nigerian wards in a news article. 85
6.2 Trained NER model not recognising Nigerian state names. 86
6.3 Trained NER model recognising Kwara State as part of a more detailed geographical description and wrongly labelling Kwara State Fire Service as place name. 86
6.4 Normalised confusion matrix of spaCy's 'en_core_web_lg' model predicting Nigerian place names. 87
6.5 Trained NER accurately differentiating between cases, where Miga is used as a Nigerian place name and a person's name. 89
6.6 Example visualising, how other place names found in the same news paper article can help to correctly disambiguate place names with geo/geo ambiguity such as Daura. 91
6.7 Example visualising, how other place names found in the same news paper article can help to correctly disambiguate place names with geo/geo ambiguity such as Daura. 92
6.8 Flood affected LGAs in Adamawa, Borno and Yobe between June and August 2022 (IOM, 2022). 93
6.9 Flood affected Nigerian states 2022 (IFRC, 2022). 94
6.10 Complex dependency structure of two numeric values part of a AV pair structure (Example 1). 96
6.11 Complex dependency structure of two numeric values part of a AV pair structure (Example 2). 97

List of Tables

2.1 Summary information extracted by various studies. Columns represent various analysed studies and rows represent, what kind of flood-related information was extracted in each study. 17

3.1 Description of GKG columns relevant for this thesis. 27
3.2 Description of the SQLite DB columns and their data types. . . . 28
3.3 Available unique GKG news articles for the months May - December before and after filtering for Nigeria-relevant news papers. 29

4.1 Distribution of labelled data for training (TRAIN, 70% of overall dataset), validation (VALID, 15%) and testing (TEST, 15%). . . 34
4.2 Distribution of annotated data for training, validation and test datasets used for place name recognition. 43
4.3 Description of dependency tags used to split sentences and to extract numeric AV pairs. 51

5.1 Performance metrics of trained text classification model tested on the validation and test dataset. 60

5.2 Performance metrics of trained text classification model on GDELT GKG test data. 61

5.3 Performance metrics of the trained NER model on the validation and test dataset. 67

5.4 Performance metrics of trained NER model on GDELT GKG test data. 68

5.5 Classification of place types by the Nominatim geocoder and their associations with administrative units in Nigeria. 69
5.6 Average cosine similarity per AV pair and per article based on 50 news articles comparing GPT-4 AV pairs to rule-based AV pairs and rule-based AV pairs to GPT-4 AV pairs. 77

5.7 Three examples of GPT-4 and rule-based AV pairs and their respective cosine similarity score. 78

6.1 Three GDELT and rule-based AV pairs exemplifying the improvement of the rule-based AV pair extraction method over the algorithm used by GDELT. 96

Chapter 1

Introduction

1.1 Motivation

Climate change is expected to lead to an increase in the frequency and severity of natural disasters such as hurricanes, floods, and heatwaves. The intensity and severity of floods have been increasing in countries, where they were rare or non existent in the past. These changes in weather patterns and anthropogenic influence are likely to have significant impacts on human communities, ecosystems and therefore can lead to economic losses and displacement of people (Pörtner et al., 2022).

Increased occurrence of extreme rainfall events driven by rising climate variability and change leads to a higher probability of flood occurrences (Adelekan and Asiyanbi, 2016). However, floods are not only caused by climatic conditions, but also due to its interactions with social, economic and political factors. Rapid urbanisation and population growth necessitates well-planned and managed infrastructure, as well as effective policies and governance frameworks to mitigate the consequences of flood events on the global population (Okyere et al., 2013). Floods in African countries are becoming more common and Nigeria, the most populous country in Africa, is no exception. Since the end of the Nigerian civil war and the accompanying oil boom in the 1970s, the country has become an increasingly urban society. People living in urban areas have gradually increased from 16% in 1970 to more than 20% in 1980 (Adekola and Lamond, 2018). This percentage increased to 53% in 2021 (Worldbank, 2023). Nigeria's urban population is expected to continue to grow relatively fast in the coming decades.

According to Bashir O. et al. (2012) the urban poor of Nigeria are the most vulnerable to flood impact, as they cannot afford land within the cities and build informal settlements in the low-lying floodplains of Lagos, such as Ibadan or Abeokuta. Other characteristics of the urban poor are that a large number of them work in the informal sector or are unemployed and live at the border of a more traditional and more modern world. Additionally, the urban poor suffers from a higher mortality rate due to insufficient healthcare and polluted environmental conditions (Chaudhuri, 2015). Uncontrolled population growth, poor infrastructure and inappropriate waste management intensify the negative impact of climate change on the frequency of floods (Bashir O. et al., 2012).

Since the 1970s up until 2011, multiple flood disasters were recorded, where people in different Nigerian states such as Abia, Adamawa, Akwa Ibom, Bayelsa, Delta, Edo, Kano, Lagos, Oyo, Taraba and Zamfara were affected (Figure 1.1). During those disasters buildings, infrastructure and properties were either damaged or destroyed (Muili and Ikotun, 2013). Therefore, flood occurrence is not a novelty in Nigeria.

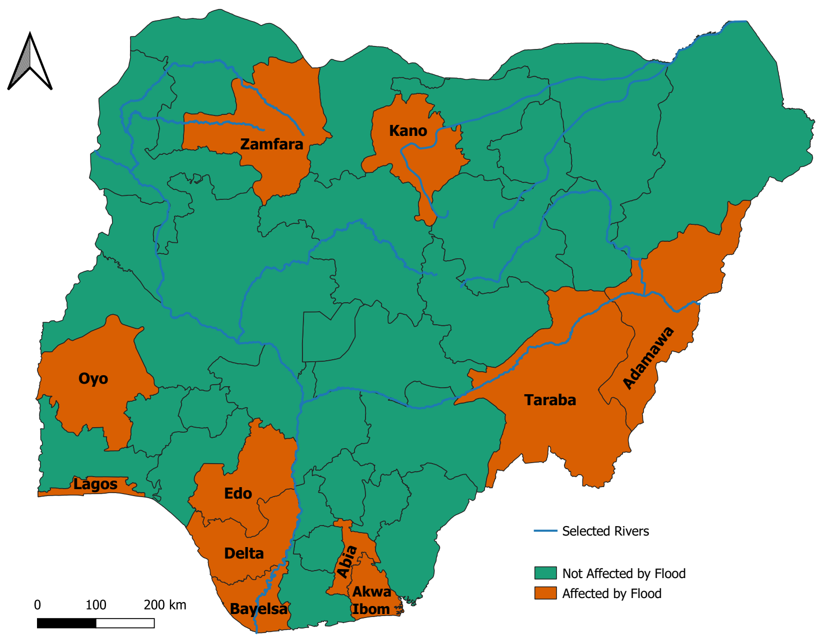


Figure 1.1: Flood affected Nigerian states between 1970 and 2011 based on Muili and Ikotun (2013), data: ©OpenStreetMap.

Nevertheless, an increase in the intensity of flood disasters in recent years (2007, 2012, 2018 and 2022) is clearly evident (Adelekan, 2011; Emeka et al., 2023; Osayomi et al., 2018; Urama et al., 2019). The floods share common characteristics, including an unanticipated influx of above-average rainfall that affected residents in 30 or more of Nigeria's 36 states. Surveys unveiled that a majority of the population was taken by surprise, receiving no or insufficient prior warning about the floods (Adelekan, 2011). This lack of prewarning systems significantly amplified the floods' impact on their livelihoods and the country's economy (Adelekan, 2011; Amangabara and Obenade, 2015; Gambo, 2018; Williams, 2022).

The exposure of the population to flood hazards can be approximated using flood-related data from past events. However, this requires access to reliable data and information sources (Guha-Sapir et al., 2011). Data collection methods vary across regions and countries based on available resources, infrastructure and institutional capabilities. Additionally, advancements in technology and data analysis techniques continue to enhance the understanding of past flood events.

Governmental organisations like Nigerian Meteorological Agency and the Nigeria Hydrological Services Agency make an effort to provide accurate weather, climate and hydrological data in a timely manner by maintaining and expanding their observation networks, acquiring the latest weather monitoring systems available and the training of personnel (Hussaini and Matazu, 2023). The Nigerian government has initiated several open data portals and platforms like the Nigeria Open Data Access portal to provide access to a wide range of data. However, this portal contains only 181 datasets, of which only 0.6% are related to disaster control. Another major issue is the poor update cycle of data in the portal (Ezema, 2023).

High resolution satellite data is commonly used for flood mapping (Bauer-Marschallinger et al., 2022; Cohen et al., 2022; Twele et al., 2016). Various remote sensing data with near global coverage is publicly available. However, mapping flooded urban areas remains a challenging problem due to its dense and complex terrain (Pierdicca et al., 2018). Additionally, the processing of radar imagery is quite labour intensive and conservative estimations assume that 50% of all optical satellite images are cloud covered, limiting their usage for mapping the extent of flood disasters (Li and Roy, 2017). Research conducted by Panteras and Cervone (2018) suggests the use of social media data (Twitter) to overcome the temporal limitations of satellite data.

Social media data has become an increasingly valuable source of information for improving natural catastrophe models. Especially in developing countries, the use of social media and internet increased between 2013 and 2018 (Poushter et al., 2018). This has led to an explosion of user-generated content, including text, images and videos, that can provide real-time information about natural disasters as they occur. However, since Elon Musk took over Twitter in October 2022, twitter data is no longer publicly available for research purposes(Hickey et al., 2023).

In order to obtain the most comprehensive understanding of a flood disaster, it is essential to utilise multiple available data sources. Relying on a single source can limit our perspective and potentially introduce biases or inaccuracies.

1.2 Research Gaps

GDELT (Global Database of Events, Language and Tone) is a machine-coded database (DB) of events that incorporates information from both international and domestic newspapers. The repository continuously monitors news media in over 100 languages worldwide, providing a comprehensive collection of news articles updated every 15 minutes. The GDELT project contains over 250 million geocoded events since 1979 (GDELT, 2022). However, the raw text of the referenced newspapers is not included in GDELT. GDELT consists of several sub datasets.

Williams (2020) found that two United Kingdom disasters of 2015, Storm Desmond and a major flood, could be timely observed in GDELT. As Williams (2020) showed, GDELT can be used to analyse various disasters retrospectively. Especially in developing countries, where the data situation is sometimes somewhat limited, GDELT could be a good addition to for example numerical hydrological models, as the impact on people can also be taken into account. Yet, to the best of my knowledge, GDELT has not yet been used to analyse flooding in Nigeria.

However, the transparency of GDELT's algorithms remains uncertain, making it difficult to assess their effectiveness in categorising information within articles. Additionally, the extent of media outlet coverage monitored by GDELT is also uncertain (Williams, 2020).

To summarise, this thesis therefore addresses two different research gaps:

1. To use GDELT as data source for past flood disasters in Nigeria and

2. to systematically extract flood-related information in an automated fashion for flood disasters in Nigeria and compare it with the flood-related information extracted by GDELT.

This thesis aims to fill these gaps by addressing the following research questions (RQ).

- [RQ1] How does the flood-related information extracted by GDELT compare to those presented in this thesis, both sourced from the same data?

- [RQ2] Can flood-related information from news articles in the GDELT DB be systematically extracted to identify flood-related hotspots in the study area of Nigeria?

To answer RQ1 the following sub-questions need to be answered iteratively.

- [RQ1.1] How can the largest number of flood-relevant news articles from GDELT be extracted for a specific area of interest?

- [RQ1.2] Is it feasible to extract place name information with a greater level of spatial detail than the place name data available in GDELT?
- [RQ1.3] Can the GDELT timestamp reliably be used to filter news articles published within an interested time frame?
- [RQ1.4] What kind of quantitative information is most commonly reported in flood-related news articles?

Chapter 2

Background

In the year 2022, Nigeria witnessed its most severe flooding in the past decade. 34 out of Nigeria's 36 states were affected by the 2022 floods. The majority of the states affected, heavily depend on agriculture, which was impacted by said floods. The death toll exceeded overall 600 and 1.4 million individuals were displaced. Concerns arose about a potential disruption in the food supply of the most populous country in Africa, as the disaster had caused extensive damage to more than 500'000 hectares of crops. Although Nigeria experiences annual flooding, the floods of 2022 brought unparalleled misery and destruction, surpassing the intensity of the worst floods to date, which occurred in 2012 (National Bureau of Statistics, 2023). To comprehend the correlation between this thesis and the problem at hand, this chapter provides an overview of the thematic framework, in which this thesis is embedded (Figure 2.1). As this thesis deals with flood-related news articles in Nigeria, the first step will involve investigating the causes of flooding in the country (Section 2.1). However, not every flood has a major impact on people and infrastructure. Therefore, Section 2.2 will address the required factors that turn a flood into a disaster. If such a flood disaster occurs, many countries and first responders rely on a disaster management cycle concept. This concept will be briefly introduced in Section 2.3 and it will be elaborated, how flood-relevant information, extracted from newspaper articles, can be used in a disaster management cycle in order to help manage a disaster. Section 2.4 will summarise, what kind of flood-relevant information can actually be extracted from news articles. Finally, Section 2.5 will introduce the research areas and methods involved in computationally extracting flood-relevant information from news articles.

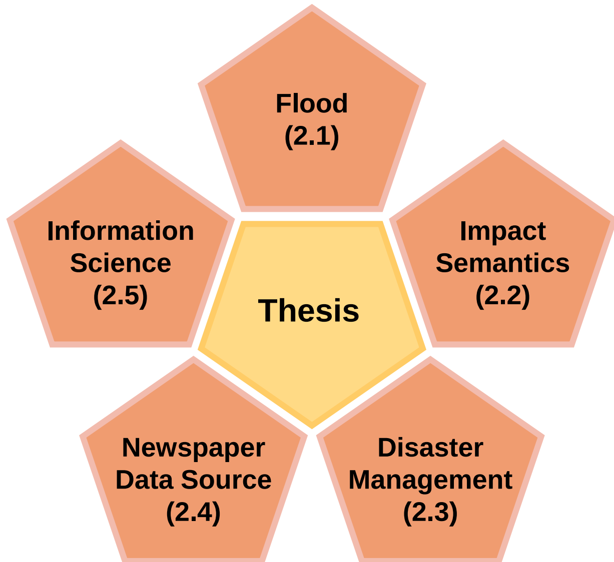


Figure 2.1: Thematic embedding of this thesis. Numbers in brackets indicate the section, in which the respective thematic foundation is discussed.

2.1 Flood

Floods are typically categorised into three types: fluvial, pluvial and coastal. Each category is discussed in more detail within the context of Nigeria in the subsequent sections. However, it is also possible for different types of floods and their causative factors to coincide, leading to what are known as compound floods (Heinrich et al., 2023).

2.1.1 Fluvial Flood

Fluvial floods or river floods occur, when the water level in a river, lake or stream rises and overflow their banks, inundating the neighbouring area and forming the waterbody's floodplain (Figure 2.2). One of the most severe river floods in Nigeria occurred during the Kano state flood disaster of 2006, impacting hundreds of thousands of lives and resulting in economic losses totaling millions of US dollars (Nkwunonwo, 2016).

This situation can occur, when the volume of water from intense rainfall surpasses the natural limit of the river channel, particularly worsening, if the channel is obstructed or restricted (Luino, 2016). In estuarine regions, elevated tide levels can obstruct the river's discharge into the sea. When flood waves

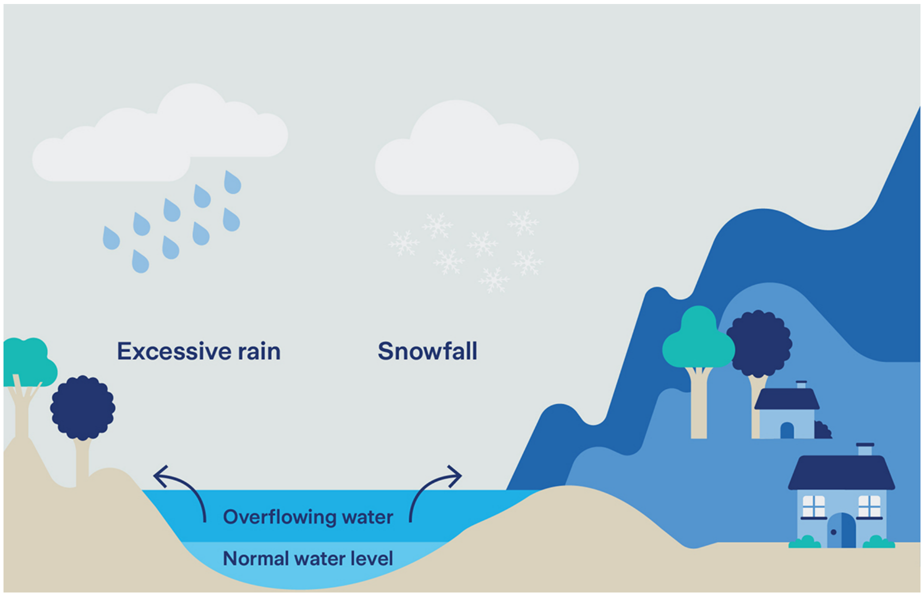


Figure 2.2: Schematic representation of fluvial flood processes by Zurich (2023).

from various tributaries converge at river confluences simultaneously, the resultant downstream flood can significantly exceed the magnitude of the individual floods. In both 2012 and 2013 Lokoja, located at the confluence of the river Niger and Benue, suffered from fluvial flooding (Badaru et al., 2014). Various rivers exhibit distinct responses to above average water inputs, influenced by several factors, including physical characteristics, such as the catchment's size and gradient, water saturation level of the landscape or soil and rock permeability. Anthropogenic alterations, such as urbanisation levels, land use changes and the capacity for floodwater retention and gradual release into lakes and floodplains further influence the discharge characteristics of rivers (Wagener et al., 2007). Dams are in place to reduce flood peaks. However, if dams retention capacity is reached or dams breach, extraordinary floods can occur, such as the disaster in 2010 affecting the Nigerian states Jigawa, Sokoto and Kebbi (Nkwunonwo, 2016).

2.1.2 Pluvial Flood

Pluvial flooding arises, when rainfall surpasses the capacity of urban stormwater drainage systems or the ground's ability to absorb it. The surplus water then spreads over the surface, accumulating in natural or man-made depressions and low-lying regions or obstructed areas (Figure 2.3).

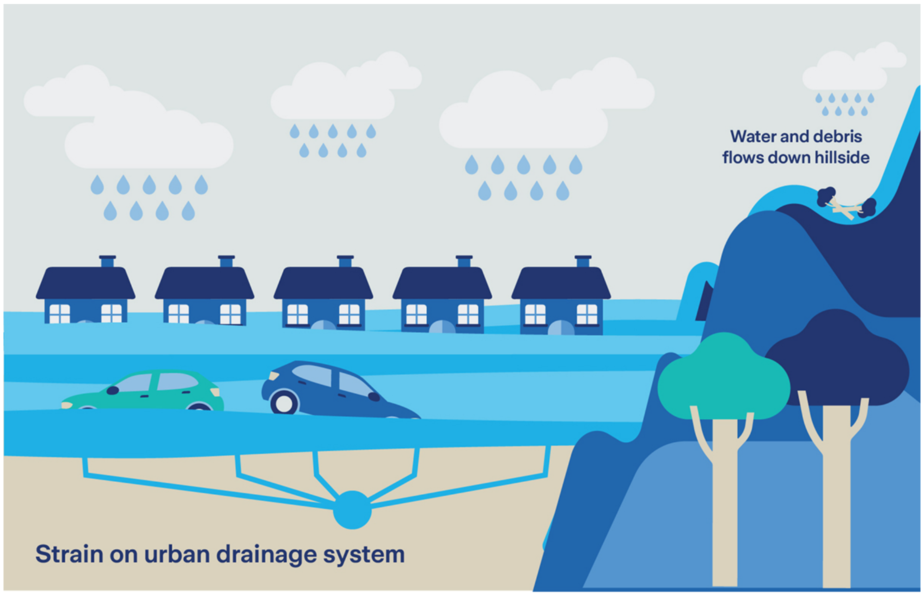


Figure 2.3: Schematic representation of pluvial flood processes by Zurich (2023).

This occurs swiftly in response to heavy rainfall, preceding the eventual entry of floodwaters into piped or natural drainage networks. Pluvial flooding is notably propelled by brief yet intense rainstorms (Rosenzweig et al., 2018). A prevalent misunderstanding about pluvial floods is the belief, that proximity to a water body is necessary to be at risk. However, pluvial flooding can occur anywhere, whether urban or rural, even in regions absent of nearby water bodies (Rözer et al., 2016). There are two common types of pluvial flooding. Surface water flooding happens, when an urban drainage system becomes overburdened, leading water to spill into streets and nearby buildings. This process unfolds gradually, allowing people time to relocate to safety and the water level typically remains shallow (Kaźmierczak and Cavan, 2011). Flash floods are marked by a sudden, forceful surge of water, propelled by heavy rainfall within a brief period of time, either locally or on nearby higher ground. They may also arise from the abrupt discharge of water from an upstream dam. Flash floods pose significant danger and devastation, not only due to the water's force, but also because of the swirling debris often carried along with the flow (Hapuarachchi et al., 2011). In 2005 and 2011 the Jalingo Local Government Area (LGA) was affected by flash floods due to extreme rainfall over Jalingo and the hills surrounding the town. Especially, houses built on the flood plains of the Mayogwoi and Lamurde rivers were destroyed, as the flash flood overflowed the river banks (Oruonye, 2012).

2.1.3 Coastal Flood (Storm Surge)

Coastal flooding refers to the submergence of land areas along the coast due to seawater. Typical triggers include powerful windstorms coinciding with high tide (storm surge) and tsunamis (Figure 2.4).

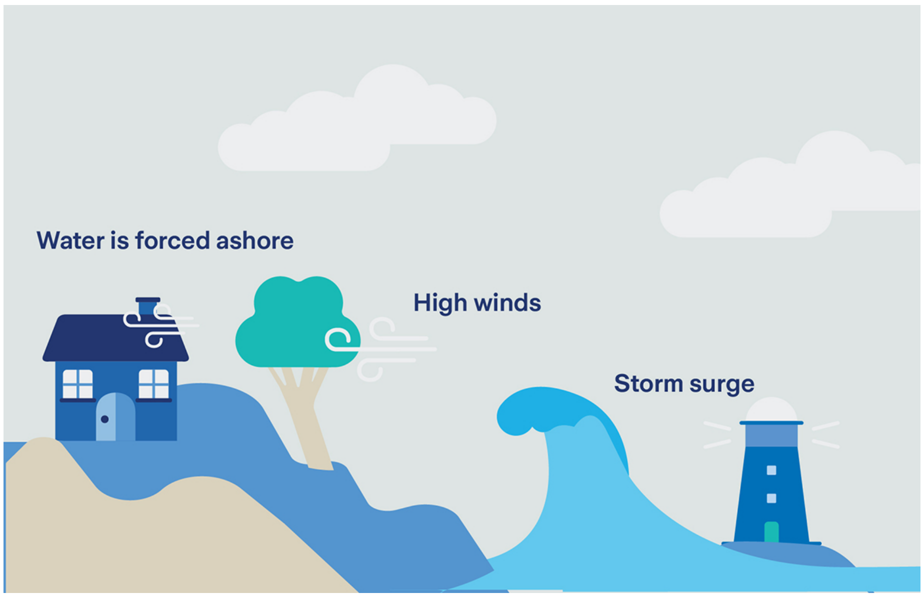


Figure 2.4: Schematic representation of coastal flood processes by Zurich (2023).

Another important factor leading to coastal floods is sea level rise due to climate change (Luque et al., 2021). Additionally, the coastal and offshore terrain plays a crucial role (Haigh et al., 2017). During such floods, water inundates low-lying areas, resulting in severe loss of life and property. In 2012, a devastating coastal surge struck Lagos Beach, resulting in significant destruction. Sixteen lives were lost and the entire beach community had to be evacuated. This problem is intensified by strong urbanisation in the area. The Lekki Peninsula for example is situated within the barrier-lagoon system along the Lagos coastline. Since 1980, the peninsula has experienced rapid urbanisation, despite its unique physical attributes being largely overlooked (Obiefuna et al., 2021; Sholademi et al., 2015).

2.2 Impact Semantics

In Section 2.1, the types of floods were described. However, for a flood disaster to occur, a flood hazard must first coincide other factors. For this reason, this section introduces the semantics of hazard, vulnerability and exposure, detailing how all three factors must converge to constitute a high flood risk, potentially resulting in a disastrous flooding. The definitions and ideas introduced in this section all follow the UN/ISDR (2009) (United Nations International Strategy for Disaster Reduction), unless specified otherwise.

Hazard

'A dangerous phenomenon, substance, human activity or condition that may cause loss of life, injury or other health impacts, property damage, loss of livelihoods and services, social and economic disruption, or environmental damage'(UN/ISDR, 2009).

Hazards are commonly considered from an anthropocentric viewpoint and are generally infrequent, unforeseen and abrupt in their nature (Gebhardt et al., 2011). The scientific perspective on natural hazards encompasses the natural processes that could be seen as the underlying causes of the observed natural disasters (Gebhardt et al., 2011). It is important to understand the difference between hazard and risk. Flood hazards can be high, while flood risk is low. A flood-prone region, where advanced and effective flood mitigation infrastructure is in place has a high flood hazard, but a lower flood risk.

Vulnerability

'The characteristics and circumstances of a community, system or asset that make it susceptible to the damaging effects of a hazard'(UN/ISDR, 2009).

The vulnerability of an element depends on their likelihood of being exposed to hazards and their capacity to withstand them (Dilley and Boudreau, 2001). The level of vulnerability of a population depends on the given hazard at a given severity level as well as the population's individual capabilities to absorb and recover from an impact (Susman et al., 2019). A wooden shed in comparison to a concrete building with flood protection in place has a higher vulnerability to be completely destroyed by a flooding. However, the vulnerability would be reduced and the resilience of the wooden shed would be increased, if it would be upgraded according to more flood proof building standards. Therefore, resilience measures: 'The ability of a system, community or society exposed to hazards to resist, absorb, accommodate, adapt to, transform and recover from the effects of a hazard in a timely and efficient manner[...]'(UN/ISDR, 2009).

Exposure

'People, property, systems, or other elements present in hazard zones that are thereby subject to potential losses'(UN/ISDR, 2009).

Measures of exposure can include the number of people or types of assets in an area. People are exposed to a flooding, when they are in the flooded area or are affected by the aftermaths of the flooding. Increased population and assets in a flood-affected area result in elevated exposure (UN/ISDR, 2009).

Risk

'The combination of the probability of an event and its negative consequences.' (UN/ISDR, 2009)

Risk should be seen as the overlap of hazard, exposure and vulnerability as shown in Figure 2.5.

Figure 2.5: Components of risk based on Crichton (2002).

Disaster

'A serious disruption of the functioning of a community or a society involving widespread human, material, economic or environmental losses and impacts, which exceeds the ability of the affected community or society to cope using its own resources' (UN/ISDR, 2009).

Other terms synonymously used with disaster are calamity and catastrophe. It is important to understand that disasters are a product of interactions between time, space and a vulnerable human population. Disasters are triggered by hazards. Hazards only have the potential to cause negative consequences, whereas a disaster signifies that the potentially negative consequences have become reality due to the occurrence of a hazard (Schneiderbauer and Ehrlich, 2004).

2.3 Disaster Management Cycle

Disaster management can a be defined as an overarching term containing a range of activities in place to maintain the control over an emergency or disaster. Disaster management is a framework, which should guide all involved people during the various stages of a disaster. As more and more countries move from more traditional approaches, which mainly focus on disaster relief immediately after a disaster happened to more recent frameworks, improving the resilience of people and infrastructure in order to minimise the number of loss and damage, available data and information become more and more important and are crucial parts of disaster management Bali (2024). The National Emergency Management Agency (NEMA) of Nigeria developed its first version of their National Disaster Response Plan in March 1999 in order to 'establish a process and structure for the systematic, coordinated, and effective delivery of Federal assistance, to address the consequences of any major disaster or emergency declared by the President of the Federal Republic of Nigeria' (NEMA, 2002).

A common concept found in the realm of disaster management is the disaster management cycle. This cycle was created to depict the continuous procedure through which involved stakeholders can build up strategies and mitigate the impact of disasters, plan responses during and immediately after a disaster and implement recovery measures following a disaster (Coetzee and Van Niekerk, 2012). Various disaster management cycles with different compositions regarding the number of included phases are currently in use (Coetzee and Van Niekerk, 2012). The World Bank (Todd, 2011) and the Government of India (NDMI, 2011) for example use a disaster management cycle composed of three phases: pre-disaster phase, disaster response phase and post-disaster phase (Figure 2.6).

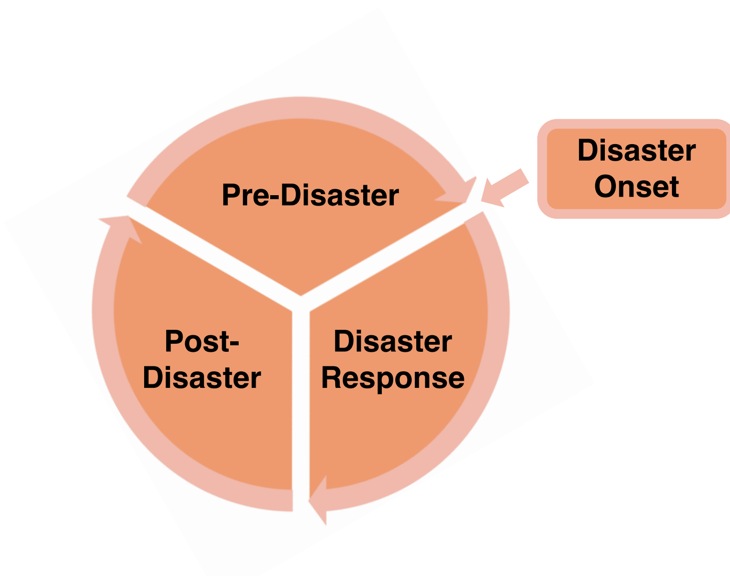


Figure 2.6: Disaster management cycle based on Todd (2011).

Due to the interconnections between phases, phases are mutually inclusive and multidimensional (Lettieri et al., 2009). The **pre-disaster phase** consists action items related to mitigation and preparedness. Whereas mitigation efforts are aimed to minimise the degree of risk, preparedness focuses on preparing responders and common people to mid- and post-disaster activities. The impact of floods can be partially mitigated due the availability of past flood data (Bali, 2024; Todd, 2011). Lu et al. (2024) for example presented an enhanced flood mitigation system for road networks based on news media, which was used for vulnerability assessment. The phase dedicated to **disaster response** involves implementing measures to handle and mitigate the diverse effects of a disaster with the goal of minimising both human and property losses. The disaster response involves executing plans and procedures developed under the pre-disaster phase. Immediately after the disaster, steps need to be initiated to evacuate the population living in the disaster zone, coordinate search and rescue missions and establish a clear and concise picture of the situation after the onset of the disaster. Additionally, intermediate communication and supply chains are required between the victims and responders. Social media data can be used by first responders, for example to have an additional source of information in order to maintain an overview of the disaster situation (Muniz-Rodriguez et al., 2020). Even global databases of historic and real-time flood events can be established using social media data (de Bruijn et al., 2019). During the **post-disaster phase**, efforts are made to reinstate conditions equal to or improved from those existing before the disaster (Bali, 2024; Todd, 2011). Bohensky and Leitch (2014) for example used news articles after the 2011 Brisbane, Australia flood in order to understand the perceived links between the flood and climate change and perceived roles of government in managing the flood.

This shows that appropriate data and information during the disaster management cycle is crucial and helps understanding the hazard and disaster, as well as supporting accurate communication and decision making (Lettieri et al., 2009). Voigt et al. (2007) state that rapid satellite mapping campaigns can support disaster relief processes, providing application examples following events such as the Indian Ocean Tsunami or the earthquake in Pakistan. Mass media has a unique role in providing information and raising awareness among people in crisis situations (Ghassabi and Zare-Farashbandi, 2023). Finally, Rausch (2014) showcase that local newspapers reflect on specific disaster related issues in their newspaper columns.

2.4 Newspapers as a Data Source

Newspapers provide comprehensive coverage of various topics, including politics, economics, social issues, disasters and weather events. Researchers can analyse news articles and reports to gather information, statistics, quotes and expert opinions related to specific events or phenomena (Ashlin and Ladle, 2007; Boettke et al., 2007; Hulme and Burgess, 2019). Compared to other disaster related data sources, newspapers offer several advantages. Firstly, newspapers provide continuous disaster information for an extended time range. Secondly, information in newspapers is relatively reliable. Most newspaper authors extract their information from authoritative sources. Thirdly, especially local newspapers can provide more information on small- and medium-scale events not present in large-scale disaster databases. Finally, newspaper archives are usually more accessible than other data sources (Du et al., 2015). However, news articles can also be driven by sensationalism. Brown et al. (2018) found that not only topics such as crime and lifestyle are sensationalised, but also topics like government affairs and science. Additionally, news agencies can apply a selective approach to decide, which topics are reported on, thus having a major influence on what topics are classified as newsworthy and therefore get published in news articles (Bennett and Townend, 2012). Compared to exact measurements like river gauge data, flood-related information extracted from newspapers is undoubtedly less precise. There is also no standardised reporting format for flood disasters in newspapers. Different articles focus on different topics related to a flood disaster. Nevertheless, historical news articles can provide valuable details about water level, flood extent or damage caused (Archer et al., 2019).

There is a substantial increase in flood-related news stories since 2000 due to climate change, improved reporting and accessibility (Devitt and O'Neill, 2017; Escobar and Demeritt, 2014). Adekola and Lamond (2018) found a similar trend for newspapers reporting about flood disasters in Nigeria. The number of flood articles generally peak around August-October, during the wet season in Nigeria (Adekola and Lamond, 2018; Gambo, 2018).

2.4.1 Reporting Style of Newspapers about Flood Disasters

A newspaper discourse study comparing two flood disasters found that national media has the tendency to decontextualise disasters, by framing them as discrete disasters. This allows journalists to present disasters as manageable problems by concentrating on the events themselves rather than exploring the underlying causes and mitigation strategies. On the other hand, local newspapers tend to provided more in-depth information about the disasters (Rashid, 2011). To describe a flood disaster, local newspapers for example use references to known local sights, which got destroyed during that disaster (Solman and Henderson, 2019).

Traditionally, news articles on flood disaster claim to report descriptively, contain factual information and describe flood disaster discretely (Devitt and O'Neill, 2017). However, Dor (2003) found, that for example headlines are designed to optimise the relevance of news articles. This includes quoting victims and using dramatic headlines (Escobar and Demeritt, 2014; Solman and Henderson, 2019). The state-of-knowledge, the beliefs and expectations of the readers are required to create a successful headline (Dor, 2003). Therefore, disaster related news articles tend to focus on the region related to the disaster. A foreign flood disaster tends to be seen as newsworthy, when nationals from the newspaper's country are victims of the flood (Gambo, 2018; Solman and Henderson, 2019). Solman and Henderson (2019) also found that United Kingdom newspapers reinforce similarities and shared values between victims and assumed readers in articles about local floods by drawing upon personal stories, emotions and suffering. Articles about foreign floods are on the other hand more distant. Nevertheless, the bulk of broadsheet newspapers claim still to be dedicated to descriptive news reports (Escobar and Demeritt, 2014; Solman and Henderson, 2019).

2.4.2 Flood-Related Information in Newspapers

The type of flood-related information extracted from newspapers can be manyfold. Rashid (2011) grouped all extracted information into the following thematic groups: hydro-meteorological characteristics of the flood, reports on flood damage, reports on emergency measures and reports on flood alleviation measures. Most keywords found were related to reports on emergency measures followed by keywords related to flood alleviation measures and hydro-meteorological characteristics. This finding is also supported by Gambo (2018), who conducted a Nigerian newspaper framing analysis for the 2012 flooding disaster. Most news articles were addressing types of relief efforts followed by facts about the flooding itself, containing insights about the time of occurrence, causes for the flood and details about the damage caused. Studies by Devitt and O'Neill (2017) and Escobar and Demeritt (2014) indicate that immediate descriptive news articles were followed by articles focusing on topics like government commitment towards flood prevention in a very politicised environment. Politicisation is characterised by conflicts between actors on political issues, possibly applying pressure on political leaders or policy makers. Such politicised articles can be used to shape social relations, condition political power and projects for change. Additionally, news articles can frame disaster management actions as a success or failure (Albrecht, 2022).

2.4.3 Structured Information Extraction from Newspapers

To ensure effective flood preparedness plans, it is crucial to enhance the comprehension of the consequences that floods have on communities. Hence, building flood resilience becomes vital for communities at risk of flooding to effectively handle future flood disasters. By examining historical flood disaster patterns and accurately projecting casualties and damages, it becomes feasible to comprehend the vulnerability of communities. In the domain of flood resilience analysis, standardised data collection plays a crucial role. To incorporate flood events from newspapers into a DB, its specific details need to be condensed into a concise set of descriptive parameters (Haltas et al., 2021; Kron et al., 2012).

Andres and Badoux (2019) for example focused on Swiss newspapers and magazines to establish a DB containing information about flood and landslide disasters between 1972 and 2007. Relevant information (Table 2.1) was extracted by an external media-monitoring company. According to Andres and Badoux (2019), it is feasible to examine newspapers with a focus on various elements, including spatial aspects, temporal dimensions, weather conditions and impacted objects. The DB contains 36 years worth of data for the whole of Switzerland with a good spatial resolution down to community level. However, the amount of damage can only be estimated and the information found in news articles is often incomplete or incorrect. Diakakis et al. (2012), Gil-Guirado et al. (2019), Vennari et al. (2016) and Zêzere et al. (2014) used a similar approach to covert unstructured disaster information into a structured DB for different regions in Europe (Table 2.1). Llasat et al. (2009) concluded that news articles can be used to get indirect estimations of a natural hazard disaster probability and its negative consequences.

Table 2.1: Summary information extracted by various studies. Columns represent various analysed studies and rows represent, what kind of flood-related information was extracted in each study.

Extracted Information	Greece Diakakis	Italy Vennari	Portugal Zêzere	Spain Gil-Guirado	Switzerland Andres Badoux
Location	x	x	x	x	x
Date	x	x	x	x	x
Hazard Type	x	x	x	x	x
Intensity				x	x
Trigger			x		x
Dead/Injured	x	x	x		x
Affected Object	x	x	x	x	x
Event Description	x	x			x
Catchment Size		x			

Andres and Badoux (2019), Diakakis et al. (2012), Gil-Guirado et al. (2019), Llasat et al. (2009), Vennari et al. (2016) and Zêzere et al. (2014) relied on manual information extraction techniques to populate their databases. In recent times, there has been a surge of interest and activity in the utilisation of computational methods to extract valuable insights from diverse collections of documents. This body of research encompasses multiple domains, including Natural Language Processing (NLP), Information Retrieval (IR), Information Extraction(IE) and Geographic Information Retrieval (GIR), which are intricately intertwined (Liu et al., 2019; Nadkarni et al., 2011a; Purves et al., 2018a). By employing NLP tools that streamline the initial assembly of a DB constructed from information in news articles, automatic text analysis can assist in replacing previously labour-intensive and time-consuming manual methods. The objective of text mining is to extract information from unstructured text and transform it into structured data, enabling the information to be organised in an easily searchable format (Lai et al., 2022; Wang and Stewart, 2015; Yzaguirre et al., 2016).

Kahle et al. (2022) compared manual interpretation of newspapers with automated text mining and NLP techniques and showed that a high consistency exists between manual and automated methods for data analysis. After extracting the general topic of a news article with NLP techniques, Panem et al. (2014) presented algorithms based on linguistic tools like Standford Typed Dependencies for numeric and textual attribute-value (AV) pair extraction from tweets. The authors hypothesise that news agencies use tweets to put out very technical information about natural hazard events. Authors of news articles on the other hand focus more on the location of the event or the number of people dead or injured.

2.5 Information Science

Since this thesis deals with the automated classification of newspaper articles and the extraction of place names and numerical AV pairs, the current state of research in these and related areas of IR, IE, GIR and NLP will be considered in this section.

2.5.1 IR and IE

Calvin Mooers first introduced the term IR in 1950 (Swanson and others, 1988) and Manning and Schütze (2009) define IR as: 'Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers)'. IR extends its scope beyond the parameters outlined in the primary definition. Examples for other unstructured information sources are: pictures, drawings, websites or sound. Additionally, it addresses a variety of data and information challenges. The term 'unstructured data' denotes information lacking a clearly defined, semantically explicit structure that is readily understandable by computers. This stands in contrast to structured data, exemplified by relational DBs (Bartschi, 1985; Manning and Schütze, 2009).

According to Sarawagi (2008) 'Information Extraction refers to the automatic extraction of structured information such as entities, relationships between entities, and attributes describing entities from unstructured sources.' IE should not be confused with IR. Cowie and Lehnert (1996) describe the difference between IR and IE based on the processing steps required to extract relevant information from newspaper articles: 'IR systems can collect the articles with relevant text. IE starts with a collection of such texts, then transforms them into information that is more readily digested and analyzed. It isolates relevant text fragments, extracts relevant information from the fragments, and then pieces together the targeted information in a coherent framework.'

The field of IE experienced rapid growth from the late 1980s onwards, with some works tracing their origins back to the 1960s. Funded by the US Navy the Message Understanding Conferences (MUC) propelled advancements in IE throughout the 1980s and 1990s. Various teams initially competed in challenges to extract information from naval messages. At each MUC conference, diverse techniques for IE were presented and discussed, encompassing rule-based systems as well as statistical methods. Additionally, evaluation methodologies for assessing the performance of IE systems were established. This included the introduction of standardised evaluation datasets, metrics and protocols to facilitate the comparison and benchmarking of different systems (Cowie and Lehnert, 1996; Gaizauskas and Wilks, 1998).

IE should not be viewed as a standalone technology, but rather as closely interconnected with related fields such as IR and NLP. Current advancements in NLP are propelling the research of IR and IE to new heights (Khurana et al., 2023).

2.5.2 GIR

Just as language can convey countless ideas by arranging words in accordance with rules, space can likewise be intricately described using the following basic components: proper nouns (e.g., Zürich, Greenwich Park, Limmatplatz), common nouns (like lake, city, street) and prepositions (such as in, at, near, between) (Acheson, 2019). Place names or toponyms are commonly categorised as a subset of proper nouns, referring specifically to places. Proper nouns belong to a linguistic category denoting specific individuals, such as people and locations and typically necessitate capitalisation in English and numerous other languages (Bennett and Agarwal, 2007). Toponyms are not the only way of referring to geographic entities or locations. Other examples are: place codes (address or postal codes), descriptions ('the largest city in England' or 'the highest mountain in Europe') and complex geographic phrases ('the bakery between the church and the river' or '100m after the big junction in the village') (Leidner and Lieberman, 2011).

GIR is a relatively young research area and originates from IR. The field of GIR has continued to evolve since it was first defined by Larson in 1996 and emphasising the geographical aspect of IR. Larson (1996) defined GIR for the first time as: 'Geographic information retrieval (GIR) is concerned with providing access to georeferenced information sources'. This rather general definition was further specified by Jones and Purves (2009) as: 'The provision of facilities to retrieve and relevance rank documents or other resources from an unstructured or partially structured collection on the basis of queries specifying both theme and geographic scope.' Jones and Purves (2009) specify that GIR deals with unstructured or only partially structured data and mainly focuses on the following challenges (Jones and Purves, 2009):

1. The detection and unique assignment of toponyms to geographical locations.

2. The ambiguous geographic terminology reflecting, how individuals perceive, resonate with and communicate through vague spatial concepts, often lacking clear sharp boundaries and how it contrasts with precise geolocation techniques using coordinates.

3. Spatial and textual indexing as well as geographical relevance making (ranking) enabling the formulation and processing of meaningful queries with a geographical reference.

4. GIR poses challenges for map-based visualisations to handle documents with extensive geographic scopes

This thesis primarily focuses on the first challenge. Solving the challenge requires sub dividing the problem into toponym recognition (geoparsing) and toponym resolution (geocoding) (Larson, 1996; McCurley, 2001).

GIR - Toponym Recognition

The aim of geoparsing is the identification of toponyms in unstructured data. The detected toponyms can be words or even whole phrases (Hu, 2018). Leidner and Lieberman (2011) differentiate among three distinct approaches for identifying toponyms in a text:

1. The most basic method involves matching each word of the text to predefined lists. Such lists are called gazetteers and Hill (2000) defines gazetteers as 'geospatial dictionaries of geographic names'. The components name, location and type are generally contained within a gazetteer. The name corresponds to the name of the toponym. The matching generally happens via this name component. The location contains the coordinates of the name, mostly stored as latitude and longitude values. The type component defines the geographic feature of the toponym like populated place or mountain. The quality of a gazetteer can vary greatly between different geographic regions (Acheson et al., 2017; Ahlers, 2013).

2. Expanding on gazetteer matching, employing regular expressions to define patterns that represent common structures of place names in a region could enhance the identification of toponyms within a text. For example, numerous place names in England end with '-ton', such as Luton, Islington, or Conington (Trubshaw, 2012). Utilising such a pattern enables the discovery of toponyms that might not be included in a gazetteer.

3. Finally, one could employ Named Entity Recognition (NER), a sub task of NLP, to determine the probability of whether a word, within a specific context, signifies a place name (see Section 2.5.3).

The task of recognising toponyms may seem straightforward to humans. However, it presents a significant challenge for computers due to the need to additionally resolve semantic ambiguities and vagueness (Ardanuy and Sporleder, 2017; Lieberman et al., 2010).

Ambiguity and vagueness are common features in spatial language. Acheson (2019) explains the difference between ambiguity and vagueness with the following example: '[...] the expression 'We are near London' is ambiguous because London has multiple potential referents (including London, England and London, Ontario, Canada) and is vague because 'near' is highly context-dependent and doesn't have a crisp region of applicability.' Amitay et al. (2004) first use the term geo/non-geo ambiguity and define it as: 'A geo/non-geo ambiguity occurs when a place name also has a non-geographic meaning, such as a person name (e.g., Berlin) or a common word (Turkey)'. If toponyms can describe place name entities as well as non place name entities, then the toponym itself becomes ambiguous in terms of its geo/non-geo ambiguity. Additional layers of ambiguity are added through the metonomical use of toponyms. In such a case the entity representing a toponym would not describe a place name, rather people, events or products related to the place name. In the sentence 'I would like a Montepulciano' Montepulciano represents a type of wine rather than the place name (Leveling and Hartrumpf, 2008; Markert and Nissim, 2002).

GIR - Toponym Resolution

After identifying the corresponding toponyms in the text accurately, whether through gazetteer matching, regular expression rules or NER, appropriate coordinates can then be assigned. Leidner (2007) defines toponym resolution as: 'computing the mapping from occurrences of names for places as found in a text to a representation of the extensional semantics of the location referred to (its referent), such as a geographic latitude/longitude footprint.' Toponym resolution addresses ambiguities too. However, unlike toponym recognition, these ambiguities are not semantic, but rather spatial in nature. Lagos is the biggest city in Nigeria. At the same time Lagos is also a village in France and a city in Faro, Portugal. Kano in the name of a state in Nigeria as well as the name of a city in the state of Kano. Such ambiguity is also termed geo/geo ambiguity (Amitay et al., 2004). Another layer of complexity adding to the challenge of toponym resolution is the presence of multiple names or slight variations in spelling for a single location. A location's name may vary across different languages. Moreover, within the same language, multiple names can refer to the same place. For instance, New York City, NYC and Big Apple all refer to the exact same location. Furthermore, place names for a given area can evolve over time, for example, New York was previously known as New Amsterdam (Ferrés, 2007). The proportion of ambiguous toponyms has already been analysed in various gazetteers. For example Smith and Crane (2001) analysed the Getty

Thesaurus of Geographic Names and the found that Asia has the highest percentage (32.7%) of places with multiple place names and North and Central America have the highest percentage (57.1%) of multiple places with the same place name. In Africa 27.0% of places have multiple place names, whereas 18.2% of different places have the same place name.

There are various methods for resolving this geo/geo ambiguity that are not mutually exclusive, but are almost always used in combination with each other. Buscaldi (2011) and Buscaldi and Rosso (2008) divide these into three different categories:

1. **Map-based:** This method leverages the spatial arrangement of candidate toponyms and, for instance, calculates their spatial distances to accurately identify the correct locations. Smith and Crane (2001) found, that methods relying on computing spatial distances relative to other mentioned toponyms in the text and subsequently selecting candidates with the smallest overall distance, can yield satisfactory outcomes. Brunner and Purves (2008) demonstrated that ambiguous toponyms frequently exhibit an autocorrelated distribution. This aspect should be considered, especially when employing a map-based approach.

2. **Knowledge-based:** This method relies on external sources of information, such as population statistics, gazetteers or online encyclopedias. Purves et al. (2018b) also showcased that many successful approaches utilise additional contextual information to achieve successful toponym resolution. Gazetteers often contain hierarchical information. These can be limited to country affiliation or also include all administrative units. Hierarchical methods utilise the appearance of hierarchically higher toponyms in the text to resolve geo/geo ambiguity (Hauptmann and Olligschlaeger, 1999). To address the ambiguity of places referenced with multiple place names, alternative place names are generally consolidated under a single entry in a gazetteer. This process is known as toponym normalisation (Leidner, 2008). Buscaldi and Magnini (2010) demonstrated that location of newspaper publishers serves as significant contextual information for resolving ambiguities. Pouliquen et al. (2004) on the other hand limited their gazetteer based on the population of each location to reduce its ambiguity. Gale et al. (1992) found that in 98% of cases only one meaning of the word occurs in the same text. Hence, it can be inferred that a word within a text consistently carries the same meaning, also when the word marks a toponym. Amitay et al. (2004) and Hauptmann and Olligschlaeger (1999) used this concept termed 'One Referent per Discourse' for toponym resolution. The majority vote method is often used, when no other toponym resolution methods are effective, however ambiguity is still present. For instance, the toponym 'London' would typically be assumed to refer to the city in England rather than the one in Ontario, Canada, due to the former's greater prominence and familiarity (Willett et al., 2012).

3. **Data-driven:** Methods using supervised and unsupervised machine learning (ML) techniques. Recent methods shifted to ML-based classifiers that use features of the text surrounding a toponym. ML systems are trained on extensive datasets comprising text documents annotated with labelled toponyms and their respective locations. Supervised learning techniques, such as support vector machines or deep neural networks are commonly employed for this task. Through analysing patterns and connections between toponyms and their contexts, these systems gain the ability to precisely interpret ambiguous references and allocate the accurate geographic coordinates (Cardoso et al., 2022; DeLozier et al., 2015; Fize et al., 2021).

Regardless of the methods employed, resolving toponym ambiguities continues to present a significant challenge in both toponym recognition and resolution, warranting further research efforts (Hu et al., 2023).

2.5.3 NLP

NLP is utilised in various methodological segments within this thesis. On one hand it is used to categorise newspaper articles into those relevant to flooding themes and those unrelated to flooding. On the other hand it is also used to extract place names and numeric AV pairs in news articles.

The origins of the field are commonly traced back to the early 1950s, emerging as a subset of both Artificial Intelligence (AI) and linguistics. Its primary goal was to address challenges associated with the automatic generation and comprehension of natural language (Arellano et al., 2015). The introduction of statistical NLP in the 1980s revolutionised the field with techniques like Hidden Markov Models and probabilistic parsing, contributing to enhanced language understanding (Anandika et al., 2021; Chater and Manning, 2006). The late 1990s saw the rise of ML approaches, particularly with the introduction of deep learning methods drawing inspiration from the workings of the human brain. Similar to the interconnected nature of neurons in the brain, neural networks operate in a similar fashion. Each neuron receives an input, processes it internally and generates an output, moving it closer to the anticipated result. This improves computers in mastering complex interpretation tasks (Kulkarni and Shivananda, 2021; Socher et al., 2012). More recent developments employ large, pretrained language models such as Bidirectional Encoder Representations from Transformer (BERT) and Generative Pretrained Transformer (GPT). These models aim to learn a broad understanding of language patterns and semantics, allowing them to perform a wide range of NLP tasks (Min et al., 2023).

A NLP pipeline refers to a sequence of processes or steps applied to a body of text to extract meaning, information or insights from it. These pipelines are designed to take raw text as input and transform it into a structured text for further analysis. spaCy[1] is an open-source NLP python library providing a wide range of tools and functionalities for various NLP tasks. Similar to pipelines guiding the flow of water, spaCy pipelines direct the flow of textual data (Figure 2.7). spaCy offers its own pretrained language models to enable the functionality of such pipelines. spaCy models are trained on general language. Therefore, SpaCy provides the possibility to retrain the models for specific use cases (spaCy, 2024; Srinivasa-Desikan, 2018).

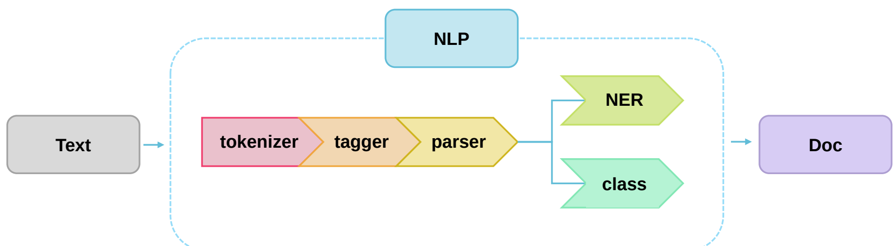


Figure 2.7: spaCy NLP pipeline for NER and text classification (class) based on spaCy (2024).

As illustrated in Figure 2.7, the NLP pipeline contains several components including a tokeniser, tagger, parser, NER and others. Therefore, prior to any processing of the input text, it needs to traverse through each of these components. SpaCy initiates processing by tokenising the text into its sub-units, such as words and punctuation marks. Tokenisation typically occurs at white space boundaries, although it adheres to language-specific rules. For instance, in English, the contraction 'don't' will be separated into 'do' and 'n't' despite the absence of white space within 'don't'. The next processing step is called Part-Of-Speech (POS) tagging. A POS represents the grammatical function of a word within a sentence, indicating its role and dependencies with other words in the sentence (Figure 2.8).

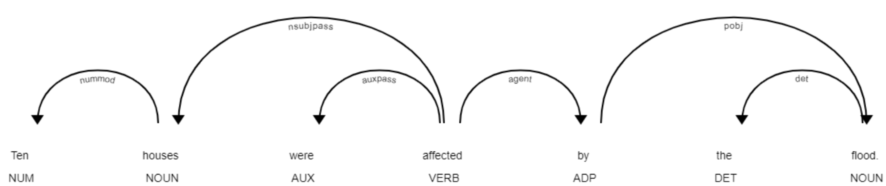


Figure 2.8: spaCy visualisation of POS tags and dependency arcs (spaCy, 2024).

Now the text is ready for further analysis such as NER or text classification.

[1] https://spacy.io/

NER involves categorising text snippets into predefined categories like places, persons or organisations. To train such NER models, training data is utilised to identify the typical contexts in which for example a place name appears (Sharma et al., 2022).

Chapter 3

Data

In order to answer the stated RQs in Section 1.2, the GDELT Global Knowledge Graph (GKG) dataset was used. Section 3.1 further introduces the GKG dataset as a whole and describes the relevant information contained in GKG for this thesis. Preprocessing and further methods that were applied in this thesis are described in Section 3.2 and Chapter 4.

3.1 GDELT - GKG

The GKG dataset comprises of news data from 2013 onwards. Open access to the raw data is provided via GDELT[1]. The text file on the website is regularly updated and includes a new link for each 15 minute update of the GKG data, available for download in .csv format. The rows in the files represent news resources and the columns contain various types of machine extracted information. The content of relevant columns for this thesis are further clarified in Table 3.1. These columns are either relevant in order to get access to the source of a news article or for comparison purposes introduced later in this thesis.

[1] http://data.gdeltproject.org/gdeltv2/masterfilelist.txt

Table 3.1: Description of GKG columns relevant for this thesis.

GKG Column Header	Description
V2.1DATE	Date in YYYYMMDDHHMMSS format, publication date of news article
V2SOURCECOMMONNAME	Human readable external source identifier, e.g. dailypost.ng
V2DOCUMENTIDENTIFIER	Full external source identifier, e.g. full URL to online news article
V1THEMES	Thematic classification by GDELT, e.g. NATURAL_HAZARD_FLOOD, full theme list[2]
V1LOCATIONS	List of all location found, including location name, country code and latitude/longitude value, e.g. Abuja, NG, 9.0833/7.53333
V2.1AMOUNTS	List of all precise numeric amounts found, e.g. 9, people lost their lives

3.2 Preprocessing of GDELT GKG Data

To organise the data downloaded from GDELT and the results derived from the methods outlined in Chapter 4, a SQLite DB was initialised. Access to the DB is provided via GitHub[3]. The columns and their data types in the DB are listed in Table 3.2. The DB includes distinct columns for all relevant GKG data for this thesis, as outlined in Table 3.1. Furthermore, this DB incorporates columns for all the data compiled as part of this thesis. Column FULLTEXT_T holds the scraped raw title and main body text from news articles referenced in V2DOCUMENTIDENTIFIER. The columns THEME_T, LOCATION_T and COORDINATES_T, DATE_T as well as ATTRIBUTES_T contain the information extracted with the methods described in Section 4.1, Section 4.2, Section 4.3 and Section 4.4 respectively.

[2] https://data.gdeltproject.org/documentation/GDELT-Global_Knowledge_Graph_CategoryList.xlsx

[3] https://github.com/etharm/GEO551

Table 3.2: Description of the SQLite DB columns and their data types.

DB Header	Data Type
V2.1DATE	integer
V2SOURCECOMMONNAME	text
V2DOCUMENTIDENTIFIER	text
V1THEMES	text
V1LOCATIONS	text
V2.1AMOUNTS	text
FULLTEXT_T	text
DATE_T	text
THEME_T	text
LOCATION_T	text
COORDINATES_T	text
ATTRIBUTES_T	text

After establishing the DB, the relevant GDELT data was downloaded. Figure 3.1 gives an overview of all carried out filtering and preprocessing steps. As Nigeria experienced in 2022 its most severe floods in the past decade (National Bureau of Statistics, 2023), the masterfile[4] was filtered for Uniform Resource Locators (URLs) containing the year 2022, the numeric representation of all months from May (05) till December (12) and the string 'gkg'. The months from May till December were chosen to cover the extended rainy season in Nigeria (Ibebuchi and Abu, 2023). A python script automatically filtered the masterfile according to specified criteria and downloaded the .csv files through a parallelised process. After removing duplicate URLs in the SQLite DB column V2DOCUMENTIDENTIFIER, GKG yielded 26'338'523 news articles for the selected time frame (Table 3.3). In the next step, all entries containing the terms 'Nigeria' and 'NG' in column 'V1LOCATIONS' were selected. This reduced the relevant number of news articles from 26'338'523 to 368'089 (Table 3.3).

[4] `http://data.gdeltproject.org/gdeltv2/masterfilelist.txt`

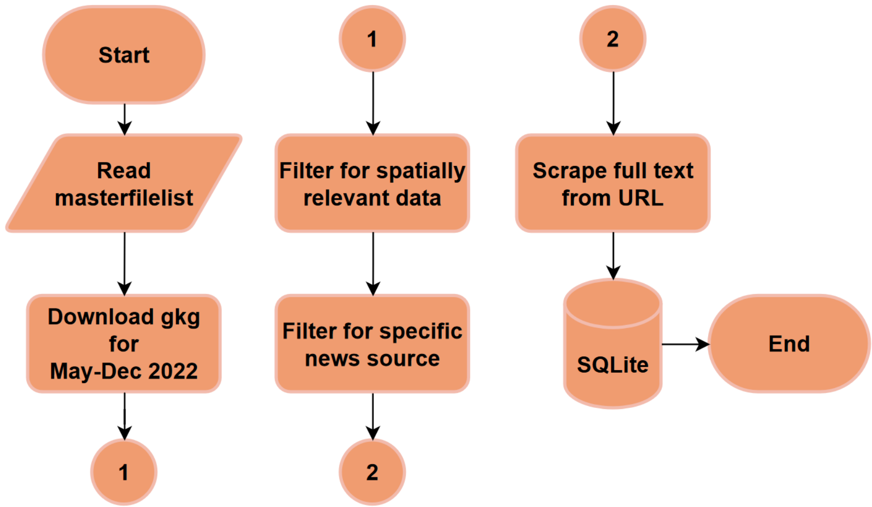


Figure 3.1: Overview of all carried out filtering and preprocessing steps.

Table 3.3: Available unique GKG news articles for the months May - December before and after filtering for Nigeria-relevant news papers.

GKG Month 2022	All News Articles	News Articles about Nigeria
May	3'619'429	56'413
June	3'422'990	51'314
July	3'299'607	48'248
August	3'353'345	44'006
September	3'263'537	43'200
October	3'254'625	43'330
November	3'119'725	41'986
December	3'005'265	39'592
Total	26'338'523	368'089

As a final preprocessing step, the title and full main body text of news articles referenced under V2DOCUMENTIDENTIFIER had to be scraped from their respective websites and stored under FULLTEXT_T in the SQLite DB. However, not all news agencies allow scraping information from their websites. As the focus of this thesis is to demonstrate the feasibility of the whole workflow, only 18'081 news articles (title and main text) from the DailyPost news agency were scraped and stored under the FULLTEXT_T column in the SQLite DB. To accomplish this final preprocessing step, only entries in the DB containing the string 'dailypost.ng' in the column V2SOURCECOMMONNAME were selected. The python library Scrapy (version 2.11.1)[5] allows for scraping information while obeying the scraping rules stated by the news agencies. The scraping rules are generally hosted on the main domain under the /robots.txt path, e.g. for DailyPost[6]. Websites are built using Hypertext Markup Language (HTML). HTML provides a way to structure content on a websites using elements. For example, the element under the tag <p> represents a paragraph, whereas <h1> represents a top level heading (Raggett et al., 1999). DailyPost contains the title of a news article also under the <h1> tag (Figure 3.2). All individual <p> paragraphs of the article main body are kept within a <div> tag, indicating a section of the website. Scrapy was used to access the text within these tags and the extracted text for each available DailyPost news article was stored under the column FULLTEXT_T in the DB.

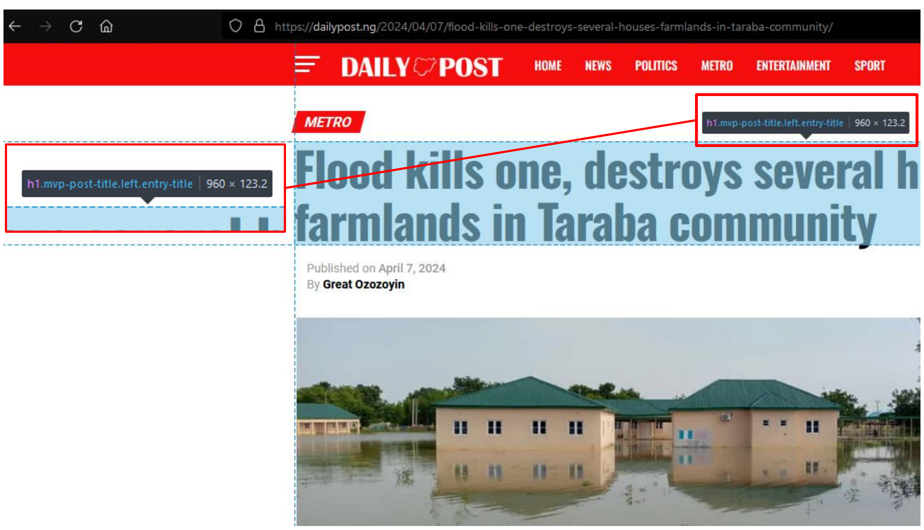


Figure 3.2: HTML structure of news article on DailyPost. The title of news articles is contained under the <h1> tag.

[5] https://scrapy.org/
[6] https://dailypost.ng/robots.txt

Chapter 4

Methods

This chapter provides a detailed overview over the methods applied to analyse and manipulate the data described in the previous chapter. The methods developed in this chapter are intended to generate the necessary results for addressing the RQs outlined in Section 1.2. Each news article in the SQLite DB, which was introduced in Chapter 3, runs through each methodological module sequentially as shown in Figure 4.1. The method in Section 4.1 categorises news articles into those relevant to floods and those that are not, based on their content. For each flood-relevant news article, the method in Section 4.2 extracts Nigerian place names from the news articles and locates them in the geographical space. Additionally, the date on which a newspaper article was published and the numeric AV pairs contained within an article are extracted by the methods described in Section 4.3 and Section 4.4.

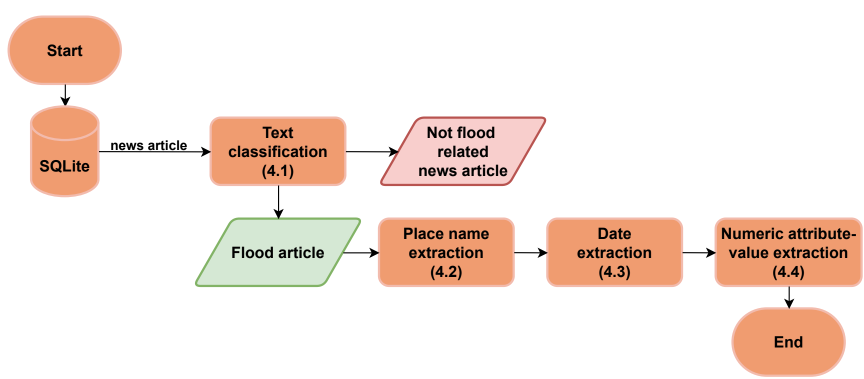


Figure 4.1: Overview of the methodological concepts to be introduced in this chapter. Numbers in brackets indicate the sections, in which the respective methods are discussed.

4.1 Text Classification

Text classification is a NLP task that involves categorising pieces of text into predefined classes. The goal of text classification is to automatically assign labels to textual data based on its content (Kowsari et al., 2019). The objective here is to develop a text classification model that utilises the content of English language news articles to categorise them as either flood relevant or not flood relevant. For such a model to function effectively, it requires training with labelled data. labelled data consists of text documents that have undergone content analysis, resulting in the assignment of specific labels or classes to each document. This labelled text data is then converted into numerical feature vectors that can be processed by ML algorithms. This process is called feature extraction and identifies the most relevant pieces of information from the text that a computer can understand and use to apply labels to unseen pieces of text. These most relevant pieces of information can include words, phrases or patterns that are for example crucial for classifying news articles into flood-relevant and not flood-relevant news articles (Shah and Patel, 2016). ML algorithms are then trained on the labelled data to learn the relationship between these input features and the output labels. Once the model is trained, it can be used to predict the category of new, unseen text data (Luo, 2021).

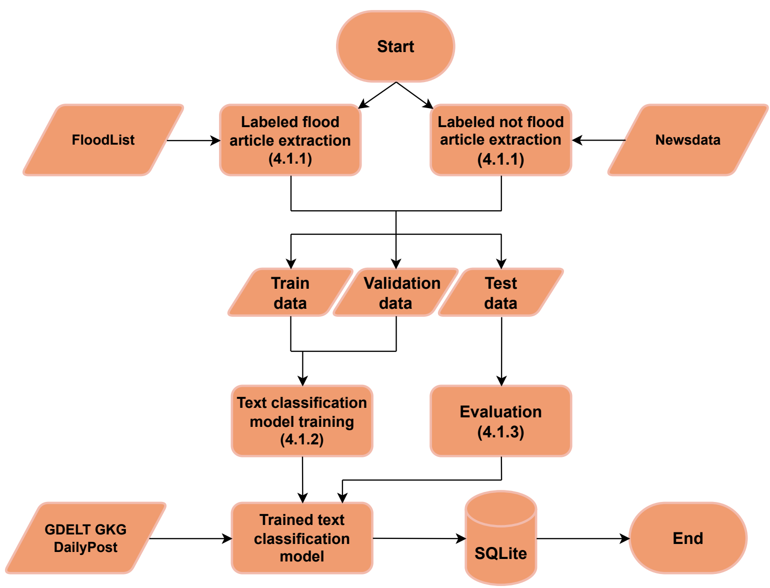


Figure 4.2: Overview of labelled data generation, text classification model training and model evaluation. Numbers in brackets indicate the sections, in which the respective parts of the workflow are discussed.

Figure 4.2 shows the workflow described in this section. Section 4.1.1 describes the generation of labelled data for the model. Section 4.1.2 elaborates on the structure of the text classification model utilised and Section 4.1.3 introduces the framework for evaluating the model's performance.

4.1.1 Labelled Training Data Generation for Text Classification

News articles from FloodList[1] and Newsdata[2] were used to compile labelled data. Articles sourced from FloodList were classified as flood relevant, while those sourced from Newsdata were classified as not flood relevant. To extract the title and main body text from FloodList news articles, a similar approach as described in Section 3.2 was chosen. FloodList also allows scraping information from their website[3]. The python library Scrapy was used to extract titles of FloodList news articles contained within the `<h1>` tag and 'entry-title' class (Figure 4.3). All main paragraphs of a news article can be found under the `<div>` tag and 'entry-content' class. Each scraped news article was stored as a text file. From Newsdata, .csv files were manually downloaded containing URLs to thematically classified news articles. The following themes were chosen to represent not flood-related news articles: business, covid and vaccine, entertainment, health, science and technology and world politics. The assumption was made that these themes do not contain any flood-related news articles. All news articles not in English language were filtered out and from the remaining URLs the content was scraped and stored as separate textfiles. News articles representing not flood-related articles are hosted on various domains. The HTML structure of news articles can vary and relevant content can be contained within various HTML tags. The python library trafilatura (version 1.6.1)[4] was used. The advantage of this library over Scrapy is that the main content of a website can be scraped with a more generalised command than specifying individual HTML tags. However, a drawback is that occasionally irrelevant information is scraped from websites. Here too, only websites that allow scraping were accessed.

[1] `https://floodlist.com/`
[2] `https://newsdata.io/datasets`
[3] `https://floodlist.com/robots.txt`
[4] `https://trafilatura.readthedocs.io`

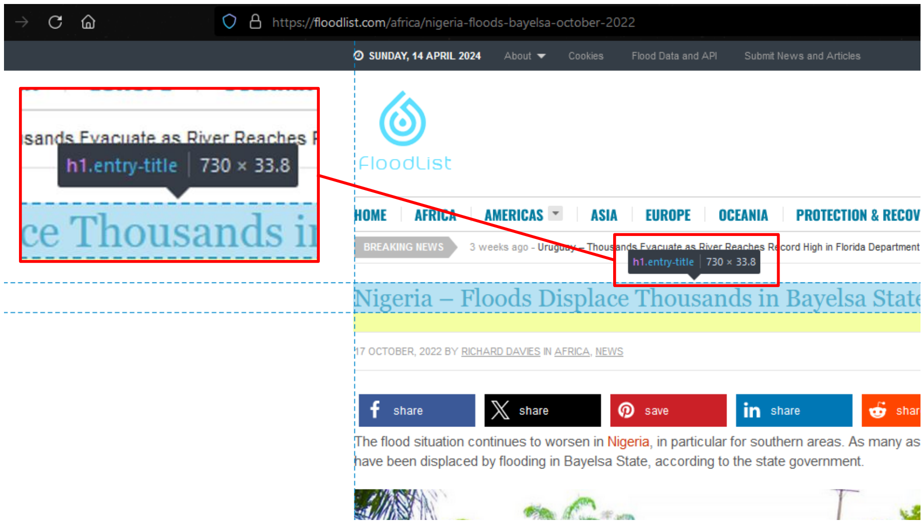


Figure 4.3: HTML structure of news articles on DailyPost. The title of news articles is contained under the `<h1>` tag.

All individual text files of flood-relevant and not flood-relevant news articles were organised into a .json file with the following structure: [[news article text,0],[news article text,1],...]. The structure represents a nested list. Each inner list contains the news article text at index 0 and the number 0 or 1 at index 1. In this context, the label '0' signifies not flood-relevant articles, while the label '1' denotes flood-relevant articles. The .json file was subsequently divided into training, validation and test datasets, as outlined in Table 4.1. The training and validation datasets were used to train the text classification model in Section 4.1.2 and the test dataset was used to evaluate the model as described in Section 4.1.3. This a classic ML approach.

Table 4.1: Distribution of labelled data for training (TRAIN, 70% of overall dataset), validation (VALID, 15%) and testing (TEST, 15%).

Data	**Class**	**Split [%]**	**Split [#]**
TRAIN	Flood Not Flood	35 35	3572 3572
VALID	Flood Not Flood	7.5 7.5	765 765
TEST	Flood Not Flood	7.5 7.5	765 765

4.1.2 Text Classification Model

The python library spaCy (version 3.7.2) was used to train a text classification model based on the labelled data created in Section 4.1.1. Before the training and validation datasets can be used to train the text classification model, spaCy tokenises the text of news articles. Tokenisation involves splitting a text into separate tokens, where each token represents a word, punctuation mark or other meaningful unit. Tokenisation rules vary depending on the language being processed. These rules encompass complex scenarios such as contractions, punctuation and special characters. Hence, spaCy offers a range of pretrained statistical models tailored to specific languages. These models serve as the foundation for tokenisation and are also employed during the training of text classification models. spaCy's 'en_core_web_lg' model[5] is used throughout Section 4.1. This is a English language specific model trained on a large corpus of written online text such as blogs, news and comments. Additionally, this model also incorporates word vectors (spaCy, 2024). Word vectors are numerical representations of words in a high-dimensional space, where each word is mapped to a vector of real numbers. These vectors capture the semantic relationships and contextual information between words based on their usage in large text corpora. Models incorporating word vectors can understand relationships between words, such as synonyms, antonyms and related terms, which enhances their ability to process and interpret text. Word vectors enable models to generalise better on unseen words or contexts (Inan et al., 2017).

spaCy provides two versions of text classification pipelines. The 'textcat_multilabel' pipeline was chosen for this thesis. This pipeline returns for each to be classified news article the following two labels 'flood_article' and 'not_flood_article' accompanied with a value between 0 and 1. If the value associated with the label 'flood_article' exceeded 0.5, the news article was classified as flood related. Conversely, if the value associated with the label 'not_flood_article' surpassed 0.5, the news article was classified as not flood related. To accomplish this classification, the 'en_core_web_lg' model was augmented and trained with the provided training data. The 'tok2vec' component in the pipeline converted each word of the already tokenised news article in the training data into a word vector representation. All individual word vectors were then aggregated into a single vector representing the whole news article. This vector was then handed over to the 'textcatbow' component of the pipeline. This component uses a Bag of Words approach, where the vector represents the frequency of words in the document. If a word appears multiple times in the news article, its corresponding element in the vector will have a higher value. This information was used by the 'textcatbow' component to finally classify a news article as either flood related or not (HaCohen-Kerner et al., 2020; spaCy, 2024). The validation data in the model training process was used to increase the generalisation capabilities of the trained model by trying to avoid overfitting. Overfitting is describing the phenomena, when a trained model cannot generalise well and fits too closely to

[5] https://spacy.io/models/en

the training dataset instead. Such a model would perform exceptionally well on the training data. However, the performance of the model would decrease, when encountering new data (Ying, 2019).

In spaCy all these pipeline specific instructions and variables are stored within a configuration file. A base configuration file for spaCy's 'textcat_multilabel' pipeline was downloaded from their website[6] as shown in Figure 4.4. This configuration file together with the training and validation data was used train the text classification model.

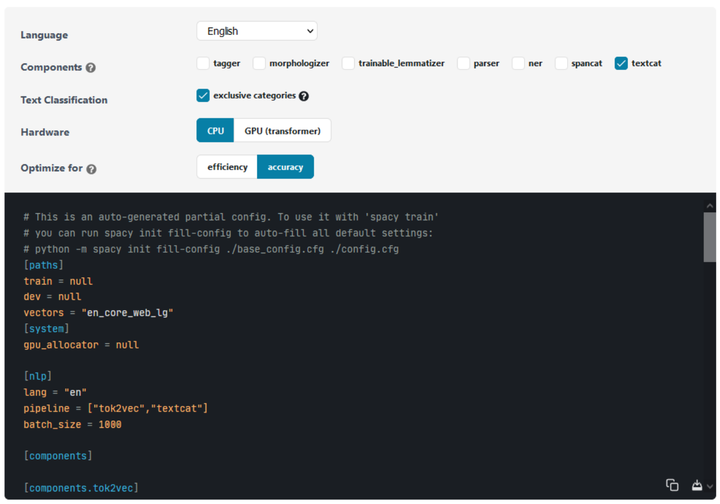


Figure 4.4: spaCy configuration file settings for an accuracy optimised'textcat _multilabel' pipeline.

4.1.3 Text Classification Model Evaluation

The performance of the fully trained text classification model described in Section 4.1.2 was then tested on the test dataset created in Section 4.1.1. The model was applied on the test dataset and a confusion matrix of correctly and incorrectly classified news articles was created (Figure 4.5). In addition to the test data, a confusion matrix was also calculated for the validation data and a subset of manually labelled GKG news articles stored in the SQLite DB. Based on such confusion matrices, the following performance metrics were calculated.

[6] https://spacy.io/usage/training#quickstart

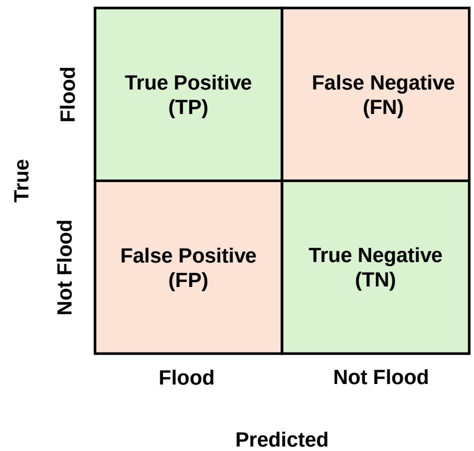


Figure 4.5: Confusion Matrix showing the relationship between predicted and true news article classifications. Green boxes show the news articles correctly predicted either as flood relevant (True Positive, TP) or not flood relevant (True Negative, TN). Red boxes show the incorrectly predicted news articles (False Negative, FN) and (False Positive, FP).

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{4.1}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4.2}$$

$$\text{F1-Score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4.3}$$

Equation 4.1 and equation 4.2 were calculated according to Powers (2020). Precision measures the accuracy of the positive predictions made by the model, specifically the proportion of correctly predicted flood-related news articles out of all news articles, which were labelled as flood relevant. Recall measures the ability of the model to correctly identify all relevant instances, specifically the proportion of correctly predicted flood-related news articles out of all actually flood-relevant news articles. The F1-Score is calculated as the harmonic mean of precision and recall, giving equal weights to both metrics as defined in Frankes and Baeza-Yates (1992).

4.2 Place Name Recognition and Resolution

NER is a NLP technique with the goal to identify and classify named entities in text. Named entities are specific, categorisable elements within a text that hold a significant semantic meaning (Yadav and Bethard, 2019). Similarly to text classification tasks (Section 4.1), a NER model requires training with labelled data. Such labelled data contains text documents with manually annotated words, marking the named entities present in the document. The data is then utilised to train a selected model type. Throughout training, the model improves its rate of correctly predicting new entities resembling those found in the labeled data categories (Nadkarni et al., 2011b). The named entities with the most significant meaning for this thesis are Nigerian place names found within English language news articles. Already existing trained NER models tend to prioritise place names from the regions or countries, from where the data is sourced. For example, an NER model trained on English language news articles contains a significant number of place names from English speaking countries like the United States or the United Kingdom (Akdemir et al., 2018). Therefore, the objective here is to train a NER model to specifically extract Nigerian place names from English language news articles.

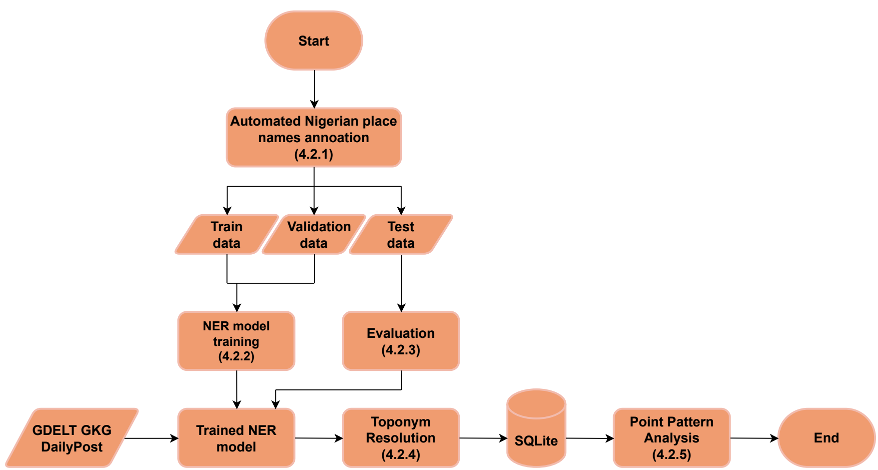


Figure 4.6: Overview of automated annotated data generation, NER model training, model evaluation and point pattern analysis. Numbers in brackets indicate the sections, in which the respective parts of the workflow are discussed.

Figure 4.6 shows the overall workflow described in this section. Section 4.2.1 introduces an approach to automatically annotate Nigerian place names in synthetically generated news articles. Section 4.2.2 elaborates on the structure of the NER model utilised and Section 4.2.3 introduces the framework for evaluating the model's performance. Section 4.2.4 elaborates on the methods used to geocode the extracted place names and Section 4.2.5 introduces the methods applied in order to visualise the extracted place names and analyse the observed patterns on the maps.

4.2.1 Automated Nigerian Place Name Annotation for NER Training Data Generation

Figure 4.7 shows the automated Nigerian place name annotation workflow to generate NER training data in more detail.

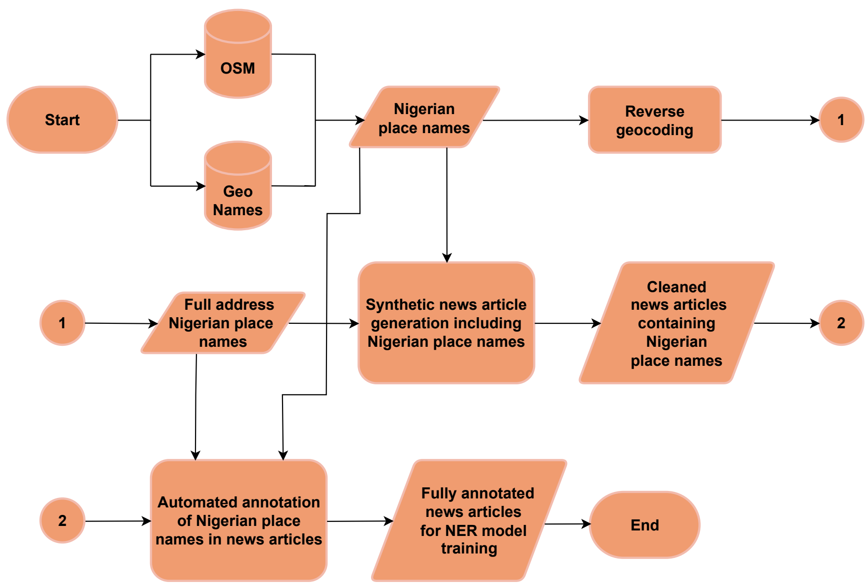


Figure 4.7: Overview of automated annotated data generation. Nigerian place names from OSM and GeoNames were used to be included in synthetically generated news articles. A string matching algorithm automatically annotated the Nigerian place names in the news articles, which were then ready to be used for NER model training.

As a first step available Nigerian place names were downloaded from Open Street Map (OSM) and GeoNames. OSM is a collaborative project creating a free editable map of the world, built by a community of mappers. Data contained within the OSM project can be mined via Overpass Turbo web tool. Overpass Turbo allows to run queries against the OSM DB using the Overpass Application Programming Interface (API). The query as shown in Figure 4.8 was used to download Nigerian place names along with latitude and longitude values from OSM[7]. The result of the query was downloaded as a GeoJSON. GeoNames is

[7]`https://overpass-turbo.eu/s/1K2Z`

a geographical DB (gazetteer) providing extensive geographical data, including place names, geographical features and administrative boundaries. GeoNames provides access to this data through web services and downloadable datasets. All available data for Nigeria was downloaded from GeoNames as a textfile file[8]. Both datasets combined resulted in 185'123 unique Nigerian place names including their corresponding latitude and longitude values. These latitude and longitude values were used to get the full address for each of these 185'123 Nigerian place names using the reverse geocoding functionality of the Nominatim geocoding service within the GeoPy python library (version 2.2.0)[9]. The Nominatim geocoding service is also based on the data within the OSM project. To obey the geocoding rules stated by Nominatim, the 185'123 place names were geocoded in batches of 1000 locations and after each geocode a sleep timer of 10s was introduced. All 185'123 place names underwent reverse geocoding, allowing for the inclusion of both place names composed of individual words and partial addresses, which were later included in the synthetically generated news articles.

```
[out:csv(::id,::type,"name", ::lat, ::lon)];
area[name="Nigeria"];
nwr(area)[name];
out center;
```

Figure 4.8: Query used to download Nigerian place names along with latitude and longitude values from OSM.

To generate the synthetic news articles, OpenAI's GPT-3.5 Turbo model was used. OpenAI[10] is a company known for developing advanced AI models such as the GPT series. The python library openai[11] was used to get access to the commercial GPT-3.5 Turbo API. The API works on a prepay based system and the consumption is tracked using an API key, which was ordered from their website[12]. One could envision the interaction with this API as a dialogue between two individuals. However, the model required priming beforehand. This included instructing the model on the persona it should imitate, thus shaping the responses it generated. The model was primed with the following prompt sentence:

'You are a journalist at a newspaper agency. You mainly write news articles about countries in Africa, for example Nigeria. In your articles you use an objective, informative, neutral and concise language and tone.'

[8] https://download.geonames.org/export/dump/
[9] https://geopy.readthedocs.io/en/stable/
[10] https://openai.com/
[11] https://github.com/openai/openai-python
[12] https://platform.openai.com/

The model replied with the following answer. The whole answer can be found in Appendix A:

'As a journalist at a newspaper agency, I specialize in writing news articles about various countries in Africa, with a focus on providing objective, informative, and neutral content. I strive to present the facts concisely and accurately, ensuring that my articles maintain a fair and balanced tone [...]'

Now the mode was ready to produce the first news article. The model was instructed with the following sentence:

'Write a news article. The topic of the article can be chosen freely. But the article must exactly contain the following 3 Nigerian place names: 'Zungoro, Shiroro, Niger, Nigeria','Lagos', 'Abuja'.'

The model responded with the following news article. Only parts of the whole news article is shown here, highlighting how the given place names were integrated in the synthetic news article. The whole answer can be found in Appendix A:

'Title: Communities in Zungoro, Shiroro, Niger, Nigeria, Rebuild after Dam Tragedy - In the wake of the devastating dam collapse tragedy that struck these communities, residents are banding together to rebuild their lives and restore normalcy. The affected areas, located in the Niger state of Nigeria, have been grappling with the aftermath of the disaster as they embark on a journey of recovery. The incident occurred several weeks ago when heavy rainfall overwhelmed the Shiroro Dam, leading to its breach and the subsequent flooding of nearby communities. The catastrophe left many without their homes, displacing families and inflicting significant damage to infrastructure. One of the hardest-hit areas was Zungoro, a small rural village situated near the banks of the Niger River [...]. Lagos and Abuja, the capital cities of Nigeria, promptly dispatched relief materials, medical supplies, and humanitarian support to assist in the recovery efforts [...]. The Nigerian government has committed to investing in long-term measures to prevent similar tragedies from occurring in the future. Plans for strengthening dam infrastructure and implementing better disaster management systems are underway. This initiative aims to safeguard not only the communities in Shiroro but also the wider region [...].'

All subsequent news articles were automatically created using the statement below. The place names to be used and the number of included place names were randomised. Specifically, 'subset_size' and 'random_subset' represent two variables used within the code, where 'subset_size' instructed the model to include between 3 and 20 place names contained within 'random_subset'. 'random_subset' contains the place names directly downloaded from OSM and GeoNames:

'Thanks sounds good. Now create another news article. Choose a new topic. But the article must contain the following subset_size Nigerian place names: "+ ', '.join(random_subset) +".'

Any kind of formatting within the synthetically produced news articles was removed and the news articles were stored in individual textfiles.

In a next step all Nigerian place names contained within the synthetically created news article were automatically annotated. For this, a custom string matching algorithm was developed. The place name dictionary was split into sub-dictionaries and a batch-based approach was used to boost the efficiency of the automated annotation. As soon as a place name, which was downloaded from OSM and GeoNames, was found within the synthetic news articles, the start and end index of each found place name was calculated and the label 'GEO' was attached to the index range. For a subset of places names within a news article, the place name was extended to the whole address and updated start and end indices for the found place name were calculated. Automatically annotated news articles were stored as .json files. The structure of these .json files is shown in Figure 4.9.

Figure 4.9: Structure of automatically annotated news articles, a JSON object with two elements: a string representing the news article and a dictionary containing information about the entities mentioned in the string. The entities are annotated with their starting and ending index positions within the news article, along with their entity type 'GEO' representing Nigerian place names.

The outlined automated workflow was used to create 1854 annotated news articles containing Nigerian place names. These articles were split into training, validation and test datasets according to Table 4.2. The training and validation datasets were used to train the NER model in Section 4.2.2 and the test dataset was used to evaluate the model as described in Section 4.2.3.

4.2.2 NER Model

The python library spaCy was as well used to train a NER model based on the annotated data created in Section 4.2.1. In a first step the training and validation datasets were tokenised as described in Section 4.1.2. However, instead of spaCy's statistical model 'en_core_web_lg', a transformer model was

Table 4.2: Distribution of annotated data for training, validation and test datasets used for place name recognition.

Data	**Split [%]**	**Split [#]**
TRAIN	70	1298
VALID	15	278
TEST	15	278

used throughout this section. Transformers are a type of neural network architecture. Neural networks are a very effective type of model to analyse complex data such as written texts. Before transformer models were developed, text data was generally analysed by a type of neural network called Recurrent Neural Network (RNN). Because the word order in written language matters, RNNs analyse bodies of texts sequentially. However, RNNs did not perform very well on long pieces of text, as RNNs tend to lose track of the meaning contained within a text, while analysing it sequentially. Additionally, RNNs do not scale well, when trained on large amount of data (Lipton et al., 2015). Transformers on the other hand can be very efficiently trained on enormous amounts of data. Transformers introduced also the concept of positional encoding. Instead of looking at words sequentially, each word in a body of text is assigned a number according to its position within a text, before it is fed into a neural network. Therefore, the neural network is able to learn the importance of word order from the data itself. Also transformers introduced the concept of self-attention, meaning that a neural network can understand a word in the context of the words around it. This for example helps to differentiate between cases, where a country name is used as an actor or as a location, where some event happened (Vaswani et al., 2017). Therefore, a transformer model was chosen to set the base of the NER model trained in this thesis, specifically the RoBERTa transformer model from HuggingFace[13].

Tokens from the training and validation data were processed by the transformer model to generate contextualised word vectors. The concept of word vectors was introduced in Section 4.1.2. These word vectors were then used by the transition based parser to identify Nigerian place names in the training and validation data. The transformer model enriched the NER component by providing deep contextual information about the news articles within both datasets (SpaCy, 2024).

A base configuration file for spaCy's transformer based NER pipeline was downloaded from their website[14] as shown in Figure 4.10. This configuration file to-

[13]https://huggingface.co/FacebookAI/roberta-base
[14]https://spacy.io/usage/training#quickstart

gether with the training and validation data was used train and tune the NER model. The NER model was trained on the Google Colab platform, as training the transformer model relies on GPU resources.

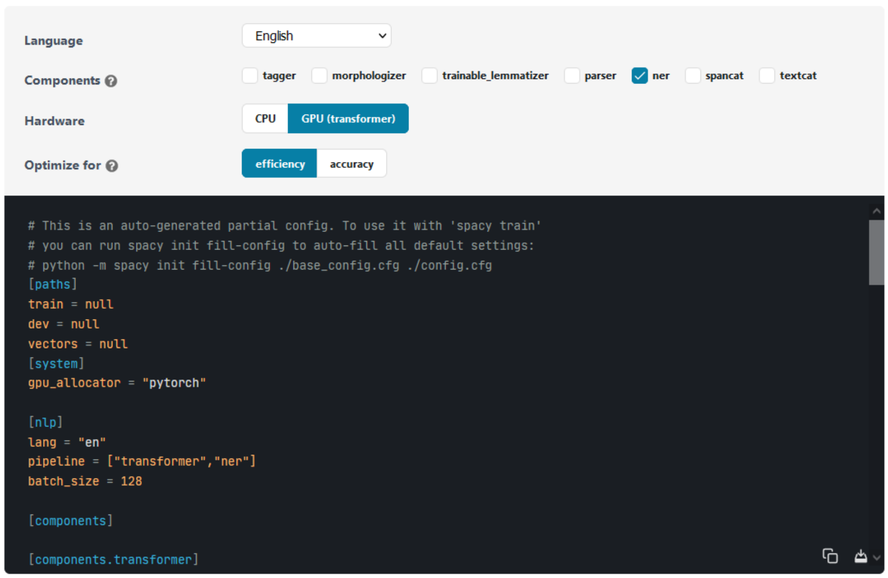


Figure 4.10: spaCy configuration file settings for a transformer based NER pipeline.

4.2.3 NER Model Evaluation

The performance of the fully trained NER model described in Section 4.2.2 was then tested on the test dataset created in Section 4.2.1. The model was applied to the test dataset and a confusion matrix of correctly and incorrectly recognised Nigerian place names was created (Figure 4.11). In addition to the test data, a confusion matrix was also calculated for the validation data and a subset of manually annotated Nigerian place names contained in the GKG news articles stored in the SQLite DB. To calculate confusion matrices for all test datasets, a token level evaluation approach was chosen (Zaratiana et al., 2022). Each token gets either predicted as 'GEO' when representing a place name or 'O' when representing any other kind of token. This approach recognises partially overlapping true and predicted entities as correct. Consequently, there might be cases, where place names are incorrectly classified as correctly predicted. This issue is visualised in an example in Figure 4.12.

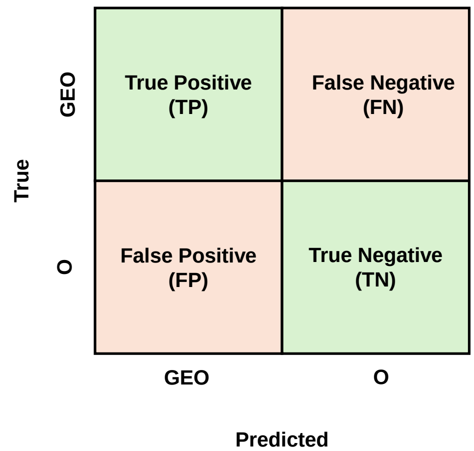
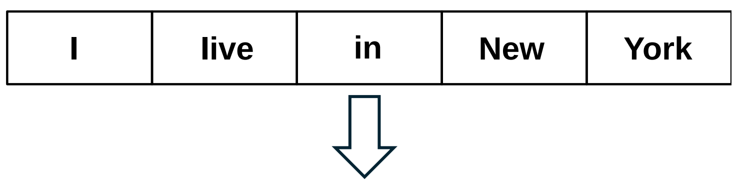


Figure 4.11: Confusion Matrix showing the relationship between predicted and true Nigerian place names. Green boxes show the correctly predicted tokens, whereas red boxes show the incorrectly predicted tokens.



	True	Predicted
I	O	O
live	O	O
in	O	O
New	GEO	O
York	GEO	GEO

Figure 4.12: Visualisation of the potential disadvantages of the token level evaluation method. Here, York gets counted as correctly predicted. However, York on its own is a different place name than New York.

Based on the confusion matrices, the following performance metrics for imbalanced datasets were calculated, since a greater number of tokens in the examined news articles represent not Nigerian place names than actual Nigerian place names. Recall was also used here as defined in Equation 4.2 in Section 4.1.3. Equation 4.4 was calculated according to Swets (1988). Here, recall measures the ability of the model to correctly identify all relevant instances, specifically the proportion of correctly recognised Nigerian place names out of all Nigerian place names actually contained within the test datasets, whereas the specificity focuses on the proportion of correctly recognised tokens not representing Nigerian place names out of all tokens not representing Nigerian place names. The balanced accuracy is an advanced performance metric not influenced by imbalanced datasets (Bekkar et al., 2013).

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{4.4}$$

$$\text{Balanced Accuracy} = \frac{\text{Recall} + \text{Specificity}}{2} \tag{4.5}$$

4.2.4 Toponym Resolution of Extracted Place Names

For all flood-related news articles in the SQLite DB possible Nigerian place names contained with the text were extracted by the model trained in Section 4.2.2 and stored in the SQLITE DB under column 'LOCATION_T'. Nominatim was used for toponym resolution of the place names contained in column 'LOCATION_T'. For place name disambiguation only latitude and longitude values within Nigeria's bounding box (4.2722,2.6711,13.8803,14.6699) were accepted. For ambiguous place names within the country borders, the candidate with the highest importance was chosen. An importance value is assigned to each place name contained within Nominatim's internal gazetteer and is calculated based on Wikipedia articles related to the place name, search and address rank values (Nominatim, 2024). For each geocoded place name in column 'LOCATION_T' its corresponding latitude and longitude value was stored in column 'COORDINATES_T'.

4.2.5 Point Pattern Analysis

All coordinates in 'COORDINATES_T' were manually checked and then plotted on a map as points in space. Point data can form patterns called point patterns. Therefore, point pattern analysis methodologies were applied in order to further analyse the map showing Nigerian place names extracted from flood-related news articles. Specifically, a set of metrics were used, which describe, how points are distributed in space and if any kind of clustering is observable. Quadrat statistics, Ripley's G function and Density Based Spatial Clustering of Applications with Noise (DBSCAN) are described here, as they were used in order to answer RQ2.

Quadrat statistics

Quadrat statistics analyse the spatial distribution of points within an area by counting the number of observations in specified cells (quadrat). The observations in this thesis are the extracted and geocoded place names from flood-related news articles. This method assesses, whether points are evenly dispersed across the cells or if they tend to cluster in certain areas. Specifically, quadrat statistics use a chi-squared χ^2 test (Rey, 2024). In order to calculate the quadrat statistics, the python library PySAL (version 24.01)[15] was used.

Ripley's G

Ripley's G function focuses on nearest neighbour distances observed in point patterns and characterises clustering within those patterns. The nearest neighbour to a point in a point pattern is the point that is closest in distance to that point. The G function evaluates the cumulative proportion of points that have their nearest neighbour within various distance thresholds. It charts these cumulative percentages over increasing distances. This distribution typically has a unique shape, when the points are randomly distributed. The core idea behind Ripley's G is to compare the observed distribution of nearest neighbour distances against a set ofincreasing distance thresholds with a theoretical distribution derived from simulated spatially random patterns, typically using a spatial Poisson point process as the reference. This comparison helps to determine, how closely the observed pattern aligns with a spatially random distribution. The g_test function within the PySAL library was used to calculate the G function (Ripley, 1988).

DBSCAN

DBSCAN is a popular clustering algorithm used to identify distinctive clusters in the data based on their density. A cluster for DBSCAN is a concentration of at least a defined number of points, each of them within a defined radius of at least another point in the cluster. DBSCAN differentiates between core, border and noise points. Core points are located within a cluster with the minimum number of defined points and radius. Border points are also located within a cluster, but only meet the radius requirement. Noise points are located outside of any clusters. The algorithm starts with an arbitrary point. If this point is a core point, it forms a cluster together with all points within the radius, including all border points. DBSCAN recursively analyses all unvisited points and classifies them either as core or noise points (Ester et al., 1996). The DBSCAN algorithm in the python library scikit-learn library was used (version 1.2.2)[16].

4.3 News Article Publication Date

In order to answer RQ1.3, it was necessary to scrape the publication dates for the analysed newspaper articles directly from the respective websites.

To scrape the publication dates for DailyPost news articles, the python library Scrapy was used again. The same workflow as described in Section 3.2 and Section 4.1.1 was used to scrape the publication date. The publication date for

[15] `https://pysal.org/`
[16] `https://scikit-learn.org/stable/`

DailyPost news articles is stored under the `<div>` HTML tag and 'mvp-post-date updated' class. The publication date returned in the yyyy-mm-dd format was stored under the 'DATE_T' column in the SQLite DB. For every news article in the GKG dataset, GDELT provides a date under the column 'V2.1DATES' (Table 3.1), which describes, when the news article was published. This date is given in a YYYYMMDDHHMMSS format. In order to enable a date comparison between the 'V2.1DATES' and 'DATE_T', the date format of column 'V2.1DATES' was aligned to match the date format in column 'DATE_T'.

4.4 Numeric Attribute-Value Pairs

Numeric AV pairs are a way of representing data, where an attribute is associated with a numeric value. The attribute is specific characteristic of an object. In a flood-related news articles, such attributes could be the affected houses or the people who died in a flood disaster. The value refers to the actual numeric value or measurement associated with the attribute, i.e. the number of affected houses and people who died in a flood disaster. Such numeric AV pairs can be used as parameters to estimate the impact of a flood on people and infrastructure without reading through the entire news article.

To automatically extract numeric AV pairs from news articles, it is necessary to develop rules that consider the relationships between words in a sentence and their grammatical properties. Such rules are always language specific. However, processing raw text presents significant challenges. For example, many words are infrequently used within a language and similarly looking words can have completely different meanings or change the meaning according to its position within a sentence. Additionally, segmenting text into its sub units like words can be challenging. For this reason spaCy, more specifically the statistical en_core_web_lg language model, was used. This model preprocessed newspaper articles, providing additional information facilitating the development of rules, which enabled an automated extraction of numeric AV pairs from the articles. The same base model was used to classify news articles based on its content as described in Section 4.1. After tokenising the raw news article, spaCy can parse and tag a given news article using its statistical model. The tagging and parsing done by spaCy for the raw input sentence '10 houses were affected by the flood' is visualised in Figure 4.13. Tagging or POS tagging is a process, where each word in an input text is assigned a tag indicating the words grammatical category. Such grammatical categories include nouns, verbs, words representing number or characters representing punctuation within a text or sentence. The tag assigned to each word is dependent on the position of the word within the larger context (SpaCy, 2024a). spaCy uses the tags from Universal Dependencies[17] (UD). The tags assigned to each word in Figure 4.13 is shown in capitalised letters under each word. The arcs connecting words in Figure 4.13 are the result of spCy's dependency parsing. Dependency parsing focuses on

[17] `https://universaldependencies.org/u/pos/`

establishing the grammatical structure of a sentence by identifying the relationships between words. In a sentence, each word is linked to another word by a directed arc. The word at the beginning of the arc is referred to as the head and the word at the end of the arc is considered as the child of the head. Every word in a sentence is connected to precisely one other word, which acts as its head, except for the root of the sentence, which is typically the main verb and does not have a head. Each of these arcs is labelled with a type of dependency (SpaCy, 2024a). In Figure 4.13 the first arc is labelled as 'nummod', indicating that '10' is the numeric modifier of the word 'houses. 'affected' is the root of the sentence and therefore has no dependency arc directed at it. The dependency types used in spaCy are also taken from UD.

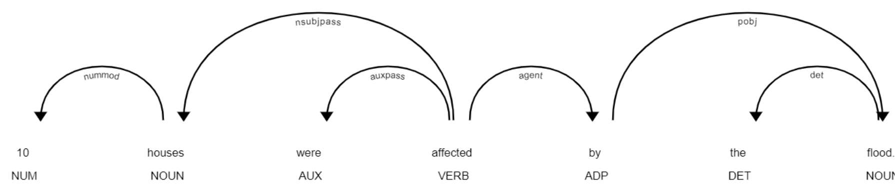


Figure 4.13: Visualisation of spaCy's POS tagging and dependency parsing used to develop rules to automatically extract numeric AV pairs from news articles.

4.4.1 Development of AV Extraction Rules

Based on the additional information as shown in Figure 4.13, rules were developed to extract numeric AV pairs from news articles. The basic set of rules was extracted from the pseudo code presented in Panem et al. (2014) and was extended with additional rules to improve the numeric AV pair extraction from news articles.

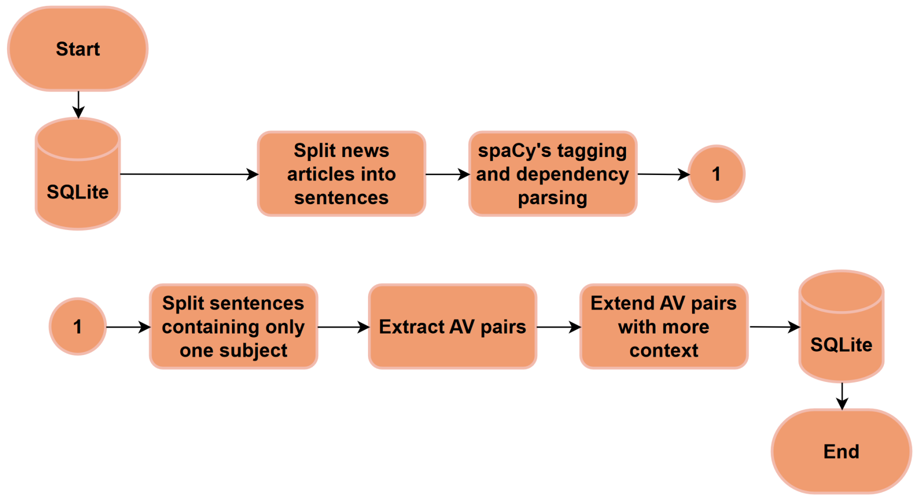


Figure 4.14: Overview of the developed algorithm to extract numeric AV pairs in news articles.

Figure 4.14 visualises a simplified overview of the algorithm developed to extract numeric AV pairs. As a first step each news article was split into single sentences using periods ('.'), exclamation marks ('!') and question marks ('?') found in the news articles. If a period was located between two characters representing numbers, such as in '2.5', no split was performed. As a next step, each sentence had to be processed by spaCy. With the additional information provided by spaCy, each sentence was further split into sub sentences, containing only one active or passive subject. Therefore, spaCy's dependency tags 'nsubj' and 'nsubjpass' were used. All dependency tags used to split sentences and to create rules to extract numeric AV pairs are described in Table 4.3. If one sub sentence still contained more than one active or passive subject, the sentence was further split at punctuations or words commonly indicating parts of a sentence, such as ',', ';', 'and' and 'where'. In total, 23 such punctuation marks or words were used. Now the news article was ready for basic AV pair extraction. To start, each sub sentence was checked for the containment of any 'nummod' dependencies. A 'nummod' dependency is shown in Figure 4.13 between the words '10' and 'houses'. If such 'nummod' dependencies were found, the head (e.g. houses) and the child (e.g. 10) of that dependency were extracted. Additionally, if a word in a sub sentence was labelled with a 'SYM' (symbol POS tag) and represented the child of a 'quantmod' dependency, the head and the child of that 'quantmod' dependency were extracted as well. Using those two rules, the basic numeric AV pairs were extracted from a news article.

Table 4.3: Description of dependency tags used to split sentences and to extract numeric AV pairs.

Dependency Tag	Description
acl	clausal modifier of noun (adnominal clause)
advmod	adverbial modifier
amod	adjectival modifier
appos	appositional modifier
compound	combinations of lexemes
dobj	direct object
nmod	nominal modifier
nsubj	nominal subject
nsubjpass	nominal subject passive
nummod	numeric modifier
pobj	prepositional object
prep	preposition
quantmod	quantifier modifier

In the next step all extracted basic numeric AV pairs were extended to provide additional context. With this added context, the extracted numeric AV pairs should become clearer, even without access to the full newspaper article. The rules in Algorithm 1 were used to extend either the attribute or value part of the extracted pair based on the additional related content contained within the analysed sub sentence.

Iterate thorugh the words of a subsentence:

Algorithm 1 Extension of AV pairs with more context using dependency parsing

Require: $sentence$ ▷ *a sub-sentence extracted containing one subject*
AVs ▷ *the corresponding AV pairs of this sub-sentence*

for $i = 0, \ldots, len(sentence) - 1$ **do** ▷ *iterate through the words*
  $token = sentence\,[\,i\,]$
  for av in AVs **do** ▷ *iterate through the AV pairs*
    $av_{attri} = av\,[\,0\,]$ ▷ *the attribute*
    $av_{val} = av\,[\,1\,]$ ▷ *the value*
    if $token.dep\_ ==$ "$nsubj$" **then**
      ▷ *if the current word has a dependency relationship "nsubj"* ◁
      if $token \subseteq av_{attri}$ OR $av_{attri} \subseteq token$ **then**
        $av_{attri} \leftarrow av_{attri} + token.head$
    if $token.dep\_ ==$ "$compound$" **then**
      if $token \subseteq av_{attri}$ OR $av_{attri} \subseteq token$ **then**
        $av_{attri} \leftarrow av_{attri} + token.head$
      else if $token.head \subseteq av_{attri}$ OR $av_{attri} \subseteq token.head$ **then**
        $av_{attri} \leftarrow token + av_{attri}$
      else if $token.head \subseteq av_{val}$ OR $av_{val} \subseteq token.head$ **then**
        $av_{val} \leftarrow token + av_{val}$
    if $token.dep\_ ==$ "$prep$" **then**
      if $token.head \subseteq av_{attri}$ OR $av_{attri} \subseteq token.head$ **then**
        $av_{attri} \leftarrow av_{attri} + token$
        for $j = 0, \ldots, len(sentence) - 1$ **do**
          ▷ *iterate through the words of the sentence again to look for the object of the preposition* ◁
          $token_{sub} = sentence\,[\,j\,]$
          if $token_{sub}.dep\_ ==$ "$pobj$" AND $abs(i - j) <= 7$ **then**
            if $token_{sub}.head \subseteq token$ OR $token \subseteq token_{sub}.head$ **then**
              $av_{attri} \leftarrow av_{attri} + token_{sub}$
    if $token.dep\_ ==$ "$pobj$" **then**
      if $token \subseteq av_{attri}$ OR $av_{attri} \subseteq token$ **then**
        if $token.head.dep\_ ==$ "$prep$" **then**
          compound_dependents = [ ]
          for $child$ in $token.head.head.children$ **do**
            if $child.dep\_ ==$ "$compound$" **then**
              compound_dependents.append($child$)
          if $len(compound\_dependents) > 0$ **then**
            $av_{attri} \leftarrow join(compound\_dependents) + token.head.head + token.head + av_{attri}$
          else
            $av_{attri} \leftarrow token.head.head + token.head + av_{attri}$

```
        else
            av_attri ← token.head + av_attri
if token.dep_ == "quantmod" then
    if token.head ⊆ av_val OR av_val ⊆ token.head then
        av_val ← token + av_val
if token.dep_ == "dobj" then
    if token ⊆ av_attri OR av_attri ⊆ token then
        if token.head.pos_ == "VERB" then
            get tokenHeadPP, past participate of token.head
        av_attri ← av_attri + tokenHeadPP
if token.dep_ == "advmod" then
    if token.head ⊆ av_val OR av_val ⊆ token.head then
        advmod_dependents = []
        for child in token.children do
            if child.dep_ == "advmod" then
                advmod_dependents.append(child)
        if len(advmod_dependents) > 0 then
            av_val ← av_val + join(advmod_dependents) + token
        else
            av_val ← token + av_val
if token.dep_ == "amod" then
    if token.head ⊆ av_val OR av_val ⊆ token.head then
        av_val ← token + av_val
    else if token.head ⊆ av_attri OR av_attri ⊆ token.head then
        av_attri ← token + av_attri
if token.dep_ == "acl" then
    if token.head ⊆ av_attri OR av_attri ⊆ token.head then
        av_attri ← av_attri + token
if token.dep_ == "appos" then
    if token.head ⊆ av_attri OR av_attri ⊆ token.head then
        av_attri ← av_attri + token
if token.dep_ == "nmod" then
    if token.head ⊆ av_val OR av_val ⊆ token.head then
        av_val ← token + av_val
if token.dep_ == "nsubjpass" then
    if token ⊆ av_attri OR av_attri ⊆ token then
        av_attri ← av_attri + token.head
```

4.4.2 Testing and Comparison of Extracted AV Pairs

The rules developed in Section 4.4.1 to extract AV pairs from news articles were tested on two datasets. An automated test was conducted on synthetic news articles generated by OpenAI's commercial GPT-4 model. The GPT-4 model was chosen, as it is a more advanced model and it was asked to come up with numeric AV pairs and incorporate them into the synthetic news articles. To generate these articles, a similar approach as described in Section 4.2.1 was chosen. The model was primed using the following instruction:

'You are a journalist at a news paper agency. You mainly write news articles about countries in Africa, for example Nigeria. In your articles you use an objective, informative, neutral and concise language and tone.'

All news articles were generated in a loop. Per loop two types of output were generated. With the instruction:

'Write another news article. The topic of the article you can choose. But the article must contain numeric AV pairs describing the number of affected people, flood height etc. You must come up with numeric values yourself.'

a news article was generated per loop. After the model returned a news article containing numeric AV pairs the model was tasked with another instruction in the same loop:

'Which numeric AV pairs did you use in the article before?'.

50 synthetic news articles containing numeric AV pairs and 50 summaries created by GPT-4 only containing the numeric AV pairs of the corresponding news article were created to be tested on the rules defined in Section 4.4.1. In order to analyse the content similarity between an AV pair extracted by GPT-4 and the same AV pair extracted by the defined rules, the cosine similarity was calculated for each AV pair comparison. Cosine similarity is based on the cosine of the angle between two vectors. The formula for cosine similarity between two vectors $\mathbf{A}$ and $\mathbf{B}$ is (Karabiber, 2024):

$$\text{Cosine Similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} \tag{4.6}$$

$\mathbf{A}$ and $\mathbf{B}$ represent here the vector representation of the AV pair extracted by GPT-4 and the AV pair extracted with the rules. In order to represent AV pairs in a vector, the 'bert-base-uncased' transformer model was used. The model was accessed using the python transformers library (version 4.36.2)[18]. $\mathbf{A} \cdot \mathbf{B}$ is the dot product of the vectors $\mathbf{A}$ and $\mathbf{B}$. $\|\mathbf{A}\|$ and $\|\mathbf{B}\|$ are the Euclidean norms of the vectors $\mathbf{A}$ and $\mathbf{B}$. The cosine similarity returns a value between -1 and 1 and was calculated using the SciPy python library (version 1.8.0)[19]. Whereas 1

[18] `https://pypi.org/project/transformers/`
[19] `https://scipy.org/`

would represent two identical AV pairs, -1 would represent two totally opposite AV pairs.

Since varying numbers of AV pairs can be extracted by GPT-4 and the defined rules for each news article, it was essential to ensure that the appropriate AV pairs were compared with each other. Therefore, the compound metric was used (Figure 4.15). The compound metric calculates the sum of two values: Comparison of numbers in AV pair and cosine similarity for compared AV. If all numbers in the compared AV match, 100% is returned. 100 is summed up with the cosine similarity returned for the compared AV pair, resulting in the compound metric. An addition was used to weigh the matching numbers higher than the cosine similarity. If the compound metric returned a value larger than 30, it was assumed that two compared AV pairs were correct. If the compound metric returned a value of 30 or below, it was assumed that two incorrect AV were compared with each other and the comparison was removed from the analysis (Figure 4.15).

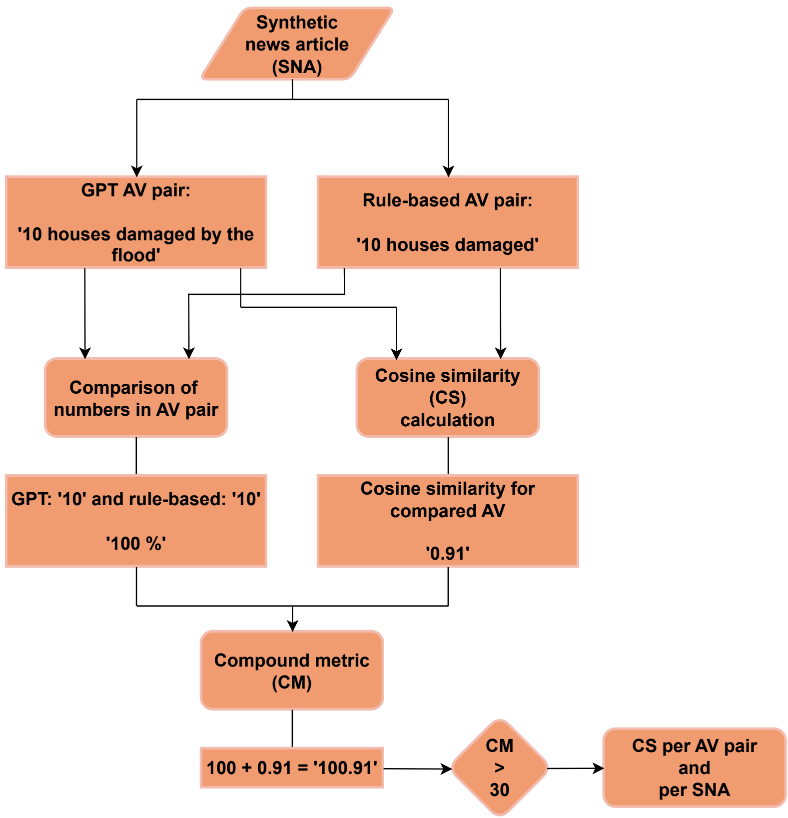


Figure 4.15: Cosine similarity comparison setup between AV pairs provided by GPT-4 and extracted by rule-based method from the same synthetic news articles. Here shown with an example (rectangular boxes), where compound metric is larger than 30. Therefore, assumption was made that correct AV pairs within an article were compared with each other.

Additionally, the extracted AV pairs, using the here proposed method, were manually analysed for 50 out of the 211 flood-relevant DailyPost news articles in the SQLite DB. The extracted numeric AV pairs were compared to the AV pairs extracted by GDELT contained in column 'V2.1AMOUNTS'.

Chapter 5

Results

In this chapter, an overview over the results is provided. Section 5.1 showcases the findings derived through the methodologies outlined in Section 4.1, aimed at addressing RQ1.1. To address RQ 1.2, Section 5.2 illustrates the results acquired through the application of the methods outlined in Section 4.2. Section 5.3 and Section 5.4 present the results to address RQ1.3 and RQ1.4. For this purpose, the methods from Section 4.3 and Section 4.4 were utilised.

5.1 Text Classification

Section 5.1.1 delves into the training process of the trained text classification model, followed by an analysis of the model's performance on validation and test data. Section 5.1.2 analyses the applicability of the trained text classification model on news articles in the GDELT GKG dataset.

5.1.1 Text Classification Model Training Process and Performance Analysis

Figure 5.1 shows the training progress of the text classification model. Throughout the training process, the training data underwent a total of 3200 iterations, as shown on the x-axis of the plot. Epochs refer to one complete pass though of the entire training dataset (Wesslen, 2022). During each epoch, the model updated its parameters based on the gradients of the loss function with respect to those parameters. The loss value quantifies the difference between the predicted output and the actual target, in this case the difference between the labelled and predicted news articles (Wang et al., 2022; Wesslen, 2022). spaCy uses a logistic loss function for text classification tasks (Honnibal, 2018). Based on the validation data, the Adaptive Moment Estimation (ADAM) optimisation algorithm aims to update the model parameters to minimise the loss of the model. ADAM uses two moving averages to adjust the parameters during training, the first moment of the loss functions gradient and the second moment of the same gradient. The first moment refers to the mean value of the gradient, whereas the the second moment refers to the uncentred variance (Kingma and Ba, 2017; SpaCy, 2024b). Within the first 200 epochs the loss value (y-axis) decreased rapidly, indicating that the model improved its ability to make accurate predictions. Between epoch 500 and 2000 the model entered a state of oscillation with a roughly 500 epochs long interval (Figure 5.1). Model oscillation refers to the behaviour, where the loss value increases and decreases within the epochs (Ruiz-Garcia et al., 2021). The model converged slowly after the first 2000 epochs. Model convergence describes the state of the model, where further training on the dataset does not significantly improve its performance by minimising the loss value (Charles and Papailiopoulos, 2018).

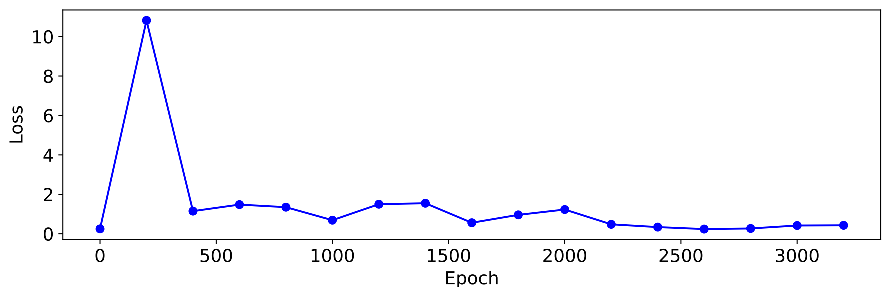


Figure 5.1: Training progress of the text classification pipeline classifying news articles into flood and not flood-relevant news articles. The blue line indicates the loss value over epoch iteration steps.

The model with a loss value of 0.43 was chosen to be tested on the validation and test datasets according to the performance metrics as described in Section 4.1.3. Figure 5.2 shows that of the 1'530 labelled news articles in the validation dataset, 1'526 news articles were either correctly classified as flood relevant or not flood relevant. Only four not flood-relevant news articles were wrongly classified as flood relevant by the trained text classification model. No flood-relevant news articles were classified as not flood relevant. The confusion matrix depicting the predicted classification of the labelled news articles in the test dataset (Figure 5.3), exhibits a comparable pattern to the confusion matrix observed on the validation data. 1'530 labelled news articles were correctly classified by the model. In each instance, one news article was misclassified.

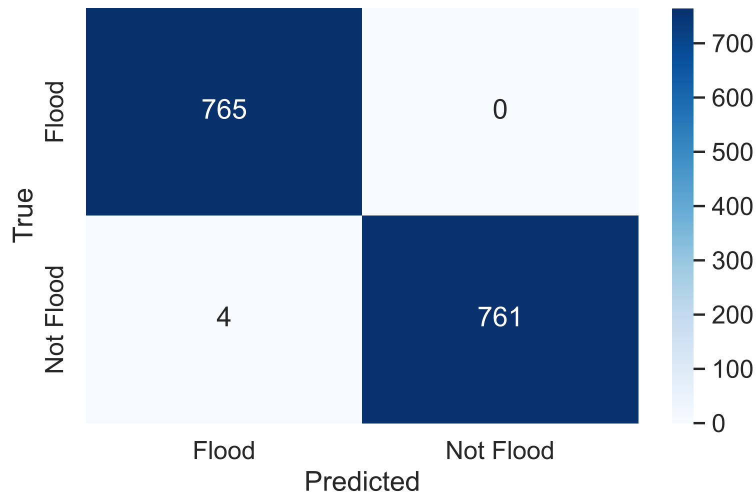


Figure 5.2: Confusion matrix of the trained text classification model tested on the validation dataset containing 1530 labelled news articles.

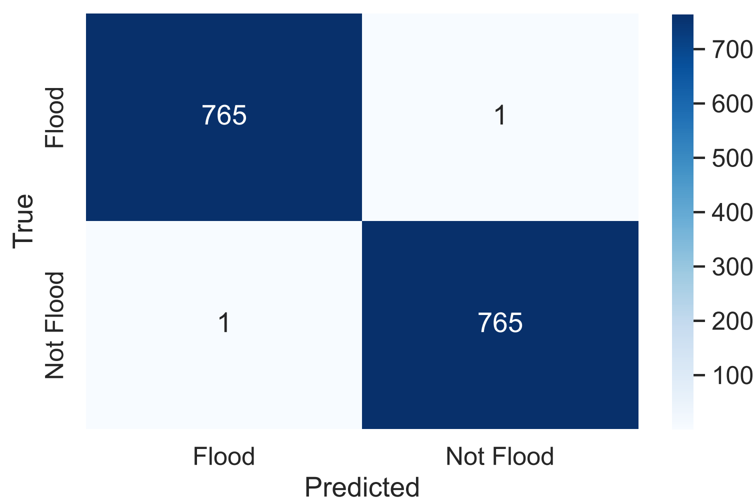


Figure 5.3: Confusion matrix of the trained text classification model tested on the test dataset containing 1532 labelled news articles.

Based on the values shown in Figure 5.2 and Figure 5.3, the performance metrics in Table 5.1 were calculated. The trained text classification model reached exceptionally high performance scores on both the validation and test dataset.

Table 5.1: Performance metrics of trained text classification model tested on the validation and test dataset.

Dataset	Precision	Recall	F1-Score
Validation Data Model	0.995	1	0.997
Test Data Model	0.999	0.999	0.999

5.1.2 Applicability of the Text Classification Model to GDELT GKG News Articles

The trained text classification model was additionally tested on news articles contained in the GDELT GKG dataset. 632 news articles were randomly selected and hand labelled as flood relevant or not. Figure 5.4 shows the 632 predictions of the text classification model in a confusion matrix. Out of the 632 news articles, 525 GKG news articles were classified correctly as either flood relevant or not. 100 actually flood-relevant news articles were wrongly classified as not flood relevant by the model. Additionally, seven not flood-relevant news articles were erroneously classified as flood relevant. Based on these values the performance metrics of the model on the GDELT GKG data were calculated (Table 5.2). For the GDELT test dataset, while the precision remained relatively high at 0.967, there was a notable drop in recall to 0.684, resulting in a corresponding decrease in the F1-Score to 0.801. Despite still achieving a good precision score, the model's ability to capture all flood-related news articles within the GDELT GKG dataset was somewhat compromised, as evidenced by the lower recall value.

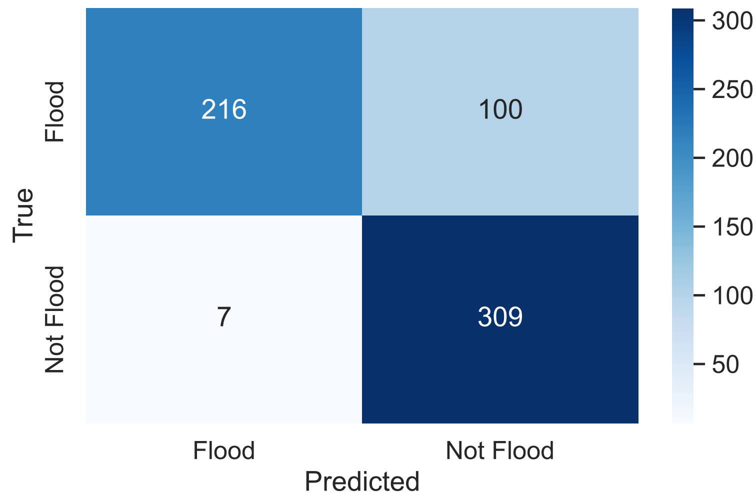


Figure 5.4: Confusion matrix of the trained text classification model tested on 632 news articles in GDELT GKG.

Table 5.2: Performance metrics of trained text classification model on GDELT GKG test data.

Dataset	**Precision**	**Recall**	**F1-Score**
Test Data GDELT	0.967	0.684	0.801

All news articles predicted as TP and FN in Figure 5.4 were manually analysed for its content and categorised into the following thematic groups: flood impact, emergency measures, politicisation, flood prevention and hydro-meteorological characteristics. Figure 5.5 shows a pie chart (A) illustrating the thematic group percentages among the 216 news articles categorised as TP and a pie chart (B) for all 100 news articles categorised as FN. Most TP news articles covered the impact of floods on both people and infrastructure, followed by reports about carried out emergency measures after the onset of flood disasters. 20.8% of the news articles focused on flood disasters within a political context. These included topics such as apportioning blame, visits by political figures to flood affected areas or how the currently reported disaster was being used as an example to change the landscape of the already existing flood protection infrastructure. Only a small portion of the news articles reported on measures that were actually planned to enhance the flood resilience of people and infrastructure in the affected area or included detailed quantitative information on hydro-meteorological parameters such as precipitation levels. However, almost 60% of all FN news articles reported on flood-related topics with a political flavour.

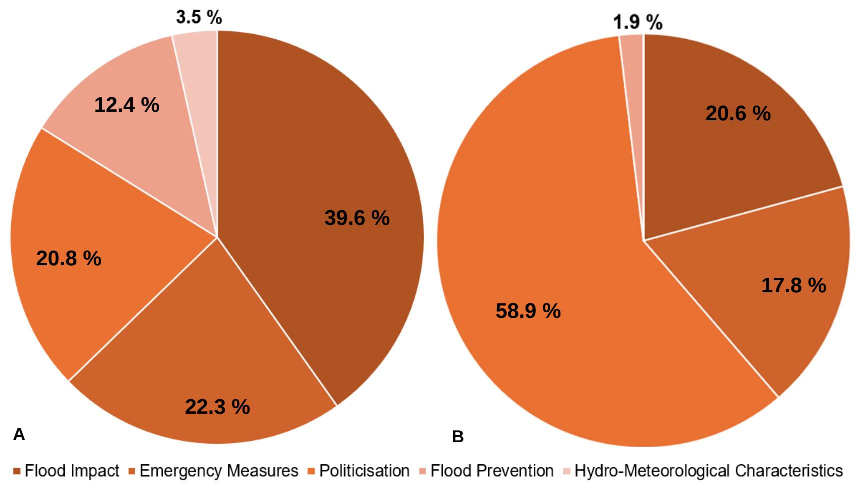


Figure 5.5: Percentage of flood-relevant topics covered in the news articles from the GDELT test dataset. A: TP news articles. B: FN news articles.

The GDELT GKG data includes a thematic classification of news articles in the column labelled as 'V1THEMES'. For the analysed news articles in the GDELT GKG dataset, GDELT on average assigned 27.1 different themes per news article. To extract flood-related news articles, the column 'V1THEMES' was filtered for one of the following terms: natural disaster flood(s)/flooded/flooding/flash flood(s), natural disaster flood water(s) and natural disaster flooded areas. A subset of 327 articles from the filtered dataset was manually analysed based on their content. Out of the 327 articles, GDELT correctly labelled 245. However, there were 81 misclassifications: some articles were incorrectly labelled as flood relevant, when they were not, while others were labelled as not flood relevant, despite being relevant.

5.2 Place Name Recognition and Resolution

Section 5.2.1 delves into the training process of the trained NER model, followed by an analysis of the model's performance on validation and test data. Section 5.2.2 analyses the applicability of the trained NER model on Nigerian place names contained within news articles in the GDELT GKG dataset.

5.2.1 NER Model Training Process and Performance Analysis

The first trained NER model with standard settings of all parameters in the configuration file[1] resulted in a loss value of 585 for the transformer component and a loss value of 1230 for the NER component within the overall pipeline. As defined in Section 5.1.1, the loss value quantifies the differences between the predicted output and the actual target, in this case the difference between the labelled and predicted Nigerian place names contained within news articles. Similarly to the text classification pipeline, a logistic loss function and the ADAM optimiser were used (Section 5.1.1). Loss values are pipeline specific. However, compared to the loss value in the text classification pipeline, as shown in Figure 5.1, the loss values of the NER pipeline with standard configuration settings were much much higher and further away from zero. To further reduce the loss values of the NER pipeline, the concept of hyperparameter tuning was applied (Agrawal, 2021). Hyperparameter tuning is the process of finding the optimal set of hyperparameters for a specific model. In contrast to the parameters, which the model adjusts during training by evaluating its performance on training and validation data, hyperparameters are predetermined before training and cannot be directly learned from the data. An example for such a hyperparameter is the number of hidden layers in a neural network. The goal of hyperparameter tuning is to improve the performance of the model by finding the best combination of hyperparameters (Agrawal, 2021). A random search approach was used to find the optimal set of hyperparameter values (Bergstra and Bengio, 2012). This approach randomly samples hyperparameter combinations from a predefined search space and is less computationally intensive as other search methods.

Figure 5.6A shows, how the loss value steadily decreased over each epoch after the hyperparameter value for maxout pieces was increased from two to three. Maxout refers to a type of layer within the neural network, in which the input is divided into groups and each group is passed through a set of linear transformations, from which the highest value is chosen to create the output of the layer. By increasing the value from two to three, the number of linear transformations within each group was adapted. This can improve the models capability of recognising more complex patterns. However, a balance must be found between model performance and overfitting (Goodfellow et al., 2013). With increasing the maxout pieces value from two to three, the loss value for the transformer component was reduced from 585 to 425 and the loss value for the NER component was reduced from 1230 to 1029. Next, various values for the hyperparameter hidden layers were experimented with. The process of finding the most suitable number of hidden layers is shown in Figure 5.7. Hidden layers are layers of neurons in a neural network. Information flows from the input layer through multiple hidden layers before reaching the final layer. In the training process, neurons in a layer assign weights to the information they receive to

[1] `https://spacy.io/usage/training#quickstart`

learn and recognise specific patterns. All neurons in the various layers are connected with each other (Agyepong and Kothari, 1997). With the reduction of the number of hidden layers, a model becomes less complex and can be beneficial, when the task does not require a highly complex model or the amount of training data is limited. However, if the number of hidden layers is reduced too much, the model may have difficulties in recognising important patterns in the data (Marrone et al., 2021; Touretzky and Pomerleau, 1989). With the increase of the maxout pieces value from two to three and the reduction of the number of hidden layers from 64 to 40, the loss for the transformer component was further reduced from 425 to 243 and the loss for the NER component was reduced from 1029 to 514 (Figure 5.6B).

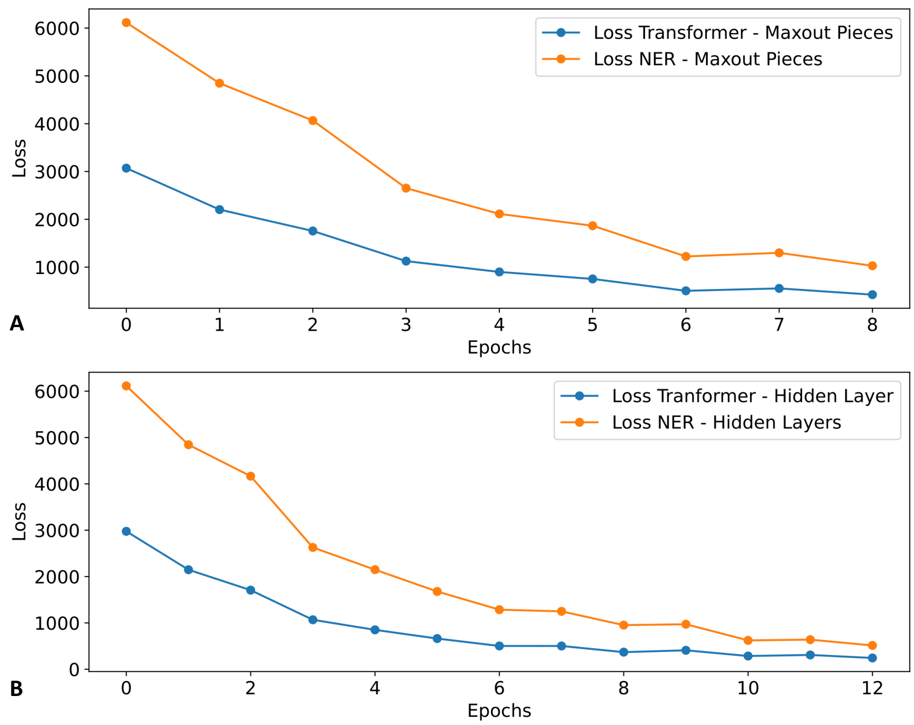


Figure 5.6: A: Loss values of both NER pipeline components after tuning the hyperparameter Maxout Pieces from two to three. B: Loss values of both NER pipeline components after tuning the hyperparameter Hidden Layers from 64 to 40 with a set Maxout Pieces value of three.

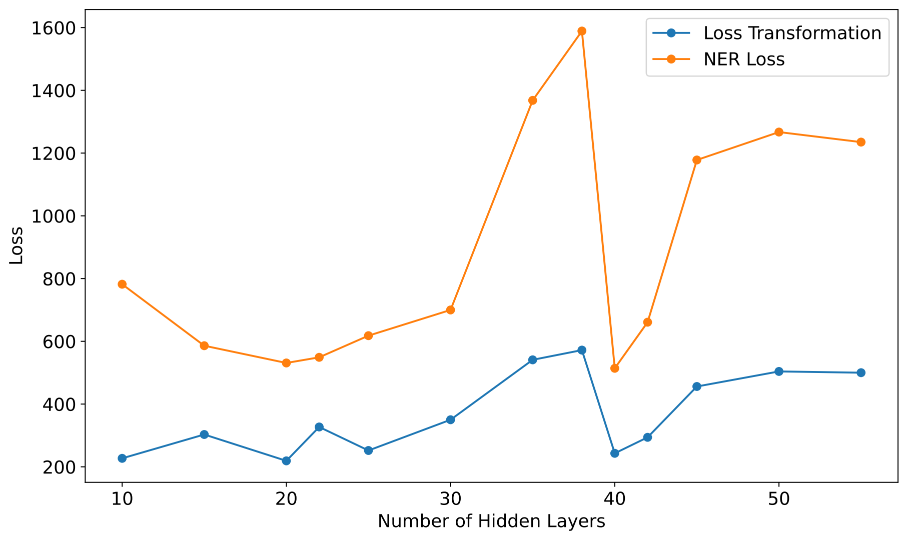


Figure 5.7: Impact of tuning the hidden layer hyperparameter value on the loss value of the NER pipeline components. The loss changes drastically between the hidden layer values of 35 and 45.

This tuned NER model was then chosen to be tested on the validation and test datasets according to the performance metrics as described in Section 4.2.3. Figure 5.8 shows that 158'952 tokens in the validation dataset were either correctly recognised as a Nigerian place name or not as a Nigerian place name. 886 tokens actually representing Nigerian place names were not correctly recognised as a Nigerian place names by the model. On the other hand, 1'554 tokens not representing Nigerian place names were erroneously recognised as Nigerian place names by the model. The confusion matrix for the test dataset shows a similar pattern (Figure 5.9). In the test dataset, less tokens were classified as FP. However, more tokens were classified as FN.

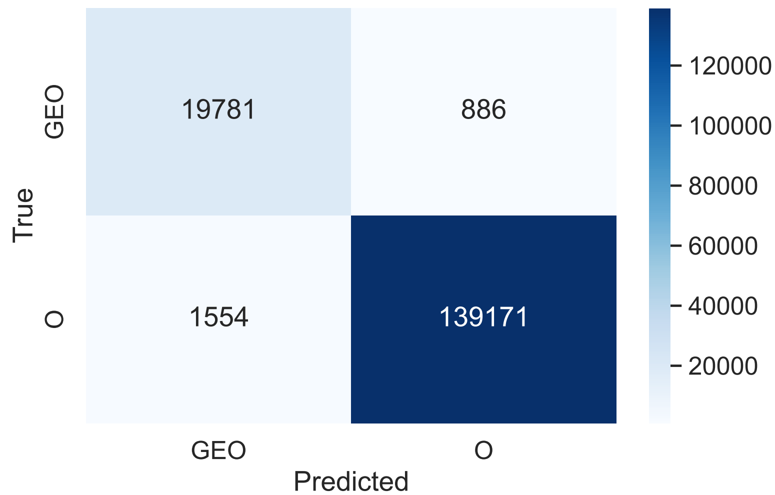


Figure 5.8: Confusion matrix of the trained NER model tested on the validation dataset containing 278 news articles. The numbers represent the classified tokens of all articles within the validation dataset.

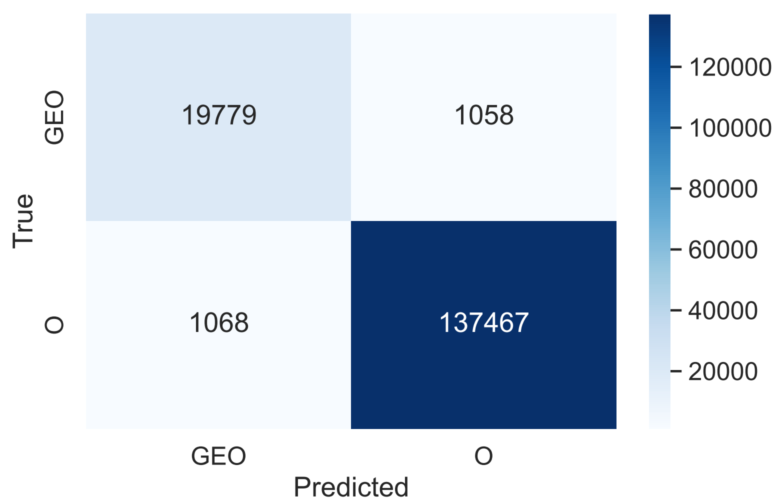


Figure 5.9: Confusion matrix of the trained NER model tested on the test dataset containing 278 news articles. The numbers represent the classified tokens of all articles within the test dataset.

Based on the values shown in Figure 5.8 and Figure 5.9, the performance metrics for imbalanced data in Table 5.3 were calculated. The trained NER model reached exceptionally high performance scores on both the validation and test dataset.

Table 5.3: Performance metrics of the trained NER model on the validation and test dataset.

Dataset	Recall	Specificity	Balanced Accuracy
Validation Data Model	0.957	0.989	0.973
Test Data Model	0.949	0.992	0.970

5.2.2 Applicability of the NER Model to GDELT GKG News Articles

The trained NER model was additionally tested on 100 flood-relevant news articles in the GDELT GKG dataset, which were randomly selected. All Nigerian place names were manually annotated using the open-source online annotation tool NER Annotator[2].

Toponym Recognition

Figure 5.10 shows the confusion matrix of the trained NER model tested on the manually annotated dataset. Based on the values in Figure 5.10, the performance metrics for imbalanced data in Table 5.4 were calculated. Compared to the values in Table 5.3, the trained NER model once more achieved a notable specificity score, affirming its precise prediction of tokens that do not signify Nigerian place names. However, the recall value dropped to 0.725. In comparison, the NER model achieved a high recall value of 0.957 and 0.949 on the validation and test dataset. This indicates that the NER model faces challenges in accurately predicting all tokens that represent Nigerian place names within the GDELT GKG data.

[2] `https://tecoholic.github.io/ner-annotator/`

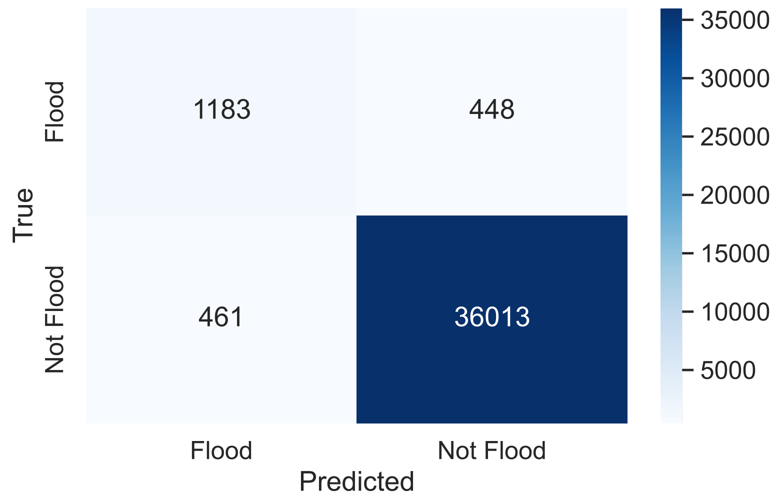


Figure 5.10: Confusion matrix of the trained NER model tested on the manually annotated test dataset containing 100 flood-relevant news articles contained within the GDELT GKG dataset. The numbers represent the classified tokens of all articles within the test dataset.

Table 5.4: Performance metrics of trained NER model on GDELT GKG test data.

Dataset	Recall	Specificity	Balanced Accuracy
Test Data GDELT	0.725	0.987	0.856

Further analysis was conducted on the place names extracted by both GDELT and the trained NER model for all the 211 flood-related DailyPost news articles analysed in the GKG dataset. The Nominatim geocoder returns a search rank value for each geocoded place name. This value is used by Nominatim's toponym resolution algorithm. Table 5.5 illustrates the correlation between search rank values and common place types corresponding to each rank. Furthermore, it includes information regarding their affiliations with administrative units in Nigeria.

Table 5.5: Classification of place types by the Nominatim geocoder and their associations with administrative units in Nigeria.

Search Rank	**Typical Place Types**	**Nigerian Admin Levels**
4	Countries	Country - Nigeria
5 - 9	States, Regions and Provinces	36 Nigerian states and one Federal Capital Territory (FCT)
10 - 12	Counties	774 LGAs
13 - 16	Cities, Municipalities and Islands	Wards
17 - 18	Towns and Boroughs	-
19	Villages and Suburbs	-
20	Hamlets, Farms and Neighbourhoods	-
21 - 25	Isolated dwellings and City blocks	-

Using this information provided by Nominatim, the sum of all place names identified by both GDELT and the trained NER model was categorised according to the search rank classification outlined in Table 5.5. Figure 5.11 distinctly illustrates that GDELT outperformed the trained NER model in identifying the country name Nigeria, along with its 36 states and the FCT. However, the trained NER model demonstrated more advanced extraction of Nigerian place names from lower level administrative units such as LGAs and wards. GDELT identified 998 Nigerian place names from the 211 examined news articles, whereas the trained NER model detected 2'338 Nigerian place names from the same set of 211 news articles. Nominatim returned a search rank for 982 out of the 998 Nigerian place names extracted by GDELT. However, out of the 2'338 Nigerian place names extracted by the trained NER model, Nominatim returned a search rank for only 1'595 of them. The type of entity for the other 743 candidates was analysed manually and categorised into one of the following groups: place names, governmental organisations, names, private companies and others. 59% of the 743 names represent place names that are either unknown to Nominatim or recognised by Nominatim under an alternative spelling. 16% of the candidates represent governmental organisation such as Cross River State Primary Healthcare Development Agency or Kwara State Fire Service, followed by 12% representing first and family names of individuals. A smaller proportion of the candidates represent names of private businesses or organisations or could not be clearly assigned to a group (Figure 5.12).

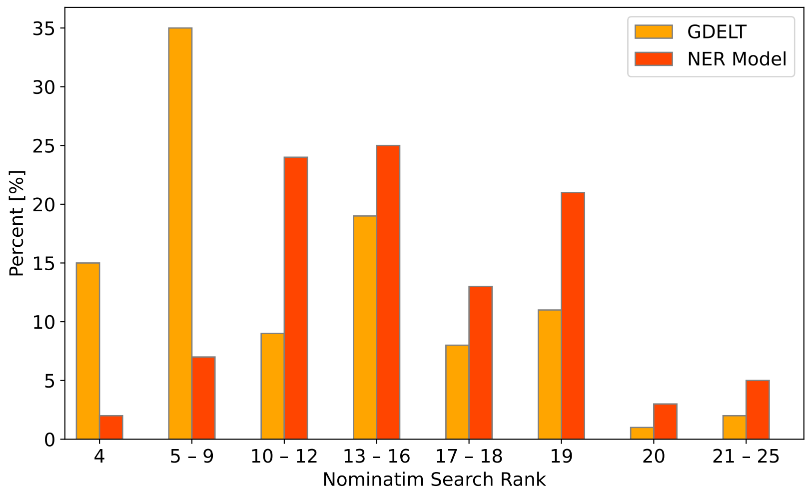


Figure 5.11: Nigerian place names extracted by GDELT and the trained NER model classified according to Nominatim's search ranks.

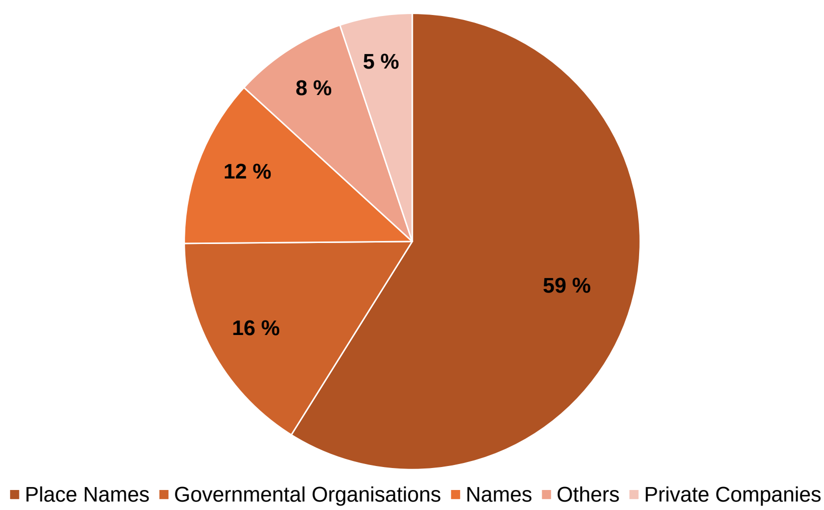


Figure 5.12: Percentage of manually analysed candidates, for which Nominatim could not provide a search rank.

Toponym Resolution

To assess the accuracy of the selected geocoding method in pinpointing Nigerian place names geographically, a comparison was conducted against manually geocoded Nigerian place names. The whole context of a news article was used to disambiguate all place names within that article. After a place name was precisely located in space, the corresponding coordinates from Google Maps[3] were saved in a table in order to calculate the spatial error between ground truth locations and locations geocoded by Nominatim. If Nominatim provided a coordinate pair indicating the centroid of a geographic area, that same location was considered as the ground truth, if no additional information could be extracted from the news article to pinpoint the place with greater detail. Place names in 221 news articles were verified through this manual process. A similar approach was used by Füglister and Purves (2020) to analyse the spatial error between machine coded and hand coded battle locations. Figure 5.13 shows both, locations geocoded by Nominatim and locations manually located in space, using all the additional context provided in a news article. The black lines indicate the spatial error for Nigerian place names, which were incorrectly geocoded by Nominatim. Some locations were situated outside the Nigerian state borders due to the utilisation of a bounding box for some of the spatial disambiguation process.

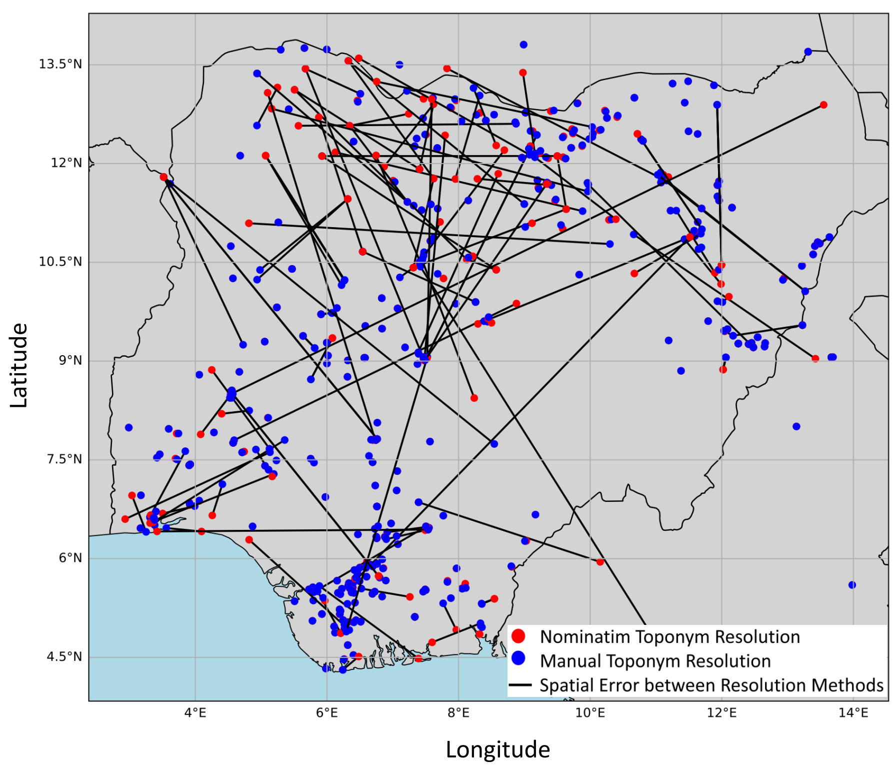


Figure 5.13: Map indicating the spatial error for Nigerian place names geocoded by Nominatim and a manual disambiguation process.

[3]https://www.google.ch/maps/

To draw conclusions about Nominatim's geocoding accuracy on the whole data, the distribution of the spatial error for manually tested locations is shown in Figure 5.14. Approximately half of the Nigerian place names in the test dataset did not exhibit any variation in geographic location between Nominatim and the manual geocoding process. This scenario frequently occurred, when the place name referred to a geographic region and the news article did not provide any more details to specify a more precise location within that area. Close to 80% of the locations in the test data exhibit a deviation of no more than 5.52 km and fall within the 75th percentile. Minor geographic discrepancies primarily occurred, when additional context such as street, bridge or specific building names was provided, allowing for a more precise reference of the place name. Significant spatial deviations occurred when Nominatim incorrectly located place names with a geo/geo ambiguity. These places could be accurately located in the manual process by utilising other place names and higher level geographical units mentioned in the text.

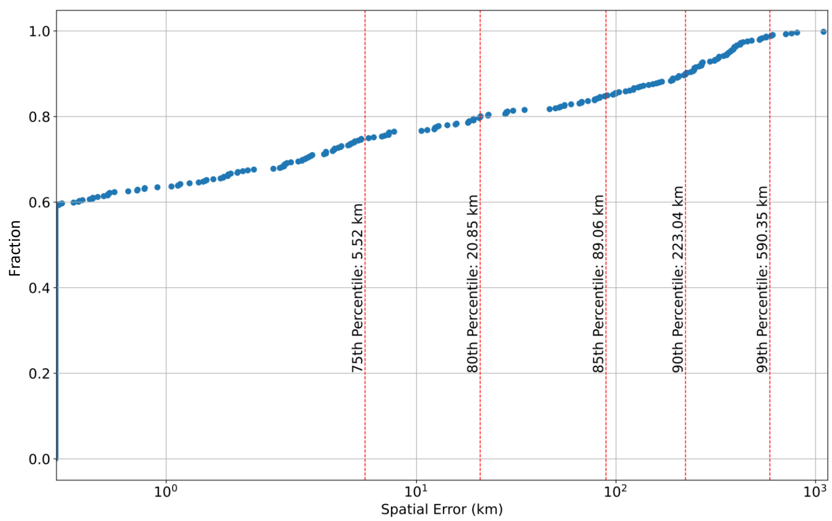


Figure 5.14: Distribution of spatial error for Nigerian place names geocoded by Nominatim and a manual disambiguation process.

Point Pattern Analysis

Figure 5.15 shows all manually adjusted geocoded locations in a two-dimensional hexagonal grid histogram with 30 hexagons in each dimension. Each hexagon visualises the number of locations contained within itself. Darker blue hexagons hold more locations than lighter blue and white hexagons. One can already see that all place names mentioned in flood-related news articles are visually not uniformly distributed in space. Quite a few darker blue hexagons are located in the southern states of Nigeria such as Anambra, Bayelsa, Delta, Imo and Rivers. All these states are located alongside the river Niger. A few individual darker blue hexagons are located in the FCT and the states of Kaduna, Kogi and Kwara, approximately at the location of the capital city of the respective state. Another patch of darker blue hexagons can be seen in the very north of the country in the states of Bauchi, Jigawa, Kano and Yobe. The river Komadugu Yobe flows through all of the states. Two darker blue hexagons are located around Yola, alongside the river Benue in Adamawa state. However, locations in other states alongside the river Benue, such as Benue, Nasarawa, Plateau and Taraba seem to be less frequently mentioned in flood related news articles.

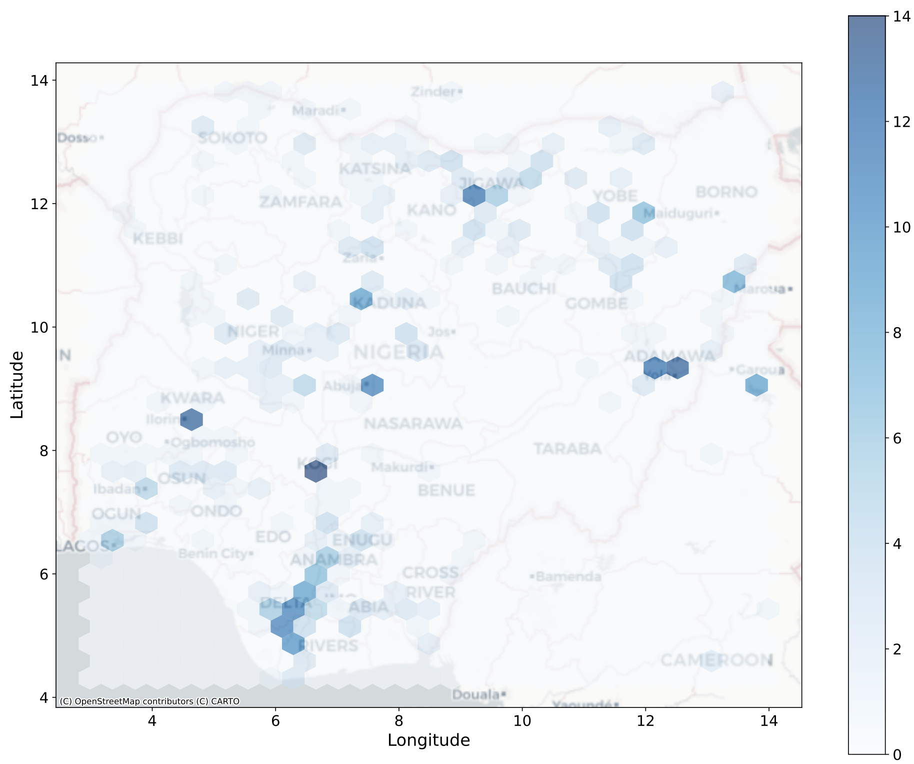


Figure 5.15: Place names mentioned in flood-related news articles clustered in hexagons representing the count of geocoded place names inside each hexagon.

The quadrat statistic (Figure 5.16) and the G function (Figure 5.17) quantitatively underline the first visual impression, that the geocoded place names are not uniformly located in Nigeria. The cells with the highest counts are roughly located along a north-east to south-west axis (Figure 5.16). The cells in the very north-west and south-east contain the least amount of points. The cell in the south-east is mainly located over Cameroon, but within the bounding box used for the disambiguation of place names. The cell in the very south-west holds the highest count, even though a substantial area of that cell is located over the Golf of Guinea. The chi-squared χ^2 test used to assess the likelihood that the distribution of points shown in Figure 5.16 matches a uniform distribution across the cells returned a low p-value of 6.887×10^{-24}, indicating a clustered point pattern. The quadrat statistic is further supported by the G function in Figure 5.17A. The distance to point (d) is shown on the x-axis, whereas the y-axis represents the fraction of nearest neighbour distances smaller than d. The empirical cumulative distribution of the nearest neighbour distances for the pattern shown in Figure 5.17B is represented by red line (Observation). The black belt represents the middle 95% of simulations representing the cumulative distribution for randomly spaced point patterns. The turquoise line is the median of all simulations. In a clustered pattern, points are typically much closer to each other compared to those in dispersed or random distribution. The red line in Figure 5.17 therefore represents a clustered point pattern, as the function rises much faster over distance than the median function representing a random distribution of points.

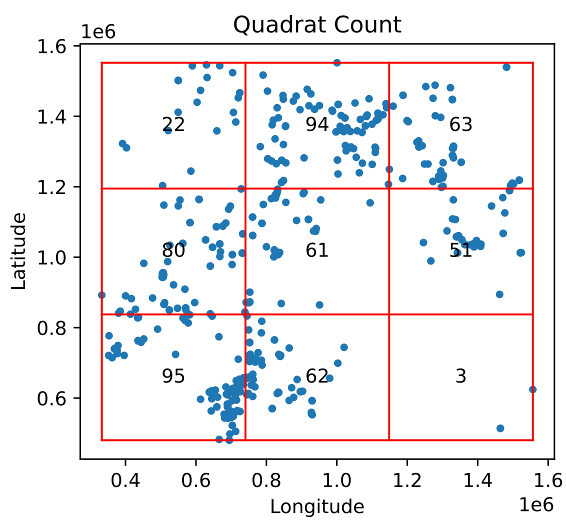


Figure 5.16: Three by three grid used to count the quadrat statistics for the underlying point pattern.

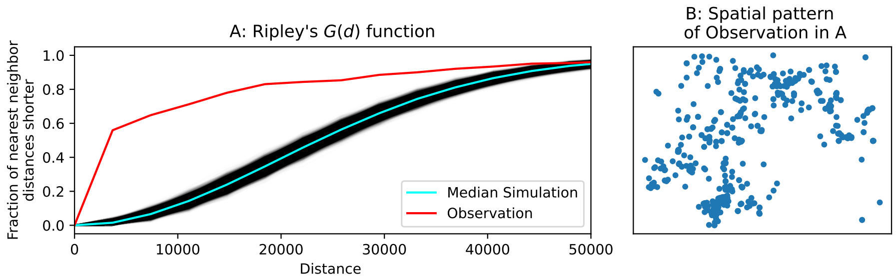


Figure 5.17: Visualisation of Ripley's G (A) for the point pattern in B.

Figure 5.18 shows the DBSCAN clusters (red points) for the analysed point pattern. Gray points represent noise points outside of any clusters. A radius value of 20km and a minimum number of points within a cluster value of 5 were chosen to compute the clusters as shown in Figure 5.18. The size and positions of the clusters resemble some similarities to the dark blue hexagons in Figure 5.15. The biggest cluster can be observed in the very south of the country.

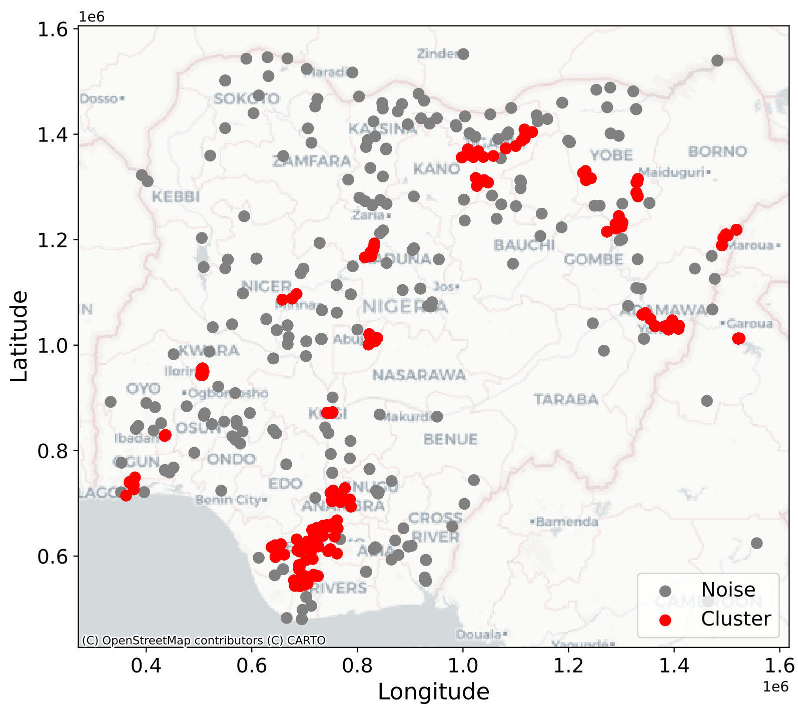


Figure 5.18: Spatial clusters for the analysed point pattern using DBSCAN.

5.3 News Article Publication Date

This section presents the results regarding the publication dates of DailyPost news articles in the GDELT GKG dataset. For 18'081 news articles the publication date of news articles provided by GDELT in column 'V2.1DATES' were compared to the publication dates directly scraped from the respective online news websites as described in Section 4.3.

No difference in the publication date could be determined for 18'056 analysed newspaper articles. A difference between the publication dates could therefore be determined for 26 news articles. In 21 cases, the difference was one day. The scraped publication date was always one day in advance, when compared to the date provided by GDELT under the 'V2.1Dates' column. The news articles with one day difference were evenly distributed between the analysed months May till December 2022. In two instances, there was a two day gap. Both news articles were released in May. Similarly, the publication dates that were scraped directly from the websites were ahead of the publication dates provided by GDELT. In one other case, the time difference was six days and in one case ten days. The news article with the 6 day difference was published in November 2022 and the one with the 10 day difference was published in December 2022. The publication dates scraped from the websites were again ahead of the publication dates provided by GDELT. In one case, no publication date could be scraped from the website. When manually entering the URL in a browser, it led to the main page of DailyPost instead to a specific news article. In the other 25 cases, the scraped publication dates and the publication dates provided by GDELT were compared to the publication dates on the respective websites. In all cases, the publication dates matched with the scraped publication dates.

5.4 Numeric Attribute-Value Pairs

This section dives into the results produced by the rule-based AV pair extraction method described in Section 4.4. The extracted AV pairs were computationally tested for content similarity with synthetically generated GPT-4 AV pairs. On the other hand, a manual comparison was done to see how well the method was able to extract AV pairs from newspaper articles in GDELT GKG.

Cosine similarity

Table 5.6 shows the cosine similarity between AV pairs extracted by the rule-based method and the AV pairs provided by GPT-4 for 50 compared news articles. The average cosine similarity per AV pair is very comparable to the average cosine similarity per article. Initially, the GPT-4 AV pairs were compared to the rule-based AV pairs. Conversely, the rule-based AV pairs were also compared to the GPT-4 AV pairs. However, no differences were found. A cosine similarity score of 0.775 and 0.776 suggests a relatively high degree of similarity between the rule-based AV pairs and the GPT-4 AV pairs on AV pair level as well as article level.

Table 5.6: Average cosine similarity per AV pair and per article based on 50 news articles comparing GPT-4 AV pairs to rule-based AV pairs and rule-based AV pairs to GPT-4 AV pairs.

	GPT-4 compared to Rule-Based	**Rule-Based compared to GPT-4**
Cosine Similarity per AV pair	0.775	0.775
Cosine Similarity per Article	0.776	0.776

Table 5.7 shows three examples of GPT-4 and rule-based AV pairs out of all AV pairs, which were used to calculate the average cosine similarities in Table 5.6. The AV pairs of GPT-4 are much easier to understand without further context than the AV pairs of the rule-based method. In some cases, the GPT-4 AV pairs almost represent entire sentences. The AV pairs extracted using the rule-based method focus much more on the pure AV pair and provide much less context. The first example only mentions that over 60 lives were claimed. However, the geographical information and the cause for this incident are missing. Although the first rule-based AV pair contains less information than the GPT-4 AV pair, the information it conveys is very clear. The second AV pair, extracted using the rule-based method, lacks clarity without additional context. This is because certain words like 'estimated' are misplaced within the word order of the sentence and other words like 'experienced' and 'with' do not contribute additional context to the AV pair. Again, the cause for the loss of 100'000 cattle is not given. However, this AV pair scored a higher similarity score, when compared to the GPT-4 AV pair than the comparison in the first row of Table 5.7. The last example in the rule-based AV pairs column is the least clear, when compared to its counterpart in column GPT-4 AV pair. This can also be seen in the lowest cosine similarity of all three examples. However, considering that cosine similarity returns values between -1 and 1, the cosine similarity value is still rather high for this comparison.

Table 5.7: Three examples of GPT-4 and rule-based AV pairs and their respective cosine similarity score.

GPT-4 AV pairs	Rule-Based AV pairs	Cosine Similarity
The landslides in Bududa district claimed over 60 lives	over 60 lives claimed	0.812
An estimated 100,000 cattle lost due to the drought	100,000 experienced with estimated cattle lost	0.849
1.5 million primary school pupils are out of school	1.5 million crisis as school pupils	0.730

Manual comparison

For 50 flood-relevant news articles from DailyPost all AV pairs were manually extracted. These 50 news articles contained a total of 215 AV pairs. In 27 out of 50 news articles analysed, numerical AV pairs are already part of the title of an article. In most cases, these numerical AV pairs are repeated again in the main text of the article and presented with more context. Out of the 215 AV pairs, 146 (67.9%) were successfully extracted using the rule-based method. In contrast, GDELT identified only 127 (59%) AV pairs for the same set of 50 news articles. Additionally, the percentage of matching AV pairs per news article was analysed. On average, 75.4% of the AV pairs extracted using the rule-based method corresponded with those manually extracted from each news article. A match was given, if the attribute and value matched exactly between the two compared AV pairs. In this comparison, GDELT also achieved a slightly lower value of 64%.

All 215 manually extracted AV pairs were additionally categorised into thematic classes, which can be seen in Figure 5.19. Over 40% of the AV pairs in these newspaper articles provide information, on how many people were affected by a flood disaster in one way or another. This typically includes quantifying the number death, injured or displaced people. On the other hand, reports can also provide details about the number of individuals successfully rescued from such emergency situations. 20% of AV pairs quantify, how infrastructure was affected by a flood. This category includes more detailed descriptions, such as the number of roofs in a building complex damaged by rainfall and the specific departments of the company located within the complex that were affected as a result. Other AV pairs roughly estimated, how many houses in a village were partially or completely damaged by the flood. This category of AV pairs is closely followed by those that provide a temporal context to the newspaper articles. For example, they may offer precise timing details, such as when the heavy rain began or when emergency services arrived at the accident scene. Conversely, less precise forecasts were also made regarding the potential arrival time of the flood, aiming to alert the readers of the newspaper article. Monetary AV pairs appeared less frequently in the newspaper articles, but primarily in relation with damaged infrastructure and the extent of the financial losses incurred. Monetary AV pair expressions were often used to provide a rough estimation of the monetary damage, for example: 'properties worth millions of naira destroyed'. Rarely exact hydro-meteorological parameters such as the amount of precipitation in mm were integrated into newspaper articles as AV pairs. 10% of the analysed AV pairs were thematically too diverse and were therefore grouped together under the category 'Others'.

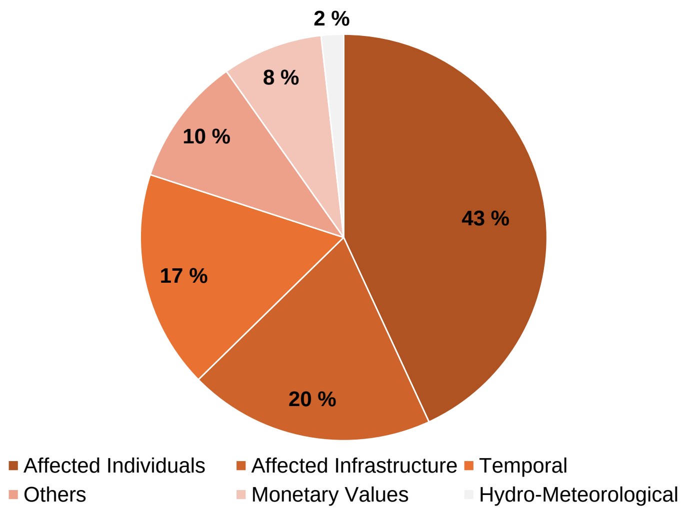


Figure 5.19: Percentage of thematic categories for numeric AV pairs extracted from 50 news articles.

Chapter 6

Discussion

This chapter analyses the results of this thesis in terms of the RQs posed in Section 1.2. The RQs are discussed in the following sections of this chapter:

- [RQ1.1] How can the largest number of flood-relevant news articles from GDELT be extracted for a specific area of interest? → Section 6.1

- [RQ1.2] Is it feasible to extract place name information with a greater level of spatial detail than the place name data available in GDELT? → Section 6.2

- [RQ1.3] Can the GDELT timestamp reliably be used to filter news articles published within an interested time frame? → Section 6.3

- [RQ1.4] What kind of quantitative information is most commonly reported in flood-related news articles? → Section 6.4

- [RQ2] Can flood-related information from news articles in the GDELT database be systematically extracted to identify flood hotspots in the study area of Nigeria? → Section 6.2

Since the RQs can be answered independently of each other, each section begins by discussing the results within a section in a coherent manner. From this, limitations are derived and considerations are made as to how the individual models and methods can be improved. Finally, each section explicitly refers back to the individual RQs.

6.1 Text Classification

In the following sections, the results of the text classification model are discussed coherently in order to summarise the strengths and limitations of the model.

6.1.1 Text Classification of News Articles in the Validation and Test Datasets

The performance of the text classification model in determining, whether a news article is flood related or not, has yielded exceptional results, as shown by the very high precision, recall and F1-Score achieved on the validation and test dataset (Table 5.1). These metrics serve as robust indicators of the model's effectiveness in correctly classifying flood-related articles, while minimising both FP and FN.

The article falsely predicted as not flood-related news article in Figure 5.3 discusses the firsthand impact of climate change on Palau, a small Pacific island nation, narrated by President Tommy Remengesau. Rising sea levels flooded his backyard, eroded coastlines and disrupted traditional agriculture. The article falsely predicted as flood-related in Figure 5.3 is very special in context. It contains a comprehensive list of states, territories and countries worldwide. Additionally, it includes a note about joining conversations on Torstar, requiring users to have a registered account. This article lacks proper content and is not flood related at all. Three out of the four FP news articles in the validation dataset (Figure 5.2) also have a very similar content as the FP classified news article in the test dataset. The fourth FP news article reported about a non-profit Christian Water Mission, deploying a team to assess safe water needs for Ukrainian refugees during the ongoing war. This analysis reveals that certain news articles utilised for training the text classification model do not accurately reflect genuine news content. Such news articles can impact the performance of the text classification model.

6.1.2 Text Classification of GDELT News Articles

The manual examination of the TP news articles from the GDELT GKG test dataset reveals that the majority of the articles focus on the impacts of floods on human lives and infrastructure. This is followed by articles that contextualise the flood impact within a political framework and discuss emergency relief measures. A smaller portion of the articles analyse the possibilities, plans and implementations of flood prevention projects. Additionally, only a minority of newspapers provide quantitative information on specific hydro-meteorological characteristics of floods (Figure 5.5). The flood-relevant topics discussed align with the authors' perspectives in Section 2.4. However, the amount of news articles within these topics varies slightly from Gambo (2018) and Rashid (2011). The findings also align with the hypothesis by Panem et al. (2014), which says that authors of news articles tend to focus more on the location of a disaster and the number of affected people. Alongside the immediate and direct effects of a flood disaster, articles also discusses consequential impacts, including outbreaks of cholera caused by contaminated drinking water or an uptick in snakebites resulting from the floods. Articles reporting about the immediate impact of flood disasters generally also include initiated emergency measures to minimise the extent of the disaster, including the organisation of emergency camps and the provision of most essential goods for daily life. A common theme that arises, is the financial support provided to those affected, whether through local funds or international organisations such as: 'UN donates $10.5m to Nigeria for flood, shelter,water'. Articles exploring the reasons behind current circumstances often incorporate a political dimension: 'But the Bayelsa State Governor waited until the floods overtook the entire state killing helpless people before he released paltry N450m for relief materials' or 'I cannot but express disappointment with the total absence of federal agencies like NEMA, as well as SEMA which would have brought in some level of professionalism into the operations' are two common examples, how individuals or governmental organisations are blamed for their inactivity. However, occasionally political actors are also commended: 'Niger Delta group applauds NDDC for fixing damaged portion of East-West Road.' As outlined by Albrecht (2022), numerous flood-related subjects are politicised within the analysed articles, aiming to influence social dynamics, shape political influence and advocate for transformative initiatives. Preventive measures range from simpler measures, such as warning the population of impending disasters to larger infrastructural measures ,such as the construction of flood defences and the maintenance of water drains. Sentences containing detailed hydro-meteorological quantities are only very infrequently encountered in news articles. Rare examples are: 'The predicted rainfall amount in southern part of the state may likely to be above 1,460 milimeters and in the eastern part is likely to be 1,260 to 1,360 milimeters while northern part will experience below 110 milimeters.' More commonly, time frames with exceptionally high rainfall intensity are mentioned.

While flood impact emerges as the predominant flood-related topic in the TP articles, less attention is given to it in the FN articles. Discussed impact topics predominantly revolve around the consequential issues arising after a flood disaster. For instance, a flood disaster led to fuel shortages, resulting in elevated fuel prices or exacerbated unemployment rates, as more individuals found themselves unemployed. While TP flood impact articles frequently include coverage of emergency measures, FN flood impact articles tend to contextualise the entire issue within a political framework. Affected people in numerous articles request from decision makers to take decisive action at last. Politicians engage in blame games, attributing responsibility to each other for the occurrence of such a significant catastrophe, while also urging readers to vote for the appropriate party to prevent future flood disasters. Flood prevention and hydro-meteorological topics are hardly discussed in these articles. It is noteworthy to mention that a specific format of newspaper articles recurs frequently among the FN articles. These articles always begin with the following sentence: 'Nigerian Newspapers: 10 things you need to know this Friday morning, Good morning! Here is today's summary from Nigerian Newspapers :' All of these articles serve as summaries of the ten most crucial current topics, each addressed in succession within a few sentences. These articles therefore report simultaneously on flood-relevant and not flood-relevant topics.

The trained text classification model only predicted a few not flood-related news articles as flood-related articles. None of these articles contain the word flood, which might have been used in a metaphorical way. Whereas the model correctly predicted two articles using flood in a metaphorical sense as not flood relevant, GDELT classified them wrongly as flood relevant. Further research on GDELT's thematic classification quality of news articles is needed. However, 24.8% of the manually analysed subset of flood filtered news articles using GDELT's classification themes in column 'V1THEMES' were not flood relevant. This discovery aligns with Williams (2020)'s finding that the thematic classes in 'V1THEMES' of GDELT's GKG dataset make it challenging to pinpoint articles exclusively or primarily focused on a natural disaster.

6.1.3 Enhancing the Text Classification Model: Limitations and Recommendations

It is noteworthy that the model's performance (Table 5.1) was evaluated on a test dataset comprised of news articles from the same source as the training and validation data. This consistency in data source possibly helped to ensure that the model's performance remained high, when tested on the validation and test dataset. The observed decrease in recall performance, when tested on the GDELT test dataset, highlights an important aspect of model's ability to generalise. The GDELT dataset consists of news articles sourced from various news agencies, introducing variations in language, writing styles and thematic coverage that may differ from the original training data. As a result, the model's ability to identify flood-related articles may be impacted due to differences in data distribution and domain specific nuances. While the precision remains relatively high for the GDELT test dataset, the lower recall suggests that the model may struggle to capture all relevant flood-related articles from this new data source. This discrepancy underscores the importance of evaluating model performance on diverse datasets to assess its generalisation capabilities across different domains and sources.

Moving forward, it is imperative to address the challenges posed by domain shift and dataset heterogeneity. This may involve retraining the model on a more diverse dataset that encompasses a wider range of sources and domains, as well as implementing robust evaluation strategies to continuously monitor and improve its performance across various contexts. Additionally, the setup for training the model could be enhanced by preventing oscillations during the training process. Since this was the very first ML model that was trained, no hyperparameter tuning was conducted due to inexperience and the very high scores achieved on the validation and test datasets.

To specifically address and answer RQ1.1, it is advisable to train the text classification model using a diverse assortment of articles related to the topic. This approach will maximise the extraction of flood-related news articles. Additionally, the model could also be specifically trained on sub-classes of flood-related news articles. With the current state of the text classification model trained here, hardly any not flood-relevant news articles are misclassified as flood-relevant. However, some flood-relevant news articles also remain undetected by the model.

6.2 Place Name Recognition and Resolution

This section aims to address, whether the NER model trained in this thesis can identify and localise place names with greater spatial resolution compared to the algorithm utilised by GDELT (Leetaru, 2012). To address this question, the discussion consolidates the findings from Section 5.2.

6.2.1 Toponym Recognition

Figure 6.1 shows, how the trained NER model can rather accurately detect Nigerian place names in a news article. Various wards in Nigeria such as Akarai, Otuoku or Umuti are labelled as GEO, indicating that the model correctly predicted those tokens as Nigerian place names. Although these place names describe rather unknown and very small scale areas in Nigeria, the model recognises these place names precisely. This method offers a significant advantage over simpler approaches that rely solely on string matching. Such methods require a comprehensive gazetteer encompassing also numerous smaller geographical locations to function efficiently (SaiKrishna et al., 2012). Such gazetteers are not commonly available for Nigeria (Onuoha and Chukwueke, 2023).

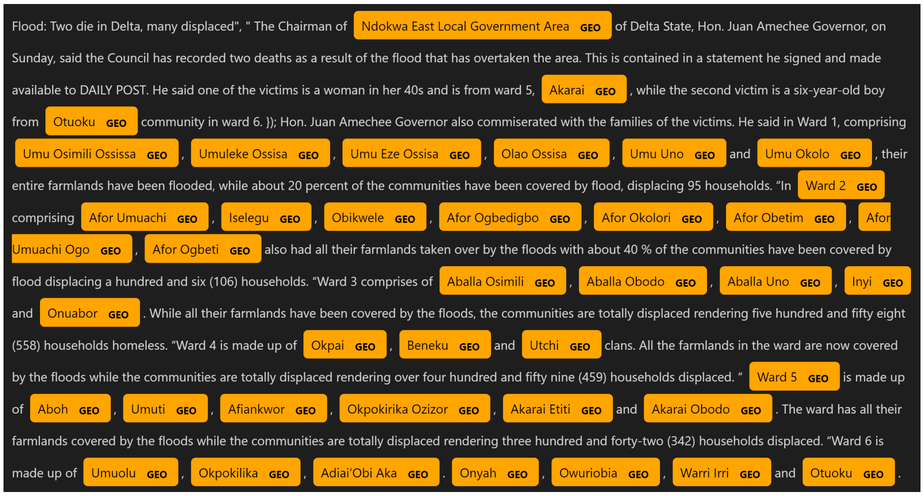


Figure 6.1: Trained NER model accurately detecting Nigerian wards in a news article.

Nevertheless, the model also exhibits weaknesses. While it successfully identifies numerous small scale locations, it fails to recognise 'Delta' and 'Delta State' in the first sentence of the news article (Figure 6.1). This corresponds to the finding shown in Figure 5.11. Although the impact of a flood on people and infrastructure can be assessed more accurately using smaller scale geographical descriptions, higher level geographical regions are also of great relevance. This is especially true with regard to toponym resolution (Leidner, 2008). The issue is further highlighted in Figure 6.2, where Nigerian states represent the only geographical information within the entire newspaper article. Although the Nigerian states are embedded in the text as an enumeration very similar to the Wards in Figure 6.1, the model does not recognise the states in the text. The same problem also exists with FCT Abuja. However, 6.3 shows, how Kwara State is correctly recognised by the model as part of a more specific geographic description. Further investigations are needed to find out, why the model struggles to recognise Nigeria, its 36 states and FCT in news articles, while simultaneously demonstrating its performance at recognising Nigerian place names at lower administrative levels. However, this way the model is easier to improve than if it would recognise the country and its states, but struggles to identify smaller place names. For example, one could simply supplement the NER model with a rule-based approach, which recognises Nigeria and its states using a string matching algorithm as used to automatically annotate training data for this NER model (Section 4.2.1).

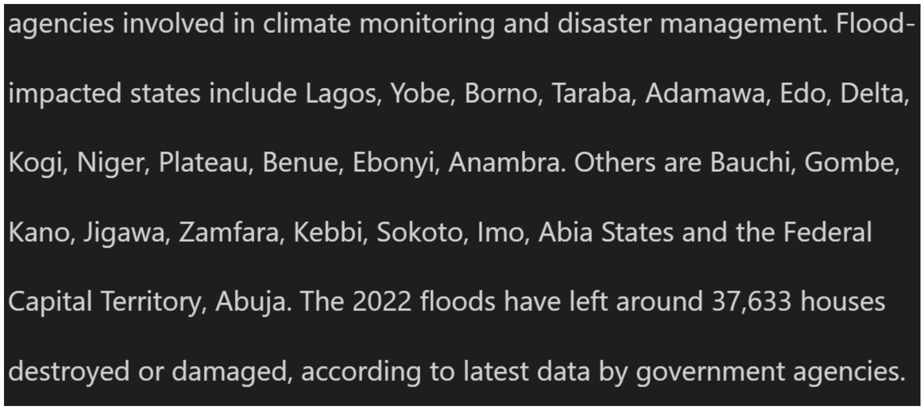

agencies involved in climate monitoring and disaster management. Flood-impacted states include Lagos, Yobe, Borno, Taraba, Adamawa, Edo, Delta, Kogi, Niger, Plateau, Benue, Ebonyi, Anambra. Others are Bauchi, Gombe, Kano, Jigawa, Zamfara, Kebbi, Sokoto, Imo, Abia States and the Federal Capital Territory, Abuja. The 2022 floods have left around 37,633 houses destroyed or damaged, according to latest data by government agencies.

Figure 6.2: Trained NER model not recognising Nigerian state names.

Another disadvantage of the model is that some state institutions are also recognised as place names, especially if these state institutions contain place names in their description (Figure 6.3). Names of individuals can also be falsely recognised as geographical locations by the model. Such misclassifications lead to an increased number of FP classifications, reducing the overall performance of the model. The situation is somewhat more intricate due to the dual role of state organisations, which can appear both as actors and geographical locations within newspaper articles. For example, the model correctly predicted 'Kogi State House of Assembly Quarters' as a place name within the whole sentence 'Our correspondent's visit to the Kogi State House of Assembly Quarters along Ganaja road this morning shows that the quarter has been totally submerged, and many of the members have relocated to an undisclosed camp within the metropolis.' In this context, 'Kogi State House of Assembly' does not represent an actor, instead, it describes the physical buildings (quarters) of the state organisation.

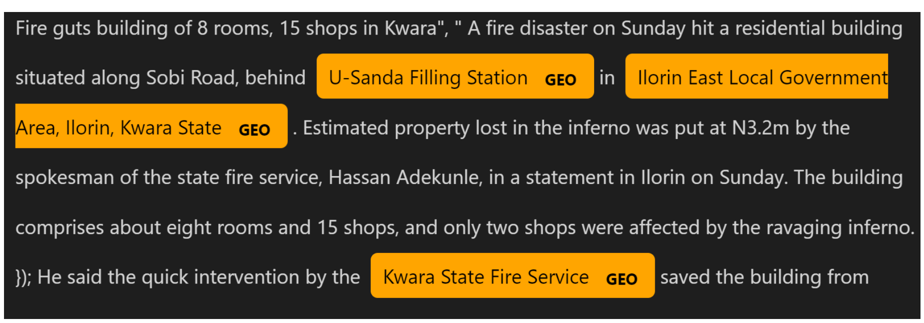

Fire guts building of 8 rooms, 15 shops in Kwara", " A fire disaster on Sunday hit a residential building situated along Sobi Road, behind U-Sanda Filling Station GEO in Ilorin East Local Government Area, Ilorin, Kwara State GEO . Estimated property lost in the inferno was put at N3.2m by the spokesman of the state fire service, Hassan Adekunle, in a statement in Ilorin on Sunday. The building comprises about eight rooms and 15 shops, and only two shops were affected by the ravaging inferno. }); He said the quick intervention by the Kwara State Fire Service GEO saved the building from

Figure 6.3: Trained NER model recognising Kwara State as part of a more detailed geographical description and wrongly labelling Kwara State Fire Service as place name.

One method of reducing the number of FP classifications would be to train the NER model to recognise the names of people and organisations as well. Tedeschi and Navigli (2022) used 15 NER categories in their NER model to improve the entity recognition and disambiguation performance of the model. For example, spaCy's 'en_core_web_lg' statistical model is trained to recognise PERSON (real and fictional peoples), NORP (nationalities or political/religious groups) and ORGANIZATION (companies, agencies and institutions) (spaCy, 2024). Figure 6.4 displays a normalised confusion matrix derived from spaCy's 'en_core_web_lg' model for predicting Nigerian place names. Although the model hardly predicts not Nigerian place names as such, it struggles to accurately recognise words representing Nigerian place names, supporting findings by Akdemir et al. (2018) that existing trained NER models tend to prioritise place names from the regions or countries, from where the data is sourced from.

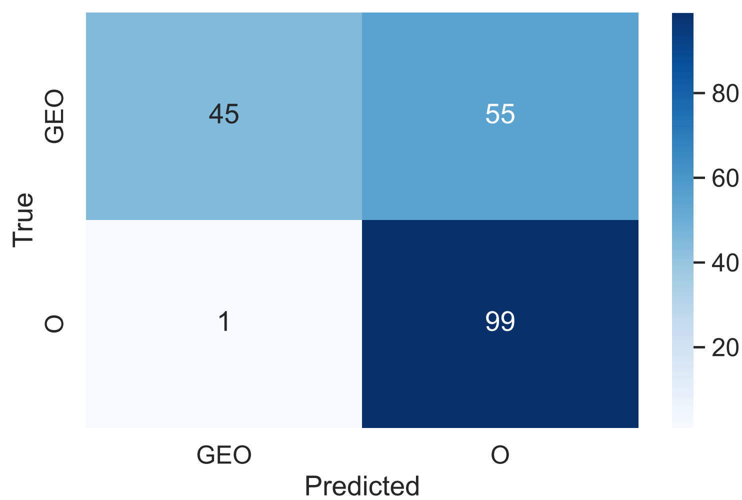


Figure 6.4: Normalised confusion matrix of spaCy's 'en_core_web_lg' model predicting Nigerian place names.

Given that the entire process of generating training data for this NER model was fully automated for this thesis (Figure 4.7), a comparable approach could be employed to embed and annotate Nigerian names and organisations within synthetic news articles. However, further analysis is needed to assess, how effectively OpenAI's GPT-3.5 Turbo Model can seamlessly integrate these additional terms into a synthetic news article. Although the model was good at incorporating Nigerian place names into synthetic news articles, there were some articles, in which the Nigerian place names were not optimally incorporated into the context of the synthetic news articles: 'Title: Dassa Kalumari Introduces Vibrant Culinary Scene, Showcasing Delectable Cuisine at De Bubbles Restaurant & Bar Date: [Current Date] Location: Dassa Kalumari, Andu, Okobotuo, Unguwar Sarkinrafi, De Bubbles Restaurant & Bar, Feromara, Ajeromi/Ifelodun Dassa Kalumari, a town nestled in the heart of Nigeria, is emerging as a culinary hotspot, captivating locals and visitors alike with its diverse and flavorful cuisine. Leading the way in this gastronomic revolution is De Bubbles Restaurant & Bar, a culinary gem situated in the vibrant neighborhood of Unguwar Sarkinrafi, captivating taste buds and showcasing the unique flavors of the region [...].' In this example, all Nigerian place names that were passed to the GPT-3.5 Turbo Model to be included in the synthetic news article were included in the first sentence one after each the other. Later in the article, these place names however appear individually and embedded in more context.

Overall, FP classifications pose fewer issues, since generally candidates falsely identified as Nigerian place names will not be spatially resolved during the geocoding process (Figure 5.12). Nevertheless, during manual toponym resolution quality checks of 1'156 candidates for Nigerian place names, it was found that Nominatim incorrectly assigned latitude and longitude values to non-place names in 23 cases. Among the 23 instances, false geocoding occurred nine times for personal names, eight times for state institutions and twice for the national currency, the Naira of Nigeria. In four cases, no clear category could be assigned. Nevertheless, the model also showcases its ability to accurately differentiate between names of individuals and place names.

However, Figure 6.5 illustrates, how the model can also accurately distinguish between location names and personal names. In this part of an article, 'Miga' is used both as a location name and a person's name. The way 'Miga' was used in context and in the sentence structure allowed the model to correctly distinguish between the use of 'Miga' as a place name and a person's name. In another case, the model correctly predicted a place name not as a Nigerian place name, because the place name was not used geographically. Instead it was used to describe a dish: 'Also, Irish potatoes, tomatoes, pepper, onions, carrots, yams and others food Items have witnessed sharp rise as a result of the flooding.' GDELT on the other hand predicted Irish Potato as a place name and listed it under column 'V1LOCATIONS' (Table 3.2).

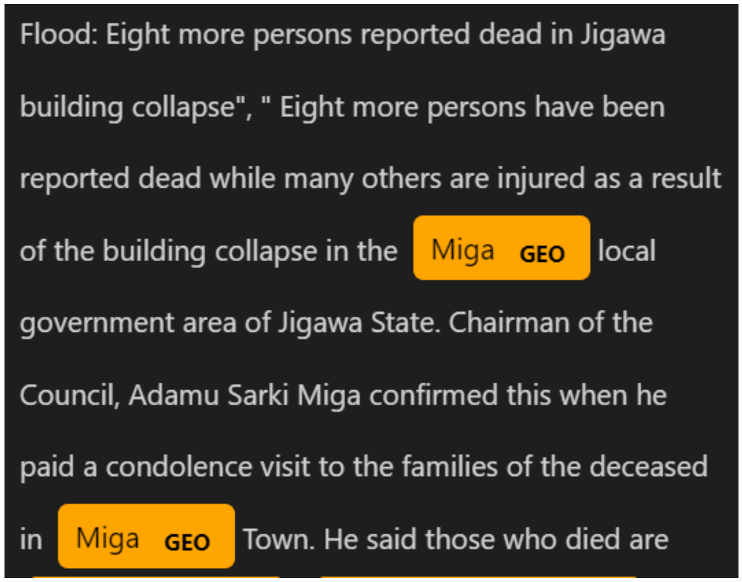


Figure 6.5: Trained NER accurately differentiating between cases, where Miga is used as a Nigerian place name and a person's name.

In general, the performance of the model can be increased even further by tuning the model even more specifically to the intended use. Only rudimentary hyperparameter tuning was conducted in this thesis due to the high computational intensity associated with retraining the NER model after each tuning step. There are many scientific studies that show, how a model can be further optimised with hyperparameter tuning (e.g. de Chavannes et al., 2021; Liao et al., 2022). Grid search, for example, rigorously explores a defined subset of hyperparameters, assessing the model's performance for every possible combination. This method requires significant computational resources. However, it guarantees a comprehensive coverage of the hyperparameter space (Shekar and Dagnew, 2019).

One notable issue of the model is the performance drop, when comparing the validation and test data (Table 5.3) against the GDELT test data (Table 5.4). In order to minimise this performance drop, the synthetic news articles must be optimised, so that they are more similar to the news articles available in the GDELT GKG data. In addition to the incorporation of state organisations and Nigerian personal names into the synthetic news articles, the process of automated news article generation can also be optimised. Currently, after each generated news article by the GPT-3.5 Turbo model, the automated workflow responded with the following sentence: 'Thanks sounds good. Now create another news article. Choose a new topic.[...]' (Section 4.2.1). The only parameters altered were the selection and quantity of Nigerian place names to be included in a synthetic news article. With varying levels of satisfaction in the responses sent to the GPT-3.5 Turbo model in reaction to the generated articles and the use of newer GPT models such as GPT-4, a greater variance in writing style could be incorporated in the synthetically created news articles. The automated annotation algorithm developed here also annotates Nigerian place names, even if they occur within a word. 'Bo' is for example a Nigerian place name. The algorithm would for example annotate 'Bo' within the word 'carbon'. Although such erroneous annotations are not used by spaCy to train the NER model and thus do not directly influence the model's performance, the algorithm could be revised, since annotated data is generated, which is then not used for model training.

6.2.2 Toponym Resolution

Although for nearly 50% of the place names, for which the spatial error was manually checked, no spatial error could be detected (Figure 5.14), the disadvantages of the toponym resolution approach used in this thesis can be divided into two groups: missing or alternatively spelled Nigerian place names in Nominatim's gazetteer and the chosen approach to disambiguate Nigerian place names in case of geo/geo ambiguities.

Figure 5.12 shows that for nearly 60% of all place names, for which Nominatim did not return latitude and longitude values, are indeed Nigerian place names. 'Atan Offot' for example is a place name located within the LGA Uyo in the state of Akwa Ibom. Whereas Nominatim did not return any coordinates for this place name, Google Maps returned (Latitude, Longitude: 5.006231334283372, 7.900175265868835) as a coordinate pair for this place name. Since this location was not contained in the gazetteer, Nominatim was also unable to provide coordinates. Another issue leading to Nominatim's inability to provide coordinates is the alternative spellings of place names. The trained NER model for example recognised 'Zankuwa' as a Nigerian place name. However, Nominatim could not provide any coordinates for this place name. 'Zankuwa' can be alternatively spelled also as 'Zenkuwa' or 'Zonkwa' (GeoNames, 2024). Nominatim only contains the alternative spelling 'Zonkwa' in its gazetteer. The OSM website[1] returns (Latitude, Longitude: 9.7874714, 8.2887424) for 'Zenkuwa' and 'Zonkwa'. Another problem for Nominatim is that the trained NER model sometimes extracts very specific place names. For example, Nominatim has an entry for 'Shiroro' in the gazetteer[2], whereas there is no entry for 'Shiroro Dam I'. In the current workflow, the toponym is not resolved at all, meaning no coordinates are returned, even though Nominatim has an entry for the less precise place name description 'Shiroro'. One approach to partially address this issue would be to use additional geocoding services alongside Nominatim, either concurrently or sequentially. For example, the used python library geopy also provides access to geocoding services such as GeoNames or Google. To at least

[1] `https://www.openstreetmap.org/node/501439761`
[2] `https://www.openstreetmap.org/node/501441530#map=14/9.9561/6.8332`

partially resolve cases like 'Shiroro Dam I', one could also program an algorithm that, for unsuccessful geocoding attempts for the entire string, would query for substrings and thus at least obtain the coordinates for 'Shiroro'.

Another disadvantage of the toponym resolution methodology used here is that other place names found in the same news article were not utilised to correctly resolve place names with geo/geo ambiguity. All recognised place names by the trained NER model in Figure 6.6 were geocoded correctly apart from 'Ringim Local Government Area' and 'Daura'. 'Ringim Local Government Area' was not geocoded by Nominatim, because its gazetteer only contains an entry for 'Ringim' and not 'Ringim Local Government Area'. 'Daura' is a place name with geo/geo ambiguity within Nigeria. On one hand it represents a place name in the state of Kano and on the other hand a place name in the state of Katsina. Nominatim returned the coordinates of 'Daura' in Kano[3], although the 'Daura' in Katsina[4] was referred to in the news article. Nominatim returned the coordinates for the wrong 'Daura', because it is ranked higher in Nominatim's ranking algorithm, which was used in addition to the bounding box of Nigeria to resolve toponym ambiguity (Section 4.2.3). However, when reading though the text in Figure 6.6, it becomes very clear, to which 'Daura' the news article was referring to. There are two possibilities, how the geo/geo ambiguity could be resolved. On one hand, one could use overarching place names found in the same text to obtain the correct coordinates for 'Daura' in the state of Katsina (Adelfio and Samet, 2013). However, this would require first to improve the NER model so that it accurately recognises all states and the FCT of Nigeria. On the other hand, one could calculate the average distances between all place names in a newspaper article and use a distance threshold to resolve the geo/geo ambiguity around place name candidates (Smith and Crane, 2001).

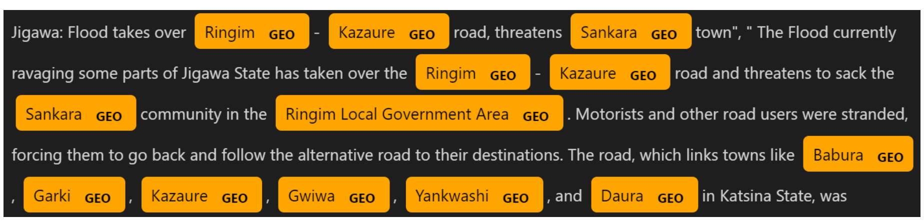


Figure 6.6: Example visualising, how other place names found in the same news paper article can help to correctly disambiguate place names with geo/geo ambiguity such as Daura.

Figure 6.7 visualises an issue that concerns both toponym recognition and toponym resolution. Although the trained NER model reliably identifies place names in the newspaper article, it would be difficult to determine exactly, where the car was recovered by the fire department solely based on the recognised place

[3]`https://www.openstreetmap.org/relation/3710422`
[4]`https://www.openstreetmap.org/relation/3711491`

names and without additional context. One would probably guess somewhere in the Asa river, close to Sobi road in Ilorin. However, the news article contains a very specific description of where the car was recovered by the fire department: 'from the Asa river, Opposite Olusola Saraki Abattoir, along Sobi Road in the Ilorin metropolis'. If the trained NER model had recognised the entire description of the recovery location as one place name, one could have answered the question of where the car was recovered very accurately, even without additional context. On the other hand, it would be rather challenging for automated toponym resolution processes to assign coordinates to this place name, since it is unlikely that this precise place name with coordinates would exist in any gazetteer.

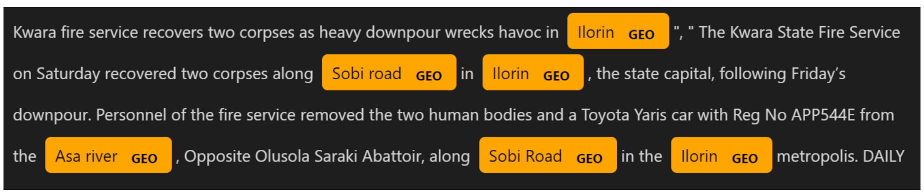


Figure 6.7: Example visualising, how other place names found in the same news paper article can help to correctly disambiguate place names with geo/geo ambiguity such as Daura.

Summary of toponym recognition and resolution

Using Figure 5.11, RQ1.2 can be answered very clearly. The trained NER model identifies more Nigerian place names than GDELT at a sub state level. However, the applied toponym recognition and toponym resolution methods would need to be revised so that, on one hand, the FCT and the states of Nigeria are recognised more reliably. On the other hand, there would be a need for a better gazetteer specifically for Nigeria, which could accurately locate these place names in the geographical space. Additionally, the method for resolving place names with geo/geo ambiguity within Nigeria would need to be revised.

6.2.3 Spatial Clustering of Analysed Point Pattern

All geocoded and manually improved place names in flood-relevant DailyPost news articles published between May and December 2022 were analysed for their spatial distribution in order to answer RQ2. The first visual impression in Figure 5.15 indicating a clustered distribution of the place names was supported by the more quantitative statistics produced by the quadrat analysis (Figure 5.16) and Ripley's G function (Figure5.17). The DBSCAN algorithm (Figure 5.18) produced a very similar amount and distribution of clusters when compared to Figure 5.15.

The biggest DBSCAN cluster is located in the south of Nigeria. This finding is supported by various other news articles stating that especially the states in the south such as Bayelsa, Delta and Rivers were affected by the floods in 2022 (BBC, 2022; France24, 2022; Vanguard, 2023). Oyero (2022) on the other hand states that Bayelsa was not under the top 10 affected states. However, Jigawa is one of the most affected state according to that news article. Figure 5.18 shows also two clusters in the state of Jigawa. Figure 6.8 shows the affected LGA's in the states of Adamawa, Borno and Yobe between June and August 2022 (IOM, 2022). Some of the DBSCAN clusters are also located in the southern LGAs of Yobe. Although place names in the northern part of Yobe appeared in the analysed news articles, they were not classified as clusters by the DBSCAN algorithm. Another DBSCAN cluster corresponds with Borno's LGAs Bama and Gwoza in the east of the state. However, hardly any place names from Damboa, Jere and Monguno appeared in the analysed news articles. This comparison shows that at the LGA level, the clusters only partially match. The comparison of Figure 6.9 with the DBSCAN clusters is interesting. Although, hardly any place names from the states Plateau, Nasarawa and Taraba appeared in the analysed news articles, resulting in no visible clusters in these areas, these states were marked as affected by the International Federation of Red Cross (IFRC) (Figure 6.9). Conversely, DBSCAN identified several smaller clusters and noise points in the south-western states like Ogun and Oyo. These were not classified as affected by the IFRC (IFRC, 2022).

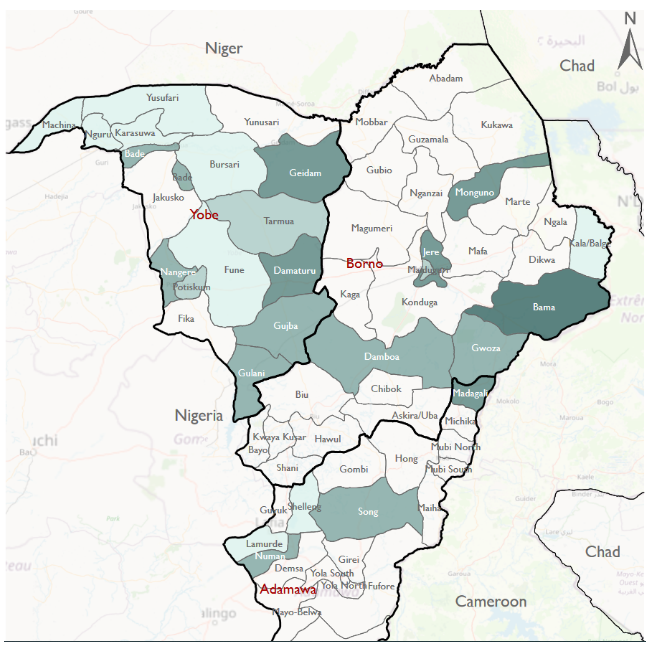


Figure 6.8: Flood affected LGAs in Adamawa, Borno and Yobe between June and August 2022 (IOM, 2022).

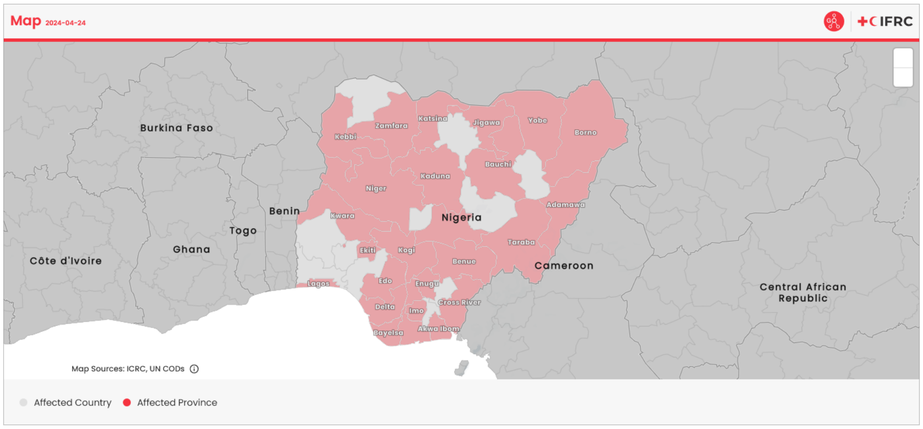


Figure 6.9: Flood affected Nigerian states 2022 (IFRC, 2022).

Based on Figure 5.18, it can be stated that clusters were identified based on the analysed data. One downside of th DBSCAN method is that the output heavily depends on the definded radius and the defined number of points to be within that radius to be counted as a cluster. However, the output of DBSCAN is also confirmed by the quantitative quadrat statistics and Ripley's G function. Qualitative comparisons show that the clusters partially correspond very well with other sources, sometimes even at the sub-state level. However, discrepancies are noted too. Therefore, RQ2 can be answered affirmatively. However, the analysis identified flood-related hotspots rather than specific flood hotspots, as the news articles analysed generally reported on topics relevant to flooding. Analysing more news articles with the methods outlined here would be beneficial. Particularly, news agencies in addition to DailyPost should also be considered.

6.3 News Article Publication Date

This section seeks to determine, if there are any discrepancies between the publication dates provided by GDELT and the dates scraped from the websites for the same news articles and therefore consolidates the findings from Section 5.3.

No discrepancies were found in the publication dates for more than 99% of the 18'081 news articles analysed. For the remaining articles, the publication dates provided by GDELT did not match those listed on the news websites. However, the difference was only one day in most instances. Only in four out of 18'081 cases the difference is more than one day. Since GDELT does not offer documentation on how the publication dates were determined, it was not possible to conduct further analyses on the discrepancies found in the publication dates. However, RQ1.3 from Section1.2 can be answered clearly. In over 99% of the cases, no difference was detected and the publication dates provided by GDELT can be reliably used for further analysis.

6.4 Numeric Attribute-Value Pairs

This section aims to consolidate the advantages and disadvantages of the rule-based AV pair extraction method (Section 4.4), when compared to the GPT-4 AV pairs and GDELT AV pairs (Section 5.4). In addition, the question of which quantitative information appears most frequently in flood-relevant news articles will be answered here.

6.4.1 Comparison with GPT-4 AV Pairs

Table 5.6 shows that the rule-based AV pair extraction method developed in this thesis can extract AV pairs from synthetic news articles, which are from a cosine similarity point of view very similar to the AV pairs provided by GPT-4. The GPT-4 AV pairs represent near perfect AV pairs, because they provide the fundamental attribute and value along with sufficient context. This allows users to gain a comprehensive overview of the quantitative information in a news article without needing to read the entire article. Even though the fundamental attribute and value can be extracted quite straight forward with POS and dependency tags, it is more challenging to extend the attribute and value part with more context. The fundamental attribute and value can be generally extracted with spaCy using the 'nummod' dependency tag. Context relevant to the attribute or value may appear significantly distant from the actual attribute or value itself. To link these components through POS and dependency tags, it is usually necessary to develop very specific rules (Section 4.4.1) that do not generalise well, when applied to another set of sentences. Rule-based AV pair extraction algorithms generally rely on a set of specific rules and heuristics designed to identify certain patters or structures in text (Sari et al., 2010). This factor, along with the need to rearrange words from a sentence into a numerical AV pair so that they logically align, means that even with a high cosine similarity, one AV pair might be straightforward for the reader to understand, while the other could be difficult to comprehend (Table 5.7). Despite the previously mentioned drawbacks of rule-based AV pair extraction methods, Table 6.1 demonstrates, how this method can augment the fundamental attribute and value with additional context for simpler sentence structures.

6.4.2 Comparison with GDELT AV Pairs

The rules outlined in Section 4.4.1 also encounter challenges, when extracting specific AV pairs from the news articles in the GDELT GKG dataset. Certain reoccurring patterns are discussed here. The model encounters difficulties, when

Table 6.1: Three GDELT and rule-based AV pairs exemplifying the improvement of the rule-based AV pair extraction method over the algorithm used by GDELT.

GDELT AV pairs	**Rule-Based AV pairs**
220, houses	220 houses in Jigawa destroyed
200, families	200 families in Jigawa displaced
over 100 demolition of houses	FCTA begins demolition of over 100 houses

two numeric values, which refer to each other, appear in the same sentence. Examples for such patterns are: ’predicted rainfall [...] and in the eastern part is likely to be 1,260 to 1,360 milimeters [...]’, ’Fifteen passengers of the 16-passenger boat that capsized [...]’ or ’[...] 90% of 264 rooms completely destroyed’. Figure 6.10 and Figure 6.11 visualise the complex dependency structure of the second and third example. GDELT extracts the second example as the following numeric AV pair: ’15, passengers on the 16-passenger’ and the rule-based method as two separate AV pairs: ’Fifteen passengers on boat found’ and ’6 passenger boat’. With a very specific set of rules, both numeric AV pairs can be extracted. However, there is a certain probability that these rules will not work on a third example, where two numeric values refer to each other.

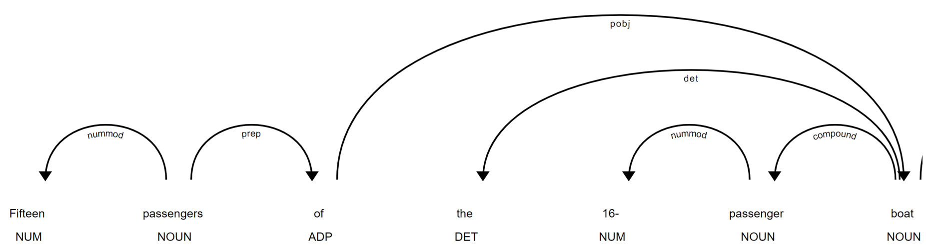


Figure 6.10: Complex dependency structure of two numeric values part of a AV pair structure (Example 1).

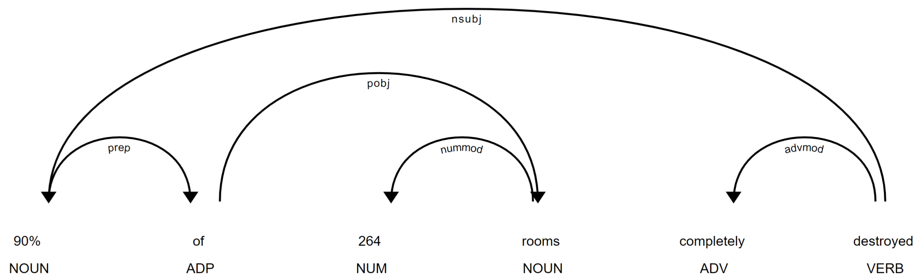


Figure 6.11: Complex dependency structure of two numeric values part of a AV pair structure (Example 2).

The established rules also face difficulties in extracting numeric AV pairs representing monetary values, such as: 'properties worth millions of naira', 'N500m worth of investment' or 'cost several millions of Naira'. GDELT extracts the first example as the following AV pair: '1000000,of naira destroyed' and the established rules did not extract the AV pair at all. Interestingly, GDELT's numeric AV pair extraction algorithm translates the term 'millions' into '1000000'. Although 'millions' and '1000000' are almost synonymous, GDELT's algorithm erroneously translates numbers that use a comma as a thousand separator. For example, the numeric AV pair '1,260 milimeters (precipitation)' is extracted by GDELT as '260, milimeters (precipitation)'. GDELT's algorithm possibly splits the sentence into sub sentences at the comma. This special case was considered in Section 4.4.1, when the rules were developed as part of this thesis.

Although the dependency structures for monetary AV pairs would be much easier to implement than the dependency structures for AV pairs that contain two numbers referencing each other, these two examples very clearly demonstrate the limitations of rule-based AV pair extraction algorithms. With each additional implemented rule, an attempt is made once more to solve another special case.

Ultimately, the specific RQ posed in the context of numeric AV pair extraction should be answered. Based on Figure 5.19, it can be stated that most numeric AV pairs in newspaper articles quantify the impact of flood disasters on affected people and infrastructure. This finding is also partially supported by the research done by Gambo (2018). However, this statement is based on a very small and manually analysed test dataset, which includes only news articles from DailyPost. To enhance the validity of the statement, the test dataset should be expanded to include several different news sources.

Chapter 7

Conclusion

Although the methods described here have their limitations, they provide a solid foundation for obtaining an initial spatial reference point in otherwise data-scarce regions, where flood-related issues are discussed in news articles. Additionally, this thesis presents an approach to extract quantitative information from flood-related articles. As the first step, a text classification model was trained to automatically categorise news articles into flood-relevant and not flood-relevant articles. Even though the model rarely misclassified articles as flood relevant when they are not, it should be optimised to more accurately identify and classify the different nuances within flood-relevant articles. Next, a NER model was trained to specifically identify Nigerian place names in news articles. An innovative approach was developed, using GPT-3.5 Turbo in conjunction with a string matching algorithm to automate the time-intensive task of generating annotated training data. To improve the automated generation of annotated data further, it is advisable to also annotate additional categories of entities that also resemble Nigerian place names. Other place names appearing in the same news article should be considered to resolve ambiguities around certain place names. A gazetteer specifically compiled for Nigeria would further support the toponym resolution process. The extracted place names were analysed for their spatial patterns over the studied period. The clustered point distribution revealed a large cluster in the south of Nigeria. Finally, a set of rules was developed to extract quantitative information from news articles, demonstrating that additional context, in which the quantitative information is embedded, is very useful to understand the extracted information. However, it was also shown that rule-based methods can only be exploited to a certain extent. It must be noted that the analyses conducted in this thesis were only based on a subset of the data available in GDELT. To make even more precise statements, the analysed timeframe should be extended using GDELT's timestamp attribute.

Bibliography

Acheson, E. (2019). *Extracting and modeling the geography of text documents : resources and applications*. PhD thesis, University of Zurich, Zürich.

Acheson, E., De Sabbata, S., and Purves, R. S. (2017). A quantitative analysis of global gazetteers: Patterns of coverage for common feature types. *Computers, Environment and Urban Systems*, 64:309–320.

Adekola, O. and Lamond, J. (2018). A media framing analysis of urban flooding in Nigeria: current narratives and implications for policy. *Regional Environmental Change*, 18(4):1145–1159.

Adelekan, I. O. (2011). Vulnerability assessment of an urban flood in Nigeria: Abeokuta flood 2007. *Natural Hazards*, 56(1):215–231.

Adelekan, I. O. and Asiyanbi, A. P. (2016). Flood risk perception in flood-affected communities in Lagos, Nigeria. *Natural Hazards*, 80(1):445–469.

Adelfio, M. D. and Samet, H. (2013). Structured toponym resolution using combined hierarchical place categories. In *Proceedings of the 7th Workshop on Geographic Information Retrieval*, GIR '13, page 49–56, New York, NY, USA. Association for Computing Machinery.

Agrawal, T. (2021). *Hyperparameter optimization in machine learning: make your machine learning and deep learning models more efficient*. Springer.

Agyepong, K. and Kothari, R. (1997). Controlling Hidden Layer Capacity Through Lateral Connections. *Neural Computation*, 9(6):1381–1402.

Ahlers, D. (2013). Assessment of the accuracy of GeoNames gazetteer data. In *Proceedings of the 7th Workshop on Geographic Information Retrieval*, GIR '13, page 74–81, New York, NY, USA. Association for Computing Machinery.

Akdemir, A., Hürriyeto\uglu, A., Yörük, E., Gürel, B., Yoltar, , and Yüret, D. (2018). Towards Generalizable Place Name Recognition Systems: Analysis and Enhancement of NER Systems on English News from India. In *Proceedings of the 12th Workshop on Geographic Information Retrieval*, GIR'18, New York, NY, USA. Association for Computing Machinery.

Albrecht, F. (2022). Natural hazards as political events: framing and politicisation of floods in the United Kingdom. *Environmental Hazards*, 21(1):17–35.

Amangabara, G. and Obenade, M. (2015). Flood Vulnerability Assessment of Niger Delta States Relative to 2012 Flood Disaster in Nigeria. *American Journal of Environmental Protection*, 3(3):76–83.

Amitay, E., Har'El, N., Sivan, R., and Soffer, A. (2004). Web-a-where: geotagging web content. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '04, page 273–280, New York, NY, USA. Association for Computing Machinery.

Anandika, A., Mishra, S. P., and Das, M. (2021). Review on Usage of Hidden Markov Model in Natural Language Processing. In Mishra, D., Buyya, R., Mohapatra, P., and Patnaik, S., editors, *Intelligent and Cloud Computing*, pages 415–423, Singapore. Springer Singapore.

Andres, N. and Badoux, A. (2019). The Swiss flood and landslide damage database: Normalisation and trends. *Journal of Flood Risk Management*, 12(S1):913–925.

Archer, D., O'Donnell, G., Lamb, R., Warren, S., and Fowler, H. J. (2019). Historical flash floods in England: New regional chronologies and database. *Journal of Flood Risk Management*, 12(S1):1–14.

Ardanuy, M. C. and Sporleder, C. (2017). Toponym disambiguation in historical documents using semantic and geographic features. In *Proceedings of the 2nd International Conference on Digital Access to Textual Cultural Heritage*, DATeCH2017, page 175–180, New York, NY, USA. Association for Computing Machinery.

Arellano, A., Zontek-Carney, E., and Austin, M. (2015). Frameworks for Natural Language Processing of Textual Requirements. *International Journal on Advances in Systems and Measurements*, 8:230–240.

Ashlin, A. and Ladle, R. J. (2007). 'Natural disasters' and newspapers: Post-tsunami environmental discourse. *Environmental Hazards*, 7(4):330–341.

Badaru, Y. U., Ishiaku, O., and Nassir, Y. M. (2014). Analysis of Sensor Imaging and Field-Validation for Monitoring, Evaluation and Control Future Flood Prone Areas along River Niger and Benue Confluence Ecology, Lokoja, Nigeria. *environment*, 4(22):90–99.

Bali, R. (2024). Disaster Management Cycle. *Asian Journal of Geographical Research*, 7(1):85–93.

Bartschi, M. (1985). An Overview of Information Retrieval Subjects. *Computer*, 18(5):67–84.

Bashir O., O., Oludare H., A., Johnson O., O., and Aloysius, B. (2012). Floods of Fury in Nigerian Cities. *Journal of Sustainable Development*, 5(7).

Bauer-Marschallinger, B., Cao, S., Tupas, M. E., Roth, F., Navacchi, C., Melzer, T., Freeman, V., and Wagner, W. (2022). Satellite-Based Flood Mapping through Bayesian Inference from a Sentinel-1 SAR Datacube. *Remote Sensing*, 14(15):1–28.

BBC (2022). Nigeria Flooding 2022: Pictures of how flood take scata communities for Rivers, Delta and Bayelsa states.

Bekkar, M., Djemaa, H. K., and Alitouche, T. A. (2013). Evaluation measures for models assessment over imbalanced data sets. *J Inf Eng Appl*, 3(10).

Bennett, B. and Agarwal, P. (2007). Semantic Categories Underlying the Meaning of 'Place'. In Winter, S., Duckham, M., Kulik, L., and Kuipers, B., editors, *Spatial Information Theory*, pages 78–95, Berlin, Heidelberg. Springer Berlin Heidelberg.

Bennett, D. and Townend, J. (2012). The Scandal of Selective Reporting. *British Journalism Review*, 23(2):60–66.

Bergstra, J. and Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of machine learning research*, 13(2).

Boettke, P., Chamlee-Wright, E., Gordon, P., Ikeda, S., Leeson, P. T., and Sobel, R. (2007). The Political, Economic, and Social Aspects of Katrina. *Southern Economic Journal*, 74(2):363–376.

Bohensky, E. L. and Leitch, A. M. (2014). Framing the flood: A media analysis of themes of resilience in the 2011 Brisbane flood. *Regional Environmental Change*, 14(2):475–488.

Brown, D. K., Harlow, S., García-Perdomo, V., and Salaverría, R. (2018). A new sensation? An international exploration of sensationalism and social media recommendations in online news publications. *Journalism*, 19(11):1497–1516.

Brunner, T. J. and Purves, R. S. (2008). Spatial autocorrelation and toponym ambiguity. In *Proceedings of the 5th Workshop on Geographic Information Retrieval*, GIR '08, page 25–26, New York, NY, USA. Association for Computing Machinery.

Buscaldi, D. (2011). Approaches to disambiguating toponyms. *SIGSPATIAL Special*, 3(2):16–19.

Buscaldi, D. and Magnini, B. (2010). Grounding toponyms in an Italian local news corpus. In *Proceedings of the 6th Workshop on Geographic Information Retrieval*, GIR '10, New York, NY, USA. Association for Computing Machinery.

Buscaldi, D. and Rosso, P. (2008). A conceptual density-based approach for the disambiguation of toponyms. *International Journal of Geographical Information Science*, 22(3):301–313.

Cardoso, A. B., Martins, B., and Estima, J. (2022). A Novel Deep Learning Approach Using Contextual Embeddings for Toponym Resolution. *ISPRS International Journal of Geo-Information*, 11(1).

Charles, Z. and Papailiopoulos, D. (2018). Stability and Generalization of Learning Algorithms that Converge to Global Optima. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 745–754. PMLR.

Chater, N. and Manning, C. D. (2006). Probabilistic models of language processing and acquisition. *Trends in Cognitive Sciences*, 10(7):335–344.

Chaudhuri, S. (2015). Urban poor, economic opportunities and sustainable development through traditional knowledge and practices. *Global Bioethics*, 26(2):86–93.

Coetzee, C. and Van Niekerk, D. (2012). Tracking the evolution of the disaster management cycle: A general system theory approach. *Jàmbá: Journal of Disaster Risk Studies*, 4(1).

Cohen, J., Heinilä, K., Huokuna, M., Metsämäki, S., Heilimo, J., and Sane, M. (2022). Satellite-based flood mapping in the boreal region for improving situational awareness. *Journal of Flood Risk Management*, 15(3):1–14.

Cowie, J. and Lehnert, W. (1996). Information extraction. *Commun. ACM*, 39(1):80–91.

Crichton, D. (2002). UK and Global Insurance Responses to Flood Hazard. *Water International*, 27(1):119–131.

de Bruijn, J. A., de Moel, H., Jongman, B., de Ruiter, M. C., Wagemaker, J., and Aerts, J. C. (2019). A global database of historic and real-time flood events based on social media. *Scientific Data*, 6(1):1–12.

de Chavannes, L. H., Kongsbak, M. G. K., Rantzau, T., and Derczynski, L. (2021). Hyperparameter Power Impact in Transformer Language Model Training. In Moosavi, N. S., Gurevych, I., Fan, A., Wolf, T., Hou, Y., Marasović, A., and Ravi, S., editors, *Proceedings of the Second Workshop on Simple and Efficient Natural Language Processing*, pages 96–118, Virtual. Association for Computational Linguistics.

DeLozier, G., Baldridge, J., and London, L. (2015). Gazetteer-Independent Toponym Resolution Using Geographic Word Profiles. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1).

Devitt, C. and O'Neill, E. (2017). The framing of two major flood episodes in the Irish print news media: Implications for societal adaptation to living with flood risk. *Public Understanding of Science*, 26(7):872–888.

Diakakis, M., Mavroulis, S., and Deligiannakis, G. (2012). Floods in Greece, a statistical and spatial approach. *Natural Hazards*, 62(2):485–500.

Dilley, M. and Boudreau, T. E. (2001). Coming to terms with vulnerability: a Critique of Definiiton of Food Security. *Food Policy*, 26(3):229–247.

Dor, D. (2003). On newspaper headlines as relevance optimizers. *Journal of Pragmatics*, 35(5):695–721.

Du, S., Gu, H., Wen, J., Chen, K., and Van Rompaey, A. (2015). Detecting flood variations in shanghai over 1949-2009 with Mann-Kendall tests and a newspaper-based database. *Water (Switzerland)*, 7(5):1808–1824.

Emeka, A., SooveBenki, W., and Benjamin, E. (2023). Impacts of the 2022 Flooding on the Residents of Yenagoa, Bayelsa State, Nigeria. *Greener Journal of Environmental Management and Public Safety, vol. 11, no. 1, pp. 1-6, 2023 By Ajumobi, VE, Womboh, SB; Ezem, SB (2023). Greener Journal of Environment Management and Public Safety*, 11(1):1–6.

Escobar, M. P. and Demeritt, D. (2014). Flooding and the framing of risk in British broadsheets, 1985–2010. *Public Understanding of Science*, 23(4):454–471.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., and others (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231.

Ezema, I. J. (2023). Availability and Access to Open Government Data in Nigeria: A Content Analysis of Government Websites and Nigerian Data Portal. *International Information and Library Review*, 55(1):15–28.

Ferrés, D. (2007). Geographical information resolution and its application to the question answering systems. *Universitat Politècnica de Catalunya*.

Fize, J., Moncla, L., and Martins, B. (2021). Deep Learning for Toponym Resolution: Geocoding Based on Pairs of Toponyms. *ISPRS International Journal of Geo-Information*, 10(12).

France24 (2022). Nigeria's worst floods in a decade kill 500, displace 1.4 million.

Frankes, W. B. and Baeza-Yates, R. (1992). Information retrieval: Data structure \& algorithms. *PrenticeHall, Englewood cliffs, NJ.*

Füglister, B. and Purves, R. (2020). From text to coordinates: machine-coding the location of historical battles to create a new spatial conflict dataset.

Gaizauskas, R. and Wilks, Y. (1998). Information extraction: beyond document retrieval. *Journal of Documentation*, 54(1):70–105.

Gale, W. A., Church, K., and Yarowsky, D. (1992). One sense per discourse. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*.

Gambo, S. N. (2018). Nigerian newspaper framing of 2012 flooding disaster in Nigeria. *The Nigerian journal of communication, the journal of the African council for communication education, Nigerian chapter*, 15(2).

GDELT (2022). The GDELT Project.

Gebhardt, H., Meyer, S., Glaser, R., Radtke, U., and Reuber, P. (2011). *Geographie: Physische Geographie und Humangeographie*. Spektrum Akademischer Verlag.

GeoNames (2024). GeoNames.

Ghassabi, F. and Zare-Farashbandi, F. (2023). The role of media in crisis management: A case study of Azarbayejan earthquake. *International Journal of Health System and Disaster Management —*, 3(January):95–102.

Gil-Guirado, S., Pérez-Morales, A., and Lopez-Martinez, F. (2019). SMC-Flood database: A high-resolution press database on flood cases for the Spanish Mediterranean coast (1960-2015). *Natural Hazards and Earth System Sciences*, 19(9):1955–1971.

Goodfellow, I., Warde-Farley, D., Mirza, M., Courville, A., and Bengio, Y. (2013). Maxout Networks. In Dasgupta, S. and McAllester, D., editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1319–1327, Atlanta, Georgia, USA. PMLR.

Guha-Sapir, D., Rodriguez-Llanes, J. M., and Jakubicka, T. (2011). Using disaster footprints, population databases and GIS to overcome persistent problems for human impact assessment in flood events. *Natural Hazards*, 58(3):845–852.

HaCohen-Kerner, Y., Miller, D., and Yigal, Y. (2020). The influence of preprocessing on text classification using a bag-of-words representation. *PLOS ONE*, 15(5):1–22.

Haigh, I. D., Ozsoy, O., Wadey, M. P., Nicholls, R. J., Gallop, S. L., Wahl, T., and Brown, J. M. (2017). An improved database of coastal flooding in the United Kingdom from 1915 to 2016. *Scientific Data*, 4:1–10.

Haltas, I., Yildirim, E., Oztas, F., and Demir, I. (2021). A comprehensive flood event specification and inventory: 1930–2020 Turkey case study. *International Journal of Disaster Risk Reduction*, 56(December 2020):102086.

Hapuarachchi, H. A. P., Wang, Q. J., and Pagano, T. C. (2011). A review of advances in flash flood forecasting. *Hydrological processes.*, 25(18).

Hauptmann, A. G. and Olligschlaeger, A. M. (1999). Using location information from speech recognition of television news broadcasts. In *ESCA Tutorial and Research Workshop (ETRW) on Accessing Information in Spoken Audio*.

Heinrich, P., Hagemann, S., Weisse, R., Schrum, C., Daewel, U., and Gaslikova, L. (2023). Compound flood events: analysing the joint occurrence of extreme river discharge events and storm surges in northern and central Europe. *Natural Hazards and Earth System Sciences*, 23(5):1967–1985.

Hickey, D., Schmitz, M., Fessler, D., Smaldino, P. E., Muric, G., and Burghardt, K. (2023). Auditing Elon Musk's Impact on Hate Speech and Bots. *Proceedings of the International AAAI Conference on Web and Social Media*, 17(1):1133–1137.

Hill, L. L. (2000). Core Elements of Digital Gazetteers: Placenames, Categories, and Footprints. In Borbinha, J. and Baker, T., editors, *Research and Advanced Technology for Digital Libraries*, pages 280–290, Berlin, Heidelberg. Springer Berlin Heidelberg.

Honnibal, M. (2018). What loss function is optimized in TextCategorizer?

Hu, X., Zhou, Z., Li, H., Hu, Y., Gu, F., Kersten, J., Fan, H., and Klan, F. (2023). Location Reference Recognition from Texts: A Survey and Comparison. *ACM Comput. Surv.*, 56(5).

Hu, Y. (2018). Geo-text data and data-driven geospatial semantics. *Geography Compass*, 12(11):e12404.

Hulme, M. and Burgess, N. (2019). London's weather and the everyday: two centuries of newspaper reports. *Weather*, 74(8):286–290.

Hussaini, A. and Matazu, B. M. (2023). An overview of key improvements by the Nigerian Meteorological Agency for the modernisation of Meteorological Services in Nigeria. *Science World Journal*, 18(1):152–157.

Ibebuchi, C. C. and Abu, I.-O. (2023). Rainfall variability patterns in Nigeria during the rainy season. *Scientific Reports*, 13(1):7888.

IFRC (2022). Nigeria : Floods - 2022.

Inan, H., Khosravi, K., and Socher, R. (2017). Tying Word Vectors and Word Classifiers: A Loss Framework for Language Modeling.

IOM (2022). Nigeria — Flood Flash Report — Borno, Adamawa and Yobe States (June - August 2022).

Jones, C. B. and Purves, R. S. (2009). Geographical Information Retrieval. In LIU, L. and ÖZSU, M. T., editors, *Encyclopedia of Database Systems*, pages 1227–1231. Springer US, Boston, MA.

Kahle, M., Kempf, M., Martin, B., and Glaser, R. (2022). Classifying the 2021 'Ahrtal' flood event using hermeneutic interpretation, natural language processing, and instrumental data analyses. *Environmental Research Communications*, 4(5).

Karabiber, F. (2024). Cosine Similarity.

Kaźmierczak, A. and Cavan, G. (2011). Surface water flooding risk to urban communities: Analysis of vulnerability, hazard and exposure. *Landscape and Urban Planning*, 103(2):185–197.

Khurana, D., Koli, A., Khatter, K., and Singh, S. (2023). Natural language processing: state of the art, current trends and challenges. *Multimedia Tools and Applications*, 82(3):3713–3744.

Kingma, D. P. and Ba, J. (2017). Adam: A Method for Stochastic Optimization.

Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., and Brown, D. (2019). Text Classification Algorithms: A Survey. *Information*, 10(4).

Kron, W., Steuer, M., Löw, P., and Wirtz, A. (2012). How to deal properly with a natural catastrophe database - Analysis of flood losses. *Natural Hazards and Earth System Science*, 12(3):535–550.

Kulkarni, A. and Shivananda, A. (2021). Deep Learning for NLP. In *Natural Language Processing Recipes: Unlocking Text Data with Machine Learning and Deep Learning Using Python*, pages 213–262. Apress, Berkeley, CA.

Lai, K., Porter, J. R., Amodeo, M., Miller, D., Marston, M., and Armal, S. (2022). A Natural Language Processing Approach to Understanding Context in the Extraction and GeoCoding of Historical Floods, Storms, and Adaptation Measures. *Information Processing and Management*, 59(1):102735.

Larson, R. R. (1996). Geographic information retrieval and spatial browsing. *Geographic information systems and libraries: patrons, maps, and spatial information [papers presented at the 1995 Clinic on Library Applications of Data Processing, April 10-12, 1995].*

Leetaru, K. (2012). Fulltext geocoding versus spatial metadata for large text archives: Towards a geographically enriched wikipedia. *D-Lib Magazine*, 18(9/10).

Leidner, J. L. (2007). Toponym resolution in text: annotation, evaluation and applications of spatial grounding. *SIGIR Forum*, 41(2):124–126.

Leidner, J. L. (2008). *Toponym resolution in text: Annotation, evaluation and applications of spatial grounding of place names*. Universal-Publishers.

Leidner, J. L. and Lieberman, M. D. (2011). Detecting geographical references in the form of place names and associated spatial natural language. *SIGSPATIAL Special*, 3(2):5–11.

Lettieri, E., Masella, C., and Radaelli, G. (2009). Disaster management: Findings from a systematic review. *Disaster Prevention and Management: An International Journal*, 18(2):117–136.

Leveling, J. and Hartrumpf, S. (2008). On metonymy recognition for geographic information retrieval. *International Journal of Geographical Information Science*, 22(3):289–299.

Li, J. and Roy, D. P. (2017). A global analysis of Sentinel-2a, Sentinel-2b and Landsat-8 data revisit intervals and implications for terrestrial monitoring. *Remote Sensing*, 9(9).

Liao, L., Li, H., Shang, W., and Ma, L. (2022). An Empirical Study of the Impact of Hyperparameter Tuning and Model Optimization on the Performance Properties of Deep Neural Networks. *ACM Trans. Softw. Eng. Methodol.*, 31(3).

Lieberman, M. D., Samet, H., and Sankaranarayanan, J. (2010). Geotagging with local lexicons to build indexes for textually-specified spatial data. In *2010 IEEE 26th International Conference on Data Engineering (ICDE 2010)*, pages 201–212.

Lipton, Z. C., Berkowitz, J., and Elkan, C. (2015). A Critical Review of Recurrent Neural Networks for Sequence Learning.

Liu, J., Kong, X., Zhou, X., Wang, L., Zhang, D., Lee, I., Xu, B., and Xia, F. (2019). Data Mining and Information Retrieval in the 21st century: A bibliographic review. *Computer Science Review*, 34.

Llasat, M. C., Llasat-Botija, M., and López, L. (2009). A press database on natural risks and its application in the study of floods in Northeastern Spain. *Natural Hazards and Earth System Science*, 9(6):2049–2061.

Lu, X., Chan, F. K. S., Chan, H. K., and Chen, W.-Q. (2024). Mitigating flood impacts on road infrastructure and transportation by using multiple information sources. *Resources, Conservation and Recycling*, 206:107607.

Luino, F. (2016). Floods. In Bobrowsky, P. T. and Marker, B., editors, *Encyclopedia of Engineering Geology*, pages 1–6. Springer International Publishing, Cham.

Luo, X. (2021). Efficient English text classification using selected Machine Learning Techniques. *Alexandria Engineering Journal*, 60(3):3401–3409.

Luque, P., Gómez-Pujol, L., Marcos, M., and Orfila, A. (2021). Coastal Flooding in the Balearic Islands During the Twenty-First Century Caused by Sea-Level Rise and Extreme Events. *Frontiers in Marine Science*, 8.

Manning, Christopher; Raghavan, P. . and Schütze, H. (2009). *An Introduction to Information Retrieval.* Cambridge University Press.

Markert, K. and Nissim, M. (2002). Towards a Corpus Annotated for Metonymies: the Case of Location Names. In *LREC*.

Marrone, S., Papa, C., and Sansone, C. (2021). Effects of hidden layer sizing on CNN fine-tuning. *Future Generation Computer Systems*, 118:48–55.

McCurley, K. S. (2001). Geospatial mapping and navigation of the web. In *Proceedings of the 10th International Conference on World Wide Web*, WWW '01, page 221–229, New York, NY, USA. Association for Computing Machinery.

Min, B., Ross, H., Sulem, E., Veyseh, A. P. B., Nguyen, T. H., Sainz, O., Agirre, E., Heintz, I., and Roth, D. (2023). Recent Advances in Natural Language Processing via Large Pre-trained Language Models: A Survey. *ACM Comput. Surv.*, 56(2).

Muili, O. B. and Ikotun, A. B. (2013). Impacts of flood disaster in Agege local government area. *International Journal of Development and Sustainability*, 2(4):2354–2367.

Muniz-Rodriguez, K., Ofori, S. K., Bayliss, L. C., Schwind, J. S., Diallo, K., Liu, M., Yin, J., Chowell, G., and Fung, I. C.-H. (2020). Social Media Use in Emergency Response to Natural Disasters: A Systematic Review With a Public Health Perspective. *Disaster Medicine and Public Health Preparedness*, 14(1):139–149.

Nadkarni, P. M., Ohno-Machado, L., and Chapman, W. W. (2011a). Natural language processing: An introduction. *Journal of the American Medical Informatics Association*, 18(5):544–551.

Nadkarni, P. M., Ohno-Machado, L., and Chapman, W. W. (2011b). Natural language processing: an introduction. *Journal of the American Medical Informatics Association*, 18(5):544–551.

National Bureau of Statistics (2023). Nigeria Impact of Flood, Recovery and Mitigation Assessment Report 2022-2023, Final Report. Abuja, Nigeria. Technical report, National Bureau of Statistics.

NDMI (2011). Hazards , Disasters and Your Community. Technical report, NDMI.

NEMA (2002). NEMA-National Disaster Response Plan. *NEMA*.

Nkwunonwo, U. C. (2016). A review of flooding and flood risk reduction in Nigeria. *Global Journal of Human-Social Science B: Geography, Geo-Sciences, Environmental Science and Disaster Management*, 16(2):22–42.

Nominatim (2024). Importance.

Obiefuna, J., Adeaga, O., Omojola, A., Atagbaza, A., and Okolie, C. (2021). Flood risks to urban development on a coastal barrier landscape of Lekki Peninsula in Lagos, Nigeria. *Scientific African*, 12:e00787.

Okyere, C. Y., Yacouba, Y., and Gilgenbach, D. (2013). The problem of annual occurrences of floods in Accra: an integration of hydrological, economic and political perspectives. *Theoretical and Empirical Researches in Urban Management*, 8(2):45–79.

Onuoha, J. and Chukwueke, C. (2023). Extent of Accessibility and Utilization of Reference Materials by Nigerian University Undergraduates. *American Journal of Operations Management and Information Systems*, 8(1):12–20.

Oruonye, E. D. (2012). Socio-economic impact assessment of flash flood in Jalingo metropolis, Taraba State, Nigeria. *International Journal of Environmental Sciences*, 1(3):135–140.

Osayomi, T., Olobo Jnr, P., Ogunwumi, T., Fatayo, O. C., Akpoterai, L. E., Mshelia, Z. H., and Abatcha, I. U. (2018). Ife Social Sciences Review "I lost all I had to the flood...": A Post-Disaster Assessment of the 2018 Kogi State Flood in Nigeria. *Ife Social Sciences Review*, 2022(2):1–20.

Oyero, K. (2022). Flooding: Bayelsa Not Among 10 Worst-Hit States, Jigawa Is Number 1 – Minister.

Panem, S., Gupta, M., and Varma, V. (2014). Structured information extraction from natural disaster events on twitter. *International Conference on Information and Knowledge Management, Proceedings*, 2014-Novem(November):1–8.

Panteras, G. and Cervone, G. (2018). Enhancing the temporal resolution of satellite-based flood extent generation using crowdsourced data for disaster monitoring. *International Journal of Remote Sensing*, 39(5):1459–1474.

Pierdicca, N., Pulvirenti, L., and Chini, M. (2018). Flood Mapping in Vegetated and Urban Areas and Other Challenges: Models and Methods. In Refice, A., D'Addabbo, A., and Capolongo, D., editors, *Flood Monitoring through Remote Sensing*, pages 135–179. Springer International Publishing, Cham.

Pörtner, H.-O., Roberts, D., Tignor, M., Poloczanska, E., Mintenbeck, K., Alegría, A., Craig, M., Langsdorf, S., Löschke, S., Möller, V., Okem, A., Rama Phd, B., Belling, D., Dieck, W., Götze, S., Kersher, T., Mangele, P., Maus, B., Mühle, A., and Weyer, N. (2022). *Climate Change 2022: Impacts, Adaptation and Vulnerability Working Group II Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change.*

Pouliquen, B., Steinberger, R., Ignat, C., and De Groeve, T. (2004). Geographical information recognition and visualization in texts written in various languages. In *Proceedings of the 2004 ACM Symposium on Applied Computing*, SAC '04, page 1051–1058, New York, NY, USA. Association for Computing Machinery.

Poushter, J., Bishop, C., and Chwe, H. (2018). Social Media Use Continues to Rise in Developing Countries but Plateaus Across Developed Ones. *Pew Research Center*, June.

Powers, D. M. W. (2020). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation.

Purves, R. S., Clough, P., Jones, C. B., Hall, M. H., and Murdock, V. (2018a). Geographic information retrieval: Progress and challenges in spatial search of text. *Foundations and Trends in Information Retrieval*, 12(2-3):164–318.

Purves, R. S., Clough, P., Jones, C. B., Hall, M. H., and Murdock, V. (2018b). Geographic Information Retrieval: Progress and Challenges in Spatial Search of Text. *Foundations and Trends® in Information Retrieval*, 12(2-3):164–318.

Raggett, D., Le Hors, A., Jacobs, I., and others (1999). HTML 4.01 Specification. *W3C recommendation*, 24.

Rashid, H. (2011). Interpreting flood disasters and flood hazard perceptions from newspaper discourse: Tale of two floods in the Red River valley, Manitoba, Canada. *Applied Geography*, 31(1):35–45.

Rausch, A. S. (2014). The Great East Japan Disaster, 2011 and the Regional Newspaper: Transitions from News to Newspaper Columns and the Creation of Public Memory. *International Journal of Mass Emergencies & Disasters*, 32(2):275–296.

Rey, S. (2024). Quadrat Based Statistical Method for Planar Point Patterns.

Ripley, B. D. (1988). *Statistical inference for spatial processes.* Cambridge university press.

Rosenzweig, B. R., McPhillips, L., Chang, H., Cheng, C., Welty, C., Matsler, M., Iwaniec, D., and Davidson, C. I. (2018). Pluvial flood risk and opportunities for resilience. *WIREs Water*, 5(6):e1302.

Rözer, V., Müller, M., Bubeck, P., Kienzler, S., Thieken, A., Pech, I., Schröter, K., Buchholz, O., and Kreibich, H. (2016). Coping with Pluvial Floods by Private Households. *Water*, 8(7).

Ruiz-Garcia, M., Zhang, G., Schoenholz, S. S., and Liu, A. J. (2021). Tilting the playing field: Dynamical loss functions for machine learning. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on*

Machine Learning, volume 139 of *Proceedings of Machine Learning Research*, pages 9157–9167. PMLR.

SaiKrishna, V., Rasool, A., and Khare, N. (2012). String matching and its applications in diversified fields. *International Journal of Computer Science Issues (IJCSI)*, 9(1):219.

Sarawagi, S. (2008). Information Extraction. *Foundations and Trends® in Databases*, 1(3):261–377.

Sari, Y., Hassan, M. F., and Zamin, N. (2010). Rule-based pattern extractor and named entity recognition: A hybrid approach. In *2010 International Symposium on Information Technology*, volume 2, pages 563–568.

Schneiderbauer, S. and Ehrlich, D. (2004). Risk, hazard and people's vulnerability to natural hazards. *A review of definitions, concepts and data. European Commission Joint Research Centre. EUR*, 21410:40.

Shah, F. P. and Patel, V. (2016). A review on feature selection and feature extraction for text classification. In *2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, pages 2264–2268.

Sharma, A., Amrita, Chakraborty, S., and Kumar, S. (2022). Named Entity Recognition in Natural Language Processing: A Systematic Review. In Gupta, D., Khanna, A., Kansal, V., Fortino, G., and Hassanien, A. E., editors, *Proceedings of Second Doctoral Symposium on Computational Intelligence*, pages 817–828, Singapore. Springer Singapore.

Shekar, B. H. and Dagnew, G. (2019). Grid Search-Based Hyperparameter Tuning and Classification of Microarray Cancer Data. In *2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP)*, pages 1–8.

Sholademi, M. O., Iso, M. O., and Lawal, K. A. (2015). Dynamics of storm surge characteristics and its devastating flooding of Nigerian coast: A case study of Lagos beach. In *14th International Workshop on Wave Hindcasting and Forecasting \& 5th Coastal Hazard Symposium*.

Smith, D. A. and Crane, G. (2001). Disambiguating Geographic Names in a Historical Digital Library. In Constantopoulos, P. and Sølvberg, I. T., editors, *Research and Advanced Technology for Digital Libraries*, pages 127–136, Berlin, Heidelberg. Springer Berlin Heidelberg.

Socher, R., Bengio, Y., and Manning, C. D. (2012). Deep learning for NLP (without magic). In *Tutorial Abstracts of ACL 2012*, ACL '12, page 5, USA. Association for Computational Linguistics.

Solman, P. and Henderson, L. (2019). Flood disasters in the United Kingdom and India: A critical discourse analysis of media reporting. *Journalism*, 20(12):1648–1664.

SpaCy (2024). Embeddings, Transformers and Transfer Learning.

spaCy (2024). Industrial-Strength Natural Language Processing in Python.

SpaCy (2024a). Linguistic Features.

SpaCy (2024b). Training Pipelines & Models.

Srinivasa-Desikan, B. (2018). *Natural Language Processing and Computational Linguistics: A practical guide to text analysis with Python, Gensim, spaCy, and Keras*. Packt Publishing Ltd.

Susman, P., O'Keefe, P., and Wisner, B. (2019). Global disasters, a radical interpretation. In *Interpretations of calamity*, pages 263–283. Routledge.

Swanson, D. R. and others (1988). Historical note: Information retrieval and the future of an illusion. *Journal of the American Society for Information Science*, 39(2):92–98.

Swets, J. A. (1988). Measuring the Accuracy of Diagnostic Systems. *Science*, 240(4857):1285–1293.

Tedeschi, S. and Navigli, R. (2022). {M}ulti{NERD}: A Multilingual, Multi-Genre and Fine-Grained Dataset for Named Entity Recognition (and Disambiguation). In Carpuat, M., de Marneffe, M.-C., and Meza Ruiz, I. V., editors, *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 801–812, Seattle, United States. Association for Computational Linguistics.

Todd, David; Todd, H. (2011). Natural Disaster Response Lessons from Evaluations of the World Bank and Others. Technical report, World Bank.

Touretzky, D. S. and Pomerleau, D. A. (1989). What's hidden in the hidden layers. *Byte*, 14(8):227–233.

Trubshaw, B. (2012). How Anglo-Saxons Found Their Way. *PDF file. Marlborough: Heart of Albion. http://www. hoap. co. uk*.

Twele, A., Cao, W., Plank, S., and Martinis, S. (2016). Sentinel-1-based flood mapping: a fully automated processing chain. *International Journal of Remote Sensing*, 37(13):2990–3004.

UN/ISDR (2009). UNISDR Terminology on Disaster Risk Reduction (2009). Technical report, UN/ISDR.

Urama, N. E., Eboh, E. C., and Onyekuru, A. (2019). Impact of extreme climate events on poverty in Nigeria: a case of the 2012 flood. *Climate and Development*, 11(1):27–34.

Vanguard (2023). Flood: Jigawa, Rivers, Taraba, Cross River, Delta are five worst affected states – FG.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, , and Polosukhin, I. (2017). Attention is All you Need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Vennari, C., Parise, M., Santangelo, N., and Santo, A. (2016). A database on flash flood events in Campania, southern Italy, with an evaluation of their spatial and temporal distribution. *Natural Hazards and Earth System Sciences*, 16(12):2485–2500.

Voigt, S., Kemper, T., Riedlinger, T., Kiefl, R., Scholte, K., and Mehl, H. (2007). Satellite image analysis for disaster and crisis-management support. *IEEE Transactions on Geoscience and Remote Sensing*, 45(6):1520–1528.

Wagener, T., Sivapalan, M., Troch, P., and Woods, R. (2007). Catchment Classification and Hydrologic Similarity. *Geography Compass*, 1(4):901–931.

Wang, Q., Ma, Y., Zhao, K., and Tian, Y. (2022). A Comprehensive Survey of Loss Functions in Machine Learning. *Annals of Data Science*, 9(2):187–212.

Wang, W. and Stewart, K. (2015). Spatiotemporal and semantic information extraction from Web news reports about natural hazards. *Computers, Environment and Urban Systems*, 50:30–40.

Wesslen, R. (2022). Understanding the different terminology in the command line output of a training pipeline.

Willett, J., Baldwin, T., Martinez, D., and Webb, J. A. (2012). Classification of study region in environmental science abstracts. In *Proceedings of the Australasian Language Technology Association Workshop 2012*, pages 118–122.

Williams, B. (2022). NIGERIA EMERGENCY FLOOD RESPONSE 3.2M affected (incl. 1.9M children). *Unicef*.

Williams, S. (2020). Exploration of the Global Database of Events, Language and Tone (GDELT), with specific application to disaster reporting. *Office for National Statistics*.

Yadav, V. and Bethard, S. (2019). A Survey on Recent Advances in Named Entity Recognition from Deep Learning models.

Ying, X. (2019). An Overview of Overfitting and its Solutions. *Journal of Physics: Conference Series*, 1168(2):22022.

Yzaguirre, A., Smit, M., and Warren, R. (2016). Newspaper archives + text mining = rich sources of historical geo-spatial data. *IOP Conference Series: Earth and Environmental Science*, 34(1).

Zaratiana, U., Tomeh, N., Holat, P., and Charnois, T. (2022). Named Entity Recognition as Structured Span Prediction. In Han, W., Zheng, Z., Lin, Z., Jin, L., Shen, Y., Kim, Y., and Tu, K., editors, *Proceedings of the Workshop on Unimodal and Multimodal Induction of Linguistic Structures (UM-IoS)*, pages 1–10, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Zêzere, J. L., Pereira, S., Tavares, A. O., Bateira, C., Trigo, R. M., Quaresma, I., Santos, P. P., Santos, M., and Verde, J. (2014). DISASTER: A GIS database on hydro-geomorphologic disasters in Portugal. *Natural Hazards*, 72(2):503–532.

Zurich (2023). Three common types of floods explained.

Appendix A

GPT Conversation

Full answer of GPT-3.5 Turbo after priming prompt:

As a journalist at a newspaper agency, I specialize in writing news articles about various countries in Africa, with a focus on providing objective, informative, and neutral content. I strive to present the facts concisely and accurately, ensuring that my articles maintain a fair and balanced tone. When covering countries like Nigeria, I meticulously gather and verify information from credible sources, such as government reports, expert opinions, and interviews with relevant individuals. I aim to present a comprehensive picture of the current socio-political, economic, and cultural landscape, shedding light on both the challenges and positive developments. To maintain objectivity, I refrain from expressing personal opinions or biases in my writing. I adhere to ethical journalistic principles, ensuring that my articles remain impartial and free from any undue influence. By providing fair and balanced coverage, I strive to inform and educate readers about the diverse aspects of Nigeria's society, economy, and politics. Conciseness is key in my articles, as I understand that readers appreciate easily digestible information. While maintaining clarity and accuracy, I condense complex issues into easily comprehensible segments, making the news accessible to a wide range of readers. By employing an objective, informative, neutral, and concise approach, I strive to produce news articles that empower readers to stay informed about Nigeria and other African countries and make well-informed decisions based on reliable information.

Full first synthetic news article created by GPT-3.5 Turbo:

Title: Communities in Zungoro, Shiroro, Niger, Nigeria, Rebuild after Dam Tragedy - In the wake of the devastating dam collapse tragedy that struck these communities, residents are banding together to rebuild their lives and restore normalcy. The affected areas, located in the Niger state of Nigeria, have been grappling with the aftermath of the disaster as they embark on a journey of recovery. The incident occurred several weeks ago when heavy rainfall overwhelmed the Shiroro Dam, leading to its breach and the subsequent flooding of nearby communities. The catastrophe left many without their homes, displacing families and inflicting significant damage to infrastructure. One of the hardest-hit areas was Zungoro, a small rural village situated near the banks of the Niger River. Rescue efforts were swiftly mobilized to evacuate residents and provide emergency relief. The community displayed resilience and courage throughout the ordeal, supporting one another and demonstrating the strength of the Nigerian spirit. Despite the destruction caused by the flooding, the government and various non-governmental organizations (NGOs) acted promptly to assess the situation and extend aid to those affected. Lagos and Abuja, the capital cities of Nigeria, promptly dispatched relief materials, medical supplies, and humanitarian support to assist in the recovery efforts. The affected communities are gradually regaining stability and rebuilding what was lost. Local and international organizations are joining forces to provide temporary shelters, clean water, and medical assistance to those in need. Reconstruction efforts have begun, with a focus on repairing damaged homes, restoring infrastructure, and revitalizing essential services. The Nigerian government has committed to investing in long-term measures to prevent similar tragedies from occurring in the future. Plans for strengthening dam infrastructure and implementing better disaster management systems are underway. This initiative aims to safeguard not only the communities in Shiroro but also the wider region. The disaster has drawn attention to the importance of sustainable urban planning and improved infrastructure across Nigeria. Efforts to reinforce preventive measures and enhance emergency response systems will be essential in mitigating the risks associated with natural disasters. While the road to recovery may be long, residents of Zungoro, Shiroro, Niger, Nigeria, are determined to rebuild their communities and reclaim their lives. Through unwavering solidarity and resilience, they hope to bounce back stronger than ever. As the nation stands united with these affected areas, the support pouring in from various sectors signals a collective commitment to restoring normalcy. Together, Nigerians will continue to lend a helping hand to ensure the revival of these communities, showing the world the spirit of unity that defines Nigeria.

Appendix B

AV Rules (Text)

- If the current word is part of a 'nsubj' dependency and matches the attribute part of the extracted AV pair, the attribute is extended to contain the the head of the dependency too.

- If the current word is part of a 'compound' dependency and matches the attribute part of the extracted AV pair, the attribute is extended to contain the head of the dependency too. If the child does not match the attribute, but the head matches the attribute, the attribute is extended to contain the current word as well. Finally, if neither the child or the head matches the attribute part, but the head matches the value part, the value is extended to contain the current word as well.

- Check if the current word is part of a 'prep' dependency matches the attribute part of the extracted AV pair. If yes, the attribute part is extended to contain the current word. The attribute is further extended with the child of a 'pobj' dependency, if that 'pobj' dependency is found in within seven words apart of the current word.

- If the current word is part of a 'pobj' dependency and the head of the dependency matches the attribute part of the extracted AV pair, the attribute is extended to contain the head of the dependency too. If the head of the current word's dependency is the child of a 'prep' dependency, the attribute part is further extended. The attribute is extended with all children in a 'compound' relationship of the head of the head of the current word.

- If the current word is part of a 'quantmod' dependency and the head of the dependency matches the value part of the extracted AV pair, the value is extended to contain the current word too.

- If the current word is part of a 'dobj' dependency and matches the attribute part of the extracted AV pair, the attribute is extended to contain the head of the current word too. If the head represents a verb (POS:

VERB), the verb is turned into its past participle form before extending the attribute.

- If the current word is part of a 'advmod' dependency and the head of the dependency matches the value part of the extracted AV pair, the value is extended to contain all children of the 'advmod' relationship.
- If the current word is part of a 'amod' dependency and the head of the dependency matches the value part of the extracted AV pair, the value is extended to contain the current word. If the head does not match the value part, but matches the attribute part, the attribute is extended to contain the current word.
- If the current word is part of a 'acl' dependency and the head of the dependency matches the attribute part of the extracted AV pair, the attribute is extended to contain the current word.
- If the current word is part of a 'appos' dependency and the head of the dependency matches the attribute part of the extracted AV pair, the attribute is extended to contain the current word.
- If the current word is part of a 'nmod' dependency and the head of the dependency matches the value part of the extracted AV pair, the value is extended to contain the current word.
- If the current word is part of a 'nsubjpass' dependency and matches the attribute part of the extracted AV pair, the attribute is extended to contain the head of the current 'nsubjpass' dependency.

Personal Declaration

I hereby declare that the submitted thesis is the result of my own, independent work. All external sources are explicitly acknowledged in the Thesis.

Eric Tharmalingam

Allenwinden, 30.04.2024