Geo-parsing and Analysis of Road Traffic Crash Incidents for Data-Driven Emergency Response Planning

Article *in* Heliyon · December 2024

DOI: 10.1016/j.heliyon.2024.e41067

Patricia Idakwo

African University of Science and Technology

2 PUBLICATIONS **0** CITATIONS

SEE PROFILE

Olubayo Adekanmbi

City, University of London

28 PUBLICATIONS **56** CITATIONS

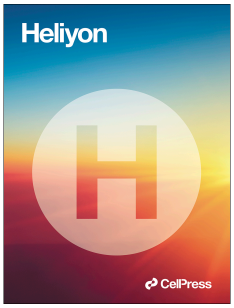
SEE PROFILE

Amos A. David

University of Lorraine

118 PUBLICATIONS **510** CITATIONS

SEE PROFILE



Geo-parsing and Analysis of Road Traffic Crash Incidents for Data-Driven Emergency Response Planning

Patricia Ojonoka Idakwo, Olubayo Adekanmbi, Anthony Soronnadi, Amos David

PII: S2405-8440(24)17098-3

DOI: https://doi.org/10.1016/j.heliyon.2024.e41067

Reference: HLY 41067

To appear in: *HELIYON*

Received Date: 10 June 2024

Revised Date: 5 November 2024

Accepted Date: 6 December 2024

Please cite this article as: P.O. Idakwo, O. Adekanmbi, A. Soronnadi, D. Amos, Geo-parsing and Analysis of Road Traffic Crash Incidents for Data-Driven Emergency Response Planning, *HELIYON*, https://doi.org/10.1016/j.heliyon.2024.e41067.

Geo-parsing and Analysis of Road Traffic Crash Incidents for Data-Driven Emergency Response Planning

Patricia Ojonoka Idakwo[a, b]*, Olubayo Adekanmbi[b], Anthony Soronnadi[b], Amos David[a]

[a]Department of Computer Science, African University of Science and Technology Abuja, Nigeria

[b] Data Science Nigeria, AI Hub, 33 Queens Street, Alagomeji, Yaba, Lagos, Nigeria

Abstract

Road traffic crashes (RTCs) are a major public health concern worldwide, particularly in Nigeria, where road transport is the most common mode of transportation. This study presents the geo-parsing approach for geographic information extraction (IE) of RTC incidents from news articles. We developed two custom, spaCy-based, RTC domain-specific named entity recognition (NER) models: RTC-NER Baseline and RTC-NER. These models were trained on a dataset of Nigerian RTC news articles. Evaluation of the models' performances shows that the RTC-NER model outperforms the RTC-NER Baseline model on both Nigerian and international test data across all three standard metrics of precision, recall and F1-score. The RTC-NER model exhibits precision, recall and F1-score values of 93.63, 93.61 and 93.62, respectively, on the Nigerian test data, and 91.9, 87.88 and 89.84, respectively, on the international test data, thus showing its versatility in IE from RTC reports irrespective of country. We further applied the RTC-NER model for feature extraction using geo-parsing techniques to extract RTC location details and retrieve corresponding geographical coordinates, creating a structured Nigeria RTC dataset for exploratory data analysis. Our study showcases the use of the RTC-NER model in IE from RTC-related reports for analysis aimed at identifying RTC risk areas for data-driven emergency response planning.

1. INTRODUCTION

Road traffic crashes (RTCs) are a development-related challenge and a major public health concern. With approximately 1.35 million deaths occurring globally in 2018, injuries from RTCs represented the eighth leading cause of death among people of all ages and the leading cause of death among children and young adults, aged between 5 and 29 years [1]. The number of deaths due to RTC injuries among young adults in low- and middle-income countries has been increasing over the years, with around 44% of these deaths occurring in lower middle-income countries such as Nigeria [2], [3]. For countries to meet the United Nations' Decade of Action for Road Safety 2021–2030 target of halving deaths from RTC injuries by 2030 through timely post-crash care [4], it is imperative that emerging technologies such as machine learning be applied to RTC data for data-driven emergency response planning.

News articles have become a popular data source for gaining insights into societal concerns and trends such as RTCs, existing as unstructured data both in print and online news articles. This raw data is usually presented in a well-structured format devoid of ambiguous or imprecise toponyms, jargons, abbreviations, improper sentence structures, misspelled words, mixed languages and grammatical errors, all of which are common drawbacks of social media text [5], [6], [7]. These data need to be transformed into computational representations through natural language processing (NLP) [8] and text mining techniques prior to information extraction (IE) into a structured format for analysis [9]. This study uses RTC reports from news articles for extraction of the geographical and human impact (casualty and injured figures) information for analytics.

There exist two approaches for geographical IE from text: the named entity recognition (NER)-based approach and the gazetteer-based approach, with the former being more suited for the well-structured written English language used in news articles [6]. NER is an NLP technique vital for IE, and it involves identification and categorisation of entities or named items such as persons, organisations and places in text [10]. IE through NER is a cost- and time-effective method and is preferred to the manual IE method of reading and input, which is characterised by time wastage and high cost and resource requirements. The NER based approach to geographic IE is implemented through a technique called geo-parsing [11].

Geo-parsing is extraction of toponyms (place names or location entities) from text and linking them to corresponding geographic coordinates in two successive steps: toponym recognition and toponym resolution [12], [13], [14]. Toponym recognition (location entity recognition) is a subset of NER and involves identification of toponyms in text; toponym resolution (geo-coding) identifies the corresponding geographic coordinates (latitude and longitude) for the toponyms identified [12], [15]. In this study, we develop RTC domain-specific NER models, RTC-NER Baseline and RTC-NER, for toponym recognition and compare their performances in recognising toponyms in RTC-related news articles.

Through implementation of this integrated approach to geo-parsing by using the better performing model, RTC-NER, on raw RTC reports collected, we curate a structured dataset for analysis. The dataset's features include the toponyms extracted - road, landmark, suburb, town/village/community, local government area (LGA), state and hospital; human impact numbers - casualty and injured; a single location feature called 'rtc_site' created by combining the identified toponyms through a rule-based selection algorithm; and coordinates (latitude and longitude) of the 'rtc_site' feature. Researchers employed this dataset for exploratory data analysis to comprehend road- and town-level RTC patterns in our dataset.

The contribution of this study entails the integrated methodology to IE, the creation of RTC model training and testing datasets, the RTC-NER models and the structured RTC dataset for analysis. Importantly, this study is expected to be of immense value to road transportation management organisations such as the Federal Road Safety Corps, Nigeria (FRSC), health care systems and policy makers. Road transportation management organisations can use the RTC-NER models to extract information from their RTC reports, and then they can analyse the extracted data for insights on risk areas (roads, towns) and hotspots that have high prevalence of RTC incidents. Consequently, these organisations can allocate adequate resources in their emergency response plans for reducing the fatality and injured figures. The analysis results of this study are expected to enable health care systems, particularly those along the Lagos–Ibadan

Expressway, to plan for hospital-based emergency medical service (EMS) delivery. Simultaneously, they might help policy makers at all levels to make policy decisions on road safety, particularly with respect to allocation of EMS resources for health coverage to RTC victims.

2. RELATED WORK

Unstructured RTC textual data from social media platforms (such as Twitter and Sina Weibo), RTC case files and news articles have proven to be valuable sources of IE for quantitative, predictive and geo-spatial RTC analysis.

Studies on extraction of RTC information, including location data from social media platforms and case files, used a combination of NLP and deep learning techniques [16], [17], [18] as well as NER and ontology [19]. Muguro et al [16] extracted data from the National Transport and Safety Authority's reports and Twitter for quantitative analysis so as to gain insights into traffic safety, cultures and practices in Kenya. For this, they used the NLP approach of topic modelling and sentiment analysis through n-gram search of keywords to classify the data into eight topics: policing, public service vehicles, robbery, infrastructure, traffic, accident, recklessness and corruption [16]. NLP and convolutional neural network long short-term memory (CNN-LSTM), which is a hybrid deep learning model, were employed by Chang et al [17] to detect RTC-related posts in the Chinese social media platform Sina Weibo for extraction of RTC and traffic-congestion locations. Their aim was to determine accident- and congestion-prone areas for spatio–temporal and semantic analysis for realisation of optimised resource allocation and mitigation measures in traffic management [17]. Addressing the issue of sparse RTC case data and poor recognition of long-text entities, Cheng et al [18] developed a NER model based on entity data enhancement (EDE)–Enhanced Representation through kNowledge IntEgration (ERNIE)–Bidirectional Gated Recurrent Unit Network (BiGRU)–conditional random field (CRF) to extract RTC information, such as time of RTC, name of person involved, plate number, type of number plate and address where accident occurred, from the text data of a real RTC case in a domestic place. Rakhmawati et al [19] used NER to extract RTC information in the Indonesian language, such as incident location, actor (such as car, truck and motorcycle), cause and time from Twitter; then domain ontology to categorize the causes of RTC.

RTC IE from news was performed through techniques such as text mining [9], NER [20] and deep learning [21]. Yang et al [9] applied a text mining technique referred to as term frequency–inverse document frequency (TF–IDF) on responder-involved RTC event news reports to comprehend the characteristics of first responder-involved incidents. NER, semantic role labelling and regular expression were employed by Pahi et al [20] to extract RTC information, such as death information (verbs), death count, injury information (verbs), injury count, location, number plate, dates and day, from news articles to populate a road accident database system. Ling et al [21] proposed the deep learning method SoftLexicon-BiLSTM-CRF for extraction of RTC information, such as time, location, subject, results and causes, from online Chinese news both on social media (Sina Weibo) and news platforms (Toutiao, CCTV News, etc.).

The approach to geo-parsing RTC incidents from social media [22] and news articles [23]has been sparsely studied in the literature. Suat-Rojas et al [22] used the spaCy Python library for extraction of location (places and addresses) and time (such as 'ayer/yesterday' or 'esta mañana/this morning') from Twitter Spanish posts and geo-coded the locations extracted. With regard to geo-parsing of RTC incidents from news article as conducted in our study, Shivakoti [5] performed geo-parsing on Norwegian news articles from Accident Investigation Board Norway and Google News. For that, they used Stanford CoreNLP toolkit with its Parts-of-Speech tagger and NER components for toponym recognition and Google Geocoding API for geo-coding, following which they visualised the results on a map [5].

Our study deviates from the commonly applied IE tools and techniques in the literature by using the spaCy library to develop custom NER models for extraction of location (road, landmark, suburb, town, LGA, state and hospital) and human impact (casualty and injured) details from RTC news and web blog articles so as to curate a structured Nigerian RTC dataset. The objective is to perform exploratory data analysis and gain insights into the road- and town-level patterns in our RTC dataset for data-driven emergency response planning.

3. METHDOLOGY

This study was performed in four main stages: data collection, RTC-NER model development, geo-parsing (feature extraction using the RTC-NER model, 'rtc_site' selection algorithm and geo-coding of 'rtc_site') and exploratory data analysis, as shown in Figure 1.

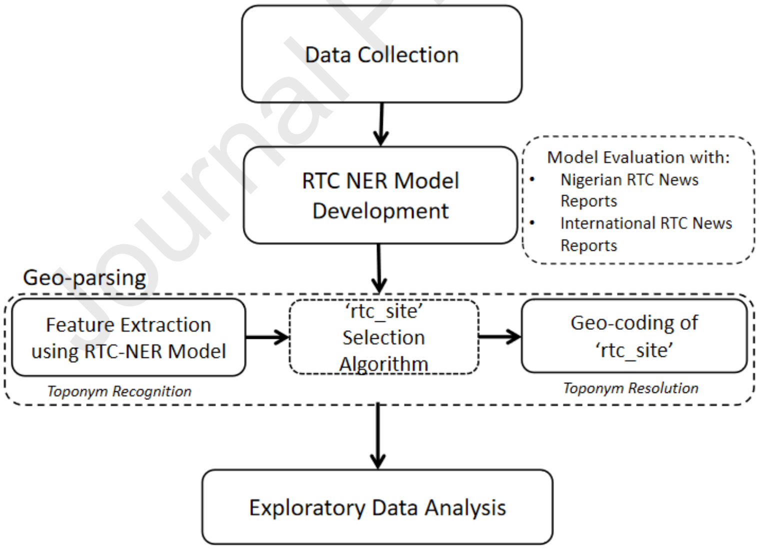

Data Collection
RTC NER Model Development
Model Evaluation with:
• Nigerian RTC News Reports
• International RTC News Reports
Geo-parsing
Feature Extraction using RTC-NER Model
'rtc_site' Selection Algorithm
Geo-coding of 'rtc_site'
Toponym Recognition
Toponym Resolution
Exploratory Data Analysis

Figure1: Methodological Framework

3.1. Data Collection

The data for this study were collected in three steps:

Step 1

RTC reports were scrapped from Nigerian news articles by using the search term- 'road accidents + crash' on the following Nigerian online print media platforms: DailyPost, the Sun and the Punch. The query yielded 856 news articles published between October 2, 2015 and

October 9, 2023. This dataset was used for the RTC-NER model's development and evaluation phase.

Step 2

RTC reports were collected from the following:

- Nairaland (https://www.nairaland.com/), a high-traffic Nigerian online discussion forum known for its high user engagement [24]. The search terms 'road accident + crash' and 'road accident + road traffic crash' yielded 198 posts published between February 6th, 2021 and June 17th, 2024. This dataset was vital to our study in reducing the effect of biases associated with data from news reports [9].
- DailyPost: The search term 'road accident + crash' was used on the Nigerian online print media platform DailyPost to scrap additional 110 RTC news reports published between July 24th, 2015 and July 1st, 2024.

Step 3

A total of 70 RTC reports from 31 countries were collected from Google News: Bolivia(1), Brazil(2), Cameroun(2), Canada(1), China(3), Egypt(3), France(2), Germany(1), Ghana(6), Hungary(1), India(7), Israel(1), Italy(1), Ivory Coast(1), Kazakhstan(5), Kenya(1), Mali(1),Pakistan(1), Philippines(3), Russia(2), Saudi Arabia(1), Scotland(3), South Africa(4), Spain(1), Sudan(1), Sweden(1), Turkey(1), UAE(1), Uganda(1), UK(7) and USA(4). The dataset was used to evaluate the performance of the models on international RTC news.

3.2. RTC-NER Model Development and Evaluation

In this stage, we first performed data pre-processing (data cleaning and annotation) prior to custom NER model development and evaluation.

3.2.1. Data Pre-processing

The dataset from step 1 in sub-section 3.1: Data Collection was cleaned by removing non-ASCII characters, extra spaces, quotation marks and question marks. For increased detail and precision, further data pre-processing was performed in a two-phased approach, as shown in Figure 2.

In Phase 1, a smaller and manageable dataset of 212 news articles was randomly selected from the RTC corpus. For annotation, the dataset was split into training (171) and test (41) data using the 80–20 rule.

Phase 2 entailed scaling to the entire RTC corpus of 856 news articles. Sentence tokenisation was then performed using a Python NLP library to split the entire corpus into 9584 sentences. In view of the sentence structure of Nigerian RTC news reporting, sentences that did not have words such as 'road', 'highway', 'expressway', 'village', 'town', 'community', 'local government area', 'lga', 'state', 'center', 'centre', 'hospital' or 'clinic' were filtered out, leaving 4218 sentences. The dataset was then split into training (3374) and test (844) data using the 80–20 rule for annotation.

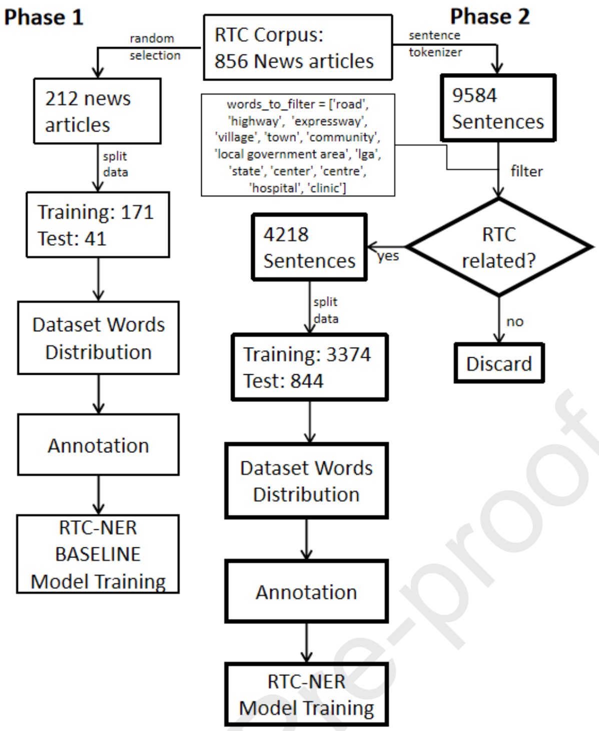


Figure 2: Two-Phased Approach to Data Pre-processing and Model Training

Dataset Word Distribution

To achieve robust and efficient NER and de-bias it to avoid shortcut learning in the RTC-NER models [15], the distribution of words that translated to entities in RTC-NER was examined. Words that represented roads, places and hospitals were accordingly grouped for the training and test datasets in each iteration, as follows: Roads [Road, Highway, Expressway]; Place [Village, Town, Community, LGA, State]; Hospital [center, centre, hospital, clinic]. The percentage distribution for RTC-related word occurrences in Table 1 showed an almost equal distribution of word groups across the training and test datasets in both iterations, with training having ~80% and test having ~20% as per the 80–20 rule for splitting data into training and test data. The balanced performance of both RTC-NER models was hence assured.

Table 1: Datasets Word Distribution

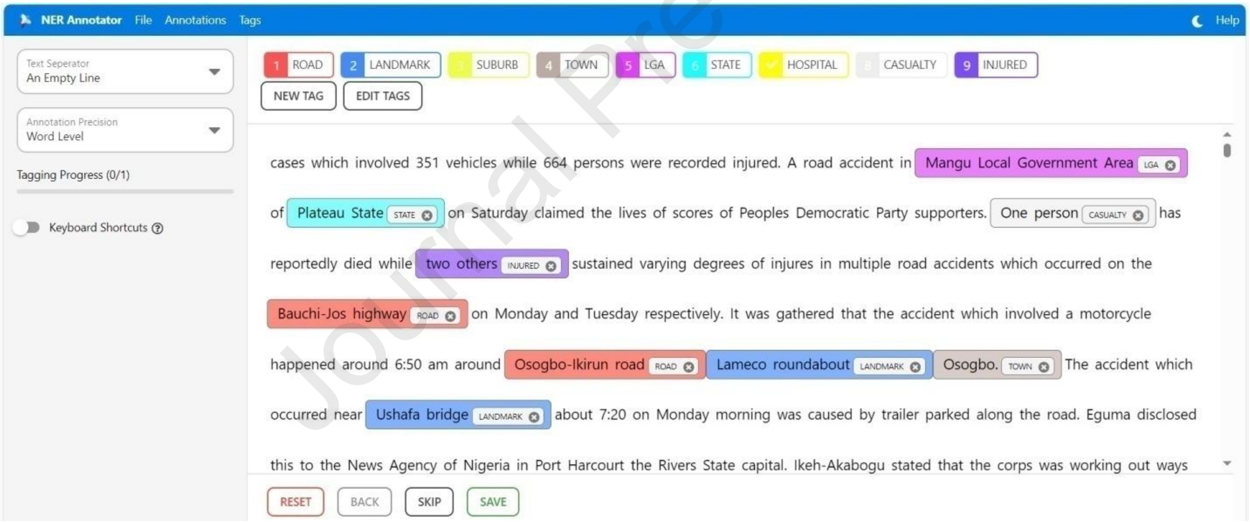
	Dataset	**Roads**	**Place**	**Hospital**	**% Roads**	**% Place**	**% Hospital**
Phase 1	Training	171	153	119	81%	81%	80%
	Test	41	37	29	19%	19%	20%
Phase 2	Training	2085	1579	622	80%	81%	80%
	Test	512	382	154	20%	19%	20%

Corpus Annotation

'NER-Annotator', a user-friendly web interface for manual annotation of entities for spaCy model training, was used [25]. We defined a set of custom tags/labels of relevance to RTC incidents, as presented in Table 2. The training and test data were converted into individual.txt files as input, while JSON files were generated as output of the NER-Annotator. A screenshot of the NER-annotator web interface is shown in Figure 3.

Table 2: Custom Tags used for our Road Traffic Crash (RTC) Corpus

Tag Name	Description	Example
Road	Name of road where RTC event occurred	Lagos–Ibadan Expressway
Landmark	Place names with spatial relation to RTC event location	near Foursquare camp
Suburb	Settlement in a town where RTC event occurred	Ogere
Town	Name of the town of RTC event or where suburb is l ocated	Ajebo
LGA	Local government area where town is located	Remo North
State	State where LGA/Town are located	Ogun State
Hospital	Hospital name where RTC victims were taken to	Victory Hospital, Ogere
Casualty	Number of people who died due to RTC event	5 deaths
Injured	Number of people injured in RTC event	3 injured



Text Seperator: An Empty Line

Annotation Precision: Word Level

Tagging Progress (0/1)

cases which involved 351 vehicles while 664 persons were recorded injured. A road accident in Mangu Local Government Area LGA of Plateau State STATE on Saturday claimed the lives of scores of Peoples Democratic Party supporters. One person CASUALTY has reportedly died while two others INJURED sustained varying degrees of injures in multiple road accidents which occurred on the Bauchi-Jos highway ROAD on Monday and Tuesday respectively. It was gathered that the accident which involved a motorcycle happened around 6:50 am around Osogbo-Ikirun road ROAD Lameco roundabout LANDMARK Osogbo. TOWN The accident which occurred near Ushafa bridge LANDMARK about 7:20 on Monday morning was caused by trailer parked along the road. Eguma disclosed this to the News Agency of Nigeria in Port Harcourt the Rivers State capital. Ikeh-Akabogu stated that the corps was working out ways

Figure 3: Interface of SpaCy Annotator for Named Entity Recognition (NER)

3.2.2. NER Training with SpaCy 3.61

SpaCy 3.61 is a Python-based open-source library for high-level NLP and has several multi-lingual, pre-trained and transformer-integrated models, such as ‘en_core_web_sm’, which can identify up to 18 entities, ranging from people, dates, city to organisations [22], [25], [26], [27], [28], [29]. In addition, spaCy provides features such as the entity recogniser NER pipeline component for development of custom NER models on domain-specific entities, as its pre-trained models fail to identify such entities in text [30], [31]. The selection of SpaCy 3.61 for the transformer-based NER model training was, therefore, based on its ease of use, cost-effectiveness, improved accuracy, flexibility and efficiency, enabling us to focus on data preparation and model development.

The training pipeline entailed use of the blank SpaCy NER model with our annotated training datasets so as to develop the RTC-NER Baseline and RTC-NER models. The RTC-NER Baseline model was trained in 4200 steps (epochs) for 6 h, while the RTC-NER model was trained in 4600 steps (epochs) for 11 h. Both models were trained on a single A100 GPU from Google Colab using the 'spacy.TransitionBasedParser.v2' architecture for NER: 'tok2vec' and 'ner' pipelines; 'spacy.Tokenizer.v1' tokenisers; 'Adam.v1' optimiser; the batch size was 1000, dropout rate 0.1, learning rate 0.001 and evaluation frequency 200.

3.3. **Geo-parsing**

At this stage of the study, actual geo-parsing was used to develop a structured RTC dataset for analysis. The geo-parsing stage entailed toponym recognition (using the RTC-NER model, which performed better, to extract location features). The output of the toponym recognition was used to create a feature referred to as 'rtc-site', while toponym resolution (identifying corresponding latitude and longitude) of 'rtc-_site' was performed using the Google Geocoding API.

3.3.1. **Feature Extraction (Toponym Recognition)**

The datasets from steps 1 and 2 of sub-section 3.1: Data Collection were merged to obtain1164 RTC reports, which were cleaned by removing duplicates, irrelevant data (e.g. non-road crashes) and reports with multiple incidents, hence resulting in a cleaned dataset of 965 usable RTC records for this phase of the study.

The RTC-NER model, which performed better than the RTC-Baseline model, was used to identify incident details, such as road, landmark, suburb, town/village, LGA, state, hospital and number of dead and injured, for each RTC report in the dataset. Several data pre-processing steps were performed to prepare the data for analysis, such as text normalisation, data conversion, stop word removal and standardisation of road names for consistency. These data transformation procedures resulted in a clean and structured dataset for further feature extraction.

3.3.2. **'rtc_site' Selection Algorithm**

A prominent characteristic of our dataset is that each RTC news report has more than one location feature, usually in non-uniform combinations. For instance, an RTC report may only include the name of the road, landmark, town and hospital; while another may only have the road, suburb, town, and LGA. As such, it was imperative to create a single location feature that would serve as the unique place name for each RTC event. Accordingly, a rule-based feature selection algorithm was designed to create the 'rtc_site' feature. This feature was constructed by combining extracted entities on the basis of specific criteria, as shown in Figure 4.

As per [11], points at the centres of towns, cities, villages or polygon/line geo-spatial features are the usual output of geo-parsing toponyms from text, resulting in a distance offset between actual event locations and geo-parsed ones. In this study, distance offsets can notably reduce the accuracy and efficiency of the RTC-NER models by introducing noise into the RTC location data, making them unreliable for further spatial analysis such as identification of RTC hotspots. To reduce the effect of distance offset, we implemented the 'rtc_site' selection algorithm, which obeyed the following order of preference for 'rct_site': 'LANDMARK', 'SUBURB', 'TOWN', 'HOSPITAL', 'ROAD + TOWN'.

3.3.3. Geo-coding

Geo-coding was performed on the ‘rtc_site’ feature using the Google Geocoding API to obtain latitude and longitude coordinates [32]. These spatial coordinates were then used for geo-spatial analysis.

The final transformed dataset had 965 rows across 12 features (‘road’, ‘landmark’, ‘suburb’, ‘town’, ‘lga’, ‘state’, ‘hospital’, ‘casualty’, ‘injured’, ‘rtc_site’, ‘rtc_lat, ‘rtc_long’). The number of records with identifiable values for each feature in the dataset was as follows: road(839), landmark(396), suburb(107), town(802), lga(271), state(921), hospital(965), casualty(837), injured(619), rtc_site(965), rtc_lat(965), rtc_long(965).

3.4. Exploratory Data Analysis

Aggregate analysis of the curated dataset of 965 RTC incidents showed a total of 4122 dead and 4451 injured persons due to RTCs along 378 unique roads, with casualty and injured figures in individual incidents ranging from 0 to 30 deaths and 0 to 83 injuries, respectively. The average number of fatalities and injuries per incident was ~5 and ~7, respectively, while the proportion of RTCs with injuries and proportion of RTCs with fatalities was 64% and 87%, respectively. This indicated a high severity rate of the RTC incidents in our dataset.

```
# Rule-based rtc_site Selection Algorithm
# For each row in our dataset,
L = row['landmark']
S = row['suburb']
T = row['town']
LG = row['lga']
ST = row[‘state’]
R = row['road']
H = row['hospital']
rtc_site = “”

# |X|>0: denotes  X is not empty,
# |X|=0: denotes  X is empty

function create_rtc_site(row):
    if |L|>0 && |S|>0 && |T|>0:
         rtc_site = L + ", " + S + ", " + T
    elif |L|>0 && |T|>0 && |ST|>0:
         rtc_site = L + ", " + T “, ” + ST
    elif L|>0 && |T|>0: rtc_site = L + ", " + T
    elif |L|>0 && |ST|>0: rtc_site = L + ", ” + ST
    elif |L|=0 && |S|>0 && |T|>0 && |ST|>0:
         rtc_site = S + ", " + T + ", " + ST
    elif |L|=0 && |S|=0 && |T|>0 && |LG|>0 && |ST|>0:
         rtc_site = T + "," + LG + "," ST
    elif |L|=0 && |S|=0 && |T|=0 && |LG|=0 && |H|>0: rtc_site = H
    elif |L|=0 && |S|=0 && |T|>0 && |LG|=0 && |H|=0 && |R|>0:
         rtc_site = R + “, ‟ + T
    else:
         rtc_site = "Unknown"
return rtc_site
```

Figure 4: ‘rtc_site’ Feature Selection Algorithm

Visualisation of RTC incident sites in Figure 5 shows the extent of RTCs across the country, with a high concentration around the south western states of Ogun, Lagos and Oyo and Osun. The categorisation of states by number of RTC incidents in Figure 6 shows that Ogun falls in the very-high category with an RTC incident count of 242. The high category has states such as Bauchi (76), Anambra (74) and Osun (65). Other states such as Ondo (42), Lagos (42), Kwara(34), Jigawa(30), Oyo(29), Kogi(28), Federal Capital Territory–FCT(28), Kano(27), Niger(26), Delta(26) and Kaduna (22) fall in the moderate category.

3.4.1. Road-level Analysis

Initial analysis focused on road-level patterns to identify high-risk areas. The statistics (number of RTC incidents, total casualty, total injured and proportion of RTC with fatalities) of the top ten roads with highest number of RTC incidents are presented in Table 3, with the results discussed in sub-section 4.2.1.

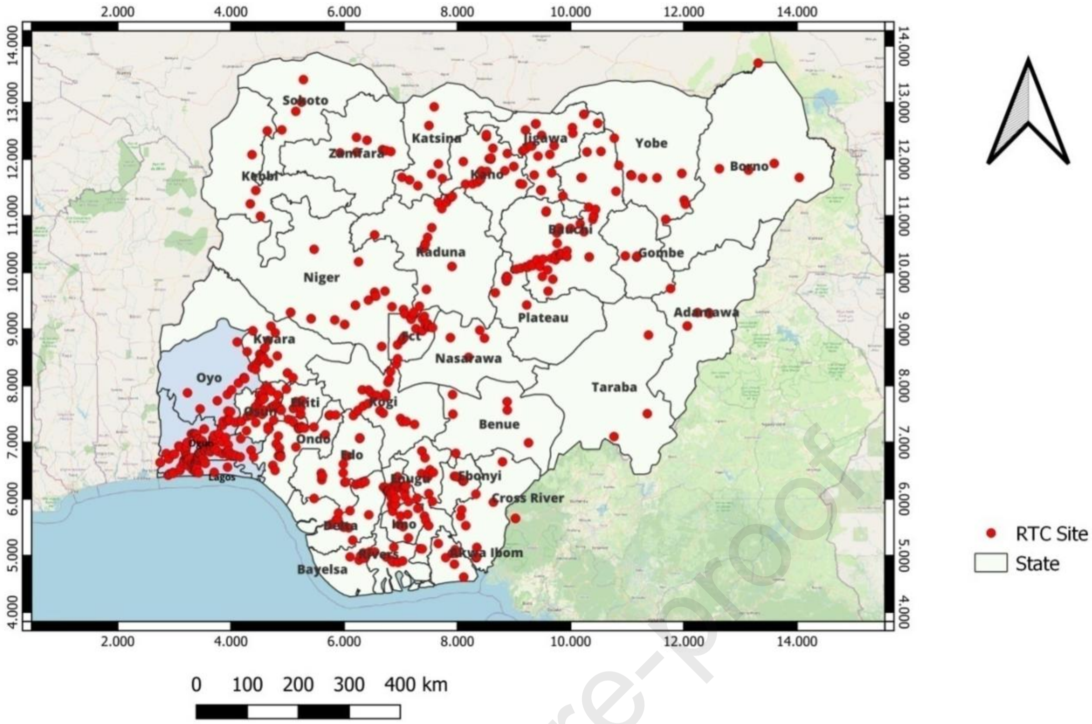


Figure 5: Road Traffic Crash (RTC) Sites in Our Dataset

The Lagos–Ibadan Expressway, a 127.6-km-long road linking Lagos and Ibadan [33], traverses the three states Lagos, Ogun and Oyo, as shown in Figure 7. Precisely, it traverses the following 14 LGAs across these states: Ikeja and Kosofe in Lagos state; Ifo, Ikenne, Shagamu, Remo North and Obafemi Owode in Ogun state; and Oluyole, Ibadan South East, Ibadan North East, Ibadan North, Lagelu, Akinyele and Egbeda in Oyo state. The RTC sites and categorisation of RTC incident counts in these LGAs are shown in Figures 8 and 9, respectively. The expressway, which is considered the busiest interstate route in Nigeria and connects the south west region to other parts of the country [34], recorded 132 fatal RTC incidents resulting in 446 deaths and 642 injuries between 2015 and 2024.

3.4.2. **Town-level Analysis**

We performed analysis at town-level to identify RTC hotspots along the expressway. The statistics (number of RTC incidents, total casualty, total injured and proportion of RTC with fatalities) of the top ten towns with highest number of RTC incidents (all in Ogun state) are presented in Table 4, with the results discussed in sub-section 4.2.2. We may remember that ‘town’ in our dataset refers to names of villages, towns or communities where an RTC incident occurred.

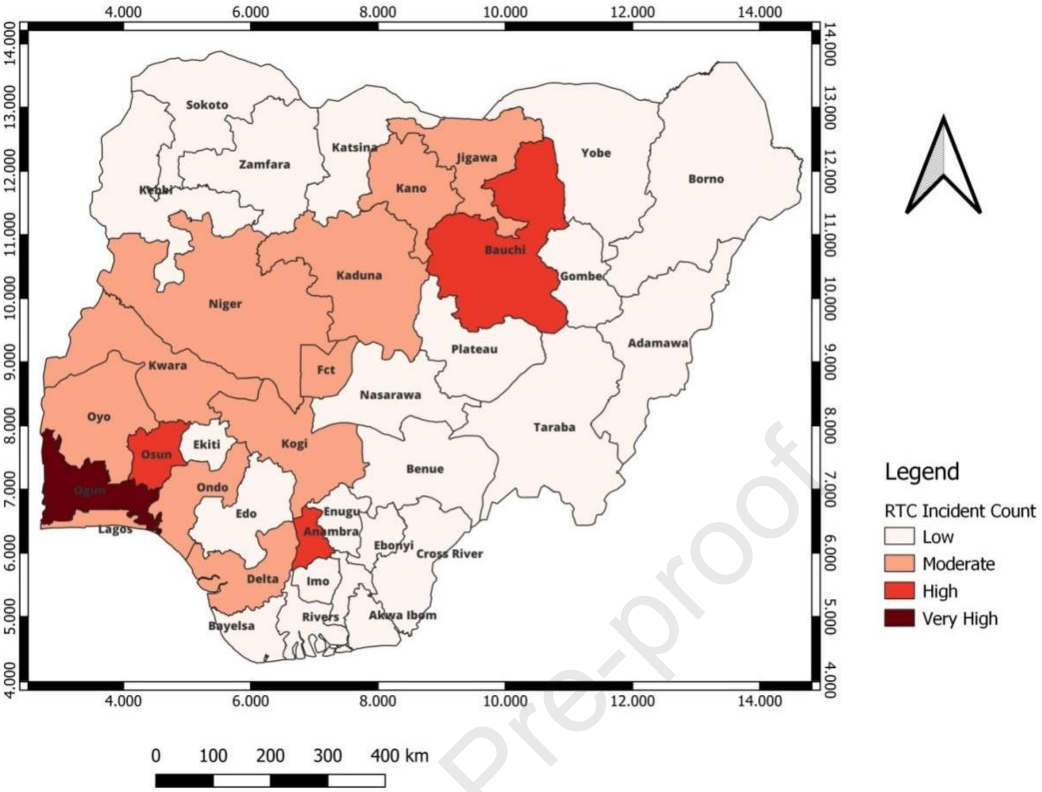


Figure 6: Distribution of RTC Incidents by State

Table 3: Top Ten Roads with Highest Number of RTC Incidents (2015 – 2024)

Road	Number of RTC Incidents	Total Fatalities	Total Injured	Proportion of RTCs with fatalities	Proportion of RTCs with injured
Lagos–Ibadan Expressway	132	446	642	82%	73%
Lagos–Abeokuta Expressway	23	53	86	83%	83%
Abuja–Lokoja Expressway	15	99	57	80%	47%
Bauchi–Jos Highway	15	46	42	87%	53%
Abeokuta–Sagamu Expressway	14	28	53	86%	79%
Abuja–Kaduna Expressway	13	115	218	92%	69%
Enugu–Onitsha Expressway	11	18	32	64%	55%
Bauchi–Kano Highway	10	56	79	100%	90%
Kano–Zaria Expressway	10	90	69	100%	60%
East–West Road	9	24	27	78%	56%

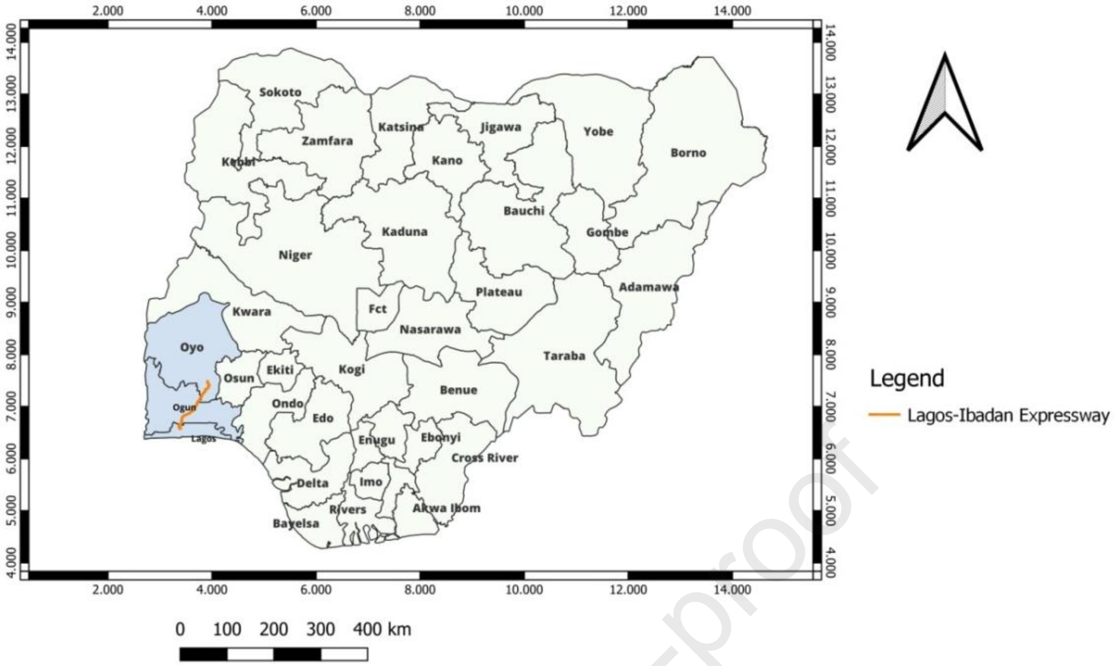


Figure 7: Lagos–Ibadan Expressway Traversing Three States

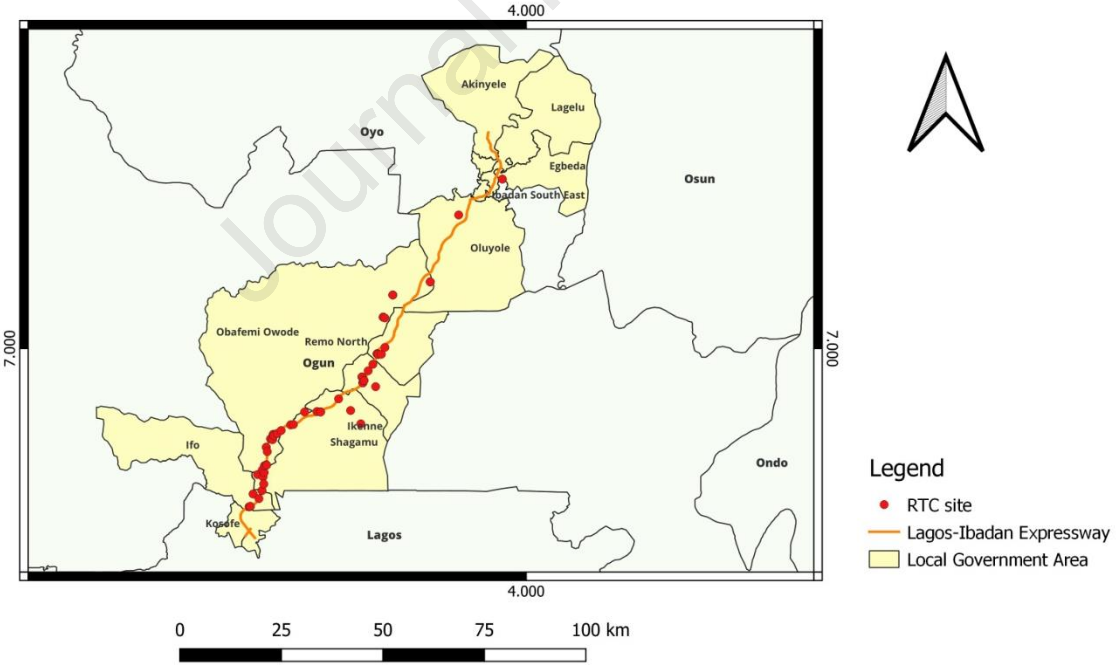


Figure 8: RTC Sites and Local Government Areas (LGAs) along the Lagos–Ibadan Expressway

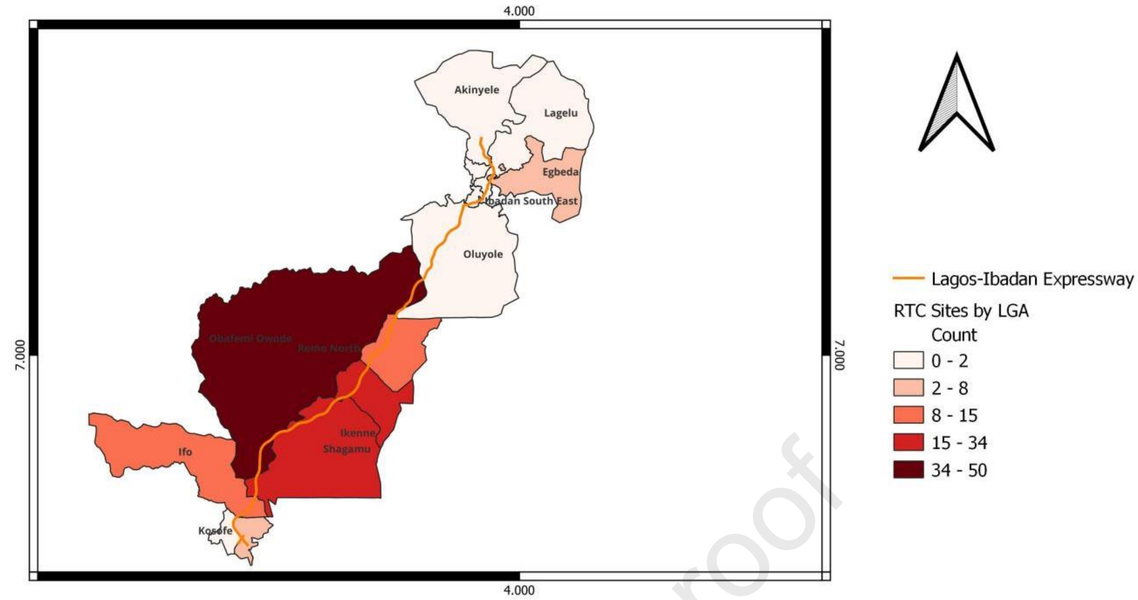


Figure 9: Categorisation of LGAs along the Lagos–Ibadan Expressway by Number of RTC Incidents

Table 4: Top Ten Towns along the Lagos–Ibadan Expressway with Highest Number of RTC Incidents (2015-2024)

Town	Number of RTC Incidents	Total Fatalities	Total Injured	Proportion of RTCs with fatalities	Proportion of RTCs with injured
Ogere	23	78	62	96%	65%
Kara	13	33	80	85%	92%
Ibafo	12	12	51	50%	75%
Mowe	12	33	75	67%	50%
Ogunmakin	10	33	61	80%	100%
Ishara	10	40	50	80%	60%
Isoku	7	41	35	100%	71%
Sagamu	7	19	21	100%	71%
Ajebo	5	15	29	100%	60%
Fidiwo	5	31	22	100%	60%

4. RESULTS AND DISCUSSION

The evaluation results of the models on Nigerian RTC test data and International RTC test data across several performance metrics, as well as exploratory data analysis results, are discussed.

4.1. Performance Evaluation of the Models

Performances of the custom RTC-NER models were evaluated using standard metrics of precision, F1-score and recall [25], [27]. The Nigerian test data used for evaluation comprised 44 RTC reports for the RTC-NER Baseline model and 844 sentences for the RTC-NER model, while the international test data comprised 70 RTC reports for both models.

Table 5 demonstrates the superior performance of the RTC-NER model versus the baseline across all three-evaluation metrics on both the Nigerian and international RTC test data. Notably, the model achieved precision rates of 93.81% (Nigeria) and 91.9% (international), indicating a low rate of false positives (incorrectly identifying non-entities). Recall values of 94.05% (Nigeria) and 87.88% (international) highlighted the model's ability to accurately identify true entities and thus minimising false negatives. The F1-score, a harmonic mean of precision and recall, further confirmed the model's overall effectiveness.

Table 5: Comparison Between the RTC-NER Baseline and RTC-NER Models

		Precision	**Recall**	**F1-Score**	**No of samples**
Nigerian RTC Reports	RTC-NER Baseline	92.37	90.15	91.25	44
	RTC-NER	**93.81**	**94.05**	**93.93**	844
International RTC Reports	RTC-NER Baseline	89.2	63.66	74.3	70
	RTC-NER	**91.9**	**87.88**	**89.84**	70

Abbreviation: NER = named entity recognition; RTC= road traffic crash

The performances of the models at the entity level on both the Nigerian and international test data as shown in Tables 6 and 7, respectively, confirmed that the RTC-NER model, with a larger training data, delivered a better overall performance versus the RTC-NER Baseline model. On the Nigerian test data, all RTC-NER entities performed better than RTC-NER Baseline entities, with ROADS having the highest F1-score of 97.75 and LANDMARK the lowest F1-score of 89.08.

Table 6: Entity-Level Performance Comparison Between the RTC-NER Baseline and RTC-NER Models on Nigerian RTC Reports

	RTC_NER_BASELINE			**RTC_NER**		
	Precision	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**
ROAD	96.08	93.78	94.92	97.63	97.86	**97.75**
LGA	88.06	81.94	84.89	95.45	96.92	**96.18**
STATE	83.8	76.77	80.13	94.7	95.03	**94.86**
HOSPITAL	91.97	87.5	89.68	95.25	95.93	**95.59**
TOWN	90.59	86.03	88.25	91.31	92.95	**92.12**
LANDMARK	87.21	85.23	86.21	92.73	85.71	**89.08**
INJURED	-	-	-	90.57	91.70	**91.13**
CASUALTY	-	-	-	90.34	90.19	**90.27**
SUBURB	-	-	-	86.84	91.67	**89.19**

Regarding recall and precision, Tables 6 and 7 show that among entities, LANDMARK achieved the lowest recalls of 85.71(Nigerian) and 76.4(international) and SUBURB earned the lowest precision ratings of 86.84(Nigerian) and 84.62(international). Thus, the likelihood of the RTC-NER model failing to identify actual 'LANDMARK' entities was highest while that of identifying non-entities as 'SUBURB' entities was highest. This was probably because the entities 'LANDMARK' and 'SUBURB' appeared the least number of times in the training datasets, as landmarks are not often used to describe RTC incident locations in news articles and RTC incidents in suburbs are not often reported in news articles owing to place bias.

The robust performance of the RTC-NER model on the international test data underscored its applicability to RTC reports from diverse regions, suggesting its potential for broader use in the field.

Table 7: Entity-Level Performance Comparison Between the RTC-NER Baseline and RTC-NER Models on International RTC Reports

	RTC_NER_BASELINE			**RTC_NER**		
	Precision	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**
ROAD	94.26	81.07	87.17	97.34	94.6	95.95
LGA	92.19	73.75	81.94	95.09	91.3	93.16
STATE	86.67	60.94	71.56	94.54	87.24	90.75
HOSPITAL	92.65	83.44	87.8	94.61	94.12	94.36
TOWN	85.88	67.26	75.43	89.14	85.95	87.52
LANDMARK	80.85	66.09	72.73	89.08	76.4	82.26
INJURED	-	-	-	87.82	82.64	85.15
CASUALTY	-	-	-	84.63	82.33	83.46
SUBURB	-	-	-	84.62	76.74	80.49

4.2. Data Analysis

The analysis result of the curated dataset of 965 RTC incidents done in sub-section 3.4 vis-à-vis road and town levels is now discussed.

4.2.1. Road-level Analysis

Of the top ten roads with highest number of RTC incidents, the Lagos–Ibadan Expressway placed first with 132 RTC incidents resulting in 446 dead and 642 injured between 2015 and 2024. However, as shown in Table 3, the proportion of RTCs with fatalities on the expressway was 82% which is less than that of roads with considerably less number of RTC incidents, such as the Kano–Zaria Expressway, Bauchi–Kano Highway, Abuja–Kaduna Expressway, Bauchi–Jos Highway and Abeokuta–Sagamu Expressway. The Kano–Zaria Expressway and Bauchi–Kano Highway, both with 10 RTC incidents, had the highest proportion of RTC with fatalities of 100%, while the Abuja–Kaduna Expressway (13 RTC incidents), Bauchi–Jos Highway (15) and Abeokuta–Sagamu Expressway (14) had their proportion of RTCs with fatalities of 92%, 87% and 86%, respectively. A study on the causes of RTC incidents on these roads will shed more light into this worrying trend.

4.2.2. Town-level Analysis

As shown in Table 4, the relationship between number of RTC incidents and proportion of RTC incidents with fatalities was not linear for the towns, as the towns with the highest proportion of RTC with fatalities of 100% were Isoku, Sagamu, Ajebo and Fidiwo, which had the least number of RTCs. Ogere, with the highest number of RTC incidents, had the proportion of RTC with fatalities of 96%, which was high. Grouping the towns in the table by their LGA yielded the following: Ikenne LGA (Ogere), Ifo LGA (Isheri), Obafemi-Owode LGA (Ibafo, Mowe, Ogunmakin, Isoku, Ajebo and Fidiwo), Remo North LGA (Ishara) and Sagamu LGA (Sagamu). Clearly, Obafemi-Owde LGA had the most number of towns. This supports the categorisation of Obafemi-Owode LGA as having the highest number of RTC incidents of all the LGAs on the Lagos–Ibadan Expressway, as shown in Figure 9.

5. LIMITATIONS

The approach to dataset curation from RTC news reports is prone to severity and time biases, as more severe and recent RTC are reported in news articles, which could result in non-representative datasets [9], [35], [36]. Hence, in step 2, more data were collected to reduce the effects of severity bias on our dataset and data collected over a 7 year period between 2015 and 2024 were aimed at time bias reduction.

Offset distance between actual RTC incident location and geo-parsed location is another limitation, which could affect emergency response planning decisions. To overcome this limitation, our 'rtc_site' selection algorithm prioritised entities such as 'LANDMARK' and 'HOSPITAL', which represent points on the earth surface, as well as 'SUBURB' and 'TOWN', both with relatively smaller surface areas such that their centres were considered near to roads of RTC incidents.

6. CONCLUSION AND FUTURE WORK

We developed an extensive methodological framework to geo-parse RTC incident locations from news articles via domain-specific NER models developed using the spaCy Python library. The impressive overall performance of the RTC-NER Baseline and RTC-NER models on the Nigerian test data and international test data demonstrated they could be satisfactorily applied for geo-parsing of RTC reports from various regions of the world. The RTC-NER model, with better performance, was used for feature extraction in the curated, structured RTC dataset made from unstructured text in Nigerian news articles for data analysis so as to identify RTC risk-areas at the road and town levels.

The analysis results offered insights into understanding the severity pattern of RTC incidents along major Nigerian roads as well as towns along the Lagos–Ibadan Expressway for data-driven emergency response planning. The outcomes of this study might be utilised by road transportation management organisations, health care systems, policy makers for emergency response planning (resource allocation and hospital-based EMS delivery) and road safety policymakers for enhancing EMS coverage to RTC victims.

The results also laid a foundation for future work to reduce biases in the dataset by including other sources of data such as social media and official RTC reports, and making the RTC-NER model to perform better on international news through inclusion of international RTC reports in the training data. Furthermore, the RTC-NER model might be optimised for better performance through hyper-parameter tuning and identification of vital entities such as cause of RTC, vehicle types involved, time and date of RTC and categories of persons involved.

CRediT authorship contribution statement

Patricia Ojonoka Idakwo: Conceptualization, Writing – original draft and review editing, Methodology, Investigation, Data curation, Formal analysis, Visualization, Validation. **Anthony Soronnadi:** Validation, Supervision, Writing – review, Project Administration. **Olubayo Adekanmbi:** Supervision, Resources, Writing – review and editing. **Amos David:** Supervision, Writing – review and editing. All authors have read and agreed to the published version of the manuscript.

Data availability

The training and test data as well as the code used for development of the RTC-NER Baseline and RTC-NER Models have been deposited under the MIT License in the Github repository at:https://github.com/PatUnoka/Geoparsing-and-Analysis-of-RTC-Locations-for-Emergency-Response-Planning.git. The RTC data employed for analysis is available on request.

Funding

This work was supported by Data Science Nigeria.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We thank Friday Idakwo Ikwoche from the University of Abuja (Nigeria) for proofreading the manuscript.

1. REFERENCES

[1] W. H. Organization, *Global status report on road safety 2018*. World Health Organization, 2019. [Online]. Available: https://books.google.com/books?hl=en&lr=&id=uHOyDwAAQBAJ&oi=fnd&pg=PR6&dq=World+Health+Organization.+(2019).+Global+status+report+on+road+safety+2018.+World+Health+Organization.&ots=2T-iZyqbT-&sig=Hqxg5tX-0TKlcMVJDmgFcY-Q1CA

[2] S. K. Ahmed *et al.*, “Road traffic accidental injuries and deaths: A neglected global health issue,” *Heal. Sci. Reports*, vol. 6, no. 5, p. e1240, Apr. 2023, doi: 10.1002/hsr2.1240.

[3] O. Awoniyi *et al.*, “Trend analysis on road traffic collision occurrence in Nigeria,” *Disaster Med. Public Health Prep.*, vol. 16, no. 4, pp. 1517–1523, Apr. 2022.

[4] H. E. Rosen *et al.*, “Global road safety 2010–18: an analysis of global status reports,” *Injury*, Apr. 2022, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0020138322005046

[5] D. Shivakoti, “Automatic Detection and Extraction of Event Locations in News Report to locate in Map.,” University of Stavanger, Norway, 2016. [Online]. Available: https://uis.brage.unit.no/uis-xmlui/handle/11250/2413910

[6] A. Kumar and J. P. Singh, “Location reference identification from tweets during emergencies: A deep learning approach,” *Int. J. Disaster Risk Reduct.*, vol. 33, pp. 365–375, Apr. 2019, doi: 10.1016/j.ijdrr.2018.10.021.

[7] F. Ali, A. Ali, M. Imran, R. A. Naqvi, M. H. Siddiqi, and K.-S. Kwak, “Traffic accident detection and condition analysis based on social networking data,” *Accid. Anal. Prev.*, vol. 151, p. 105973, Apr. 2021, doi: 10.1016/j.aap.2021.105973.

[8] M. S. Mredula, N. Dey, M. S. Rahman, I. Mahmud, and Y. Z. Cho, “A Review on the Trends in Event Detection by Analyzing Social Media Platforms’ Data,” *Sensors*, vol. 22, no. 12, pp. 1–41, 2022, doi: 10.3390/s22124531.

[9] C. Yang, J. Liu, X. Li, and T. Barnett, “Analysis of first responder-involved traffic incidents by mining news reports,” *Accid. Anal. Prev.*, vol. 192, no. January, p. 107261, 2023, doi: 10.1016/j.aap.2023.107261.

[10] A. Jolly, V. Pandey, I. Singh, and N. Sharma, “Exploring Biomedical Named Entity Recognition via SciSpaCy and BioBERT Models,” *Open Biomed. Eng. J.*, vol. 18, no. 1, pp. 1–13, 2024, doi: 10.2174/0118741207289680240510045617.

[11] S. Wang *et al.*, “New Era for Geo-Parsing to Obtain Actual Locations: A Novel Toponym Correction Method Based on Remote Sensing Images,” *Remote Sens.*, vol. 14, no. 19, p. 4725, Apr. 2022, [Online]. Available: https://www.mdpi.com/2072-4292/14/19/4725

[12] Z. Liu, K. Janowicz, L. Cai, R. Zhu, G. Mai, and M. Shi, “Geoparsing: Solved or biased? an evaluation of geographic biases in geoparsing,” *Agil. GIScience Ser.*, vol. 3, p. 9, Apr. 2022, [Online]. Available: https://agile-giss.copernicus.org/articles/3/9/2022/

[13] J. Wang and Y. Hu, “Enhancing spatial and textual analysis with EUPEG: An extensible and unified platform for evaluating geoparsers,” *Trans. GIS*, vol. 23, no. 6, pp. 1393–1419, Apr. 2019, doi: 10.1111/tgis.12579.

[14] N. Lytvynenko*, S. Lienkov, O. Lytvynenko, O. Banzak, and H. Banzak, “Development of Geoinformation Technology for Monitoring Events on the Basis of Data from Unstructured Web Resource Text,” *Int. J. Innov. Technol. Explor. Eng.*, vol. 9, no. 5, pp. 1160–1165, 2020, doi: 10.35940/ijitee.e2764.039520.

[15] R. Ma, X. Wang, X. Zhou, Q. Zhang, and X.-J. Huang, “Towards Building More Robust NER datasets: An Empirical Study on NER Dataset Bias from a Dataset Difficulty View,” Apr. 2023, pp. 4616–4630. [Online]. Available: https://aclanthology.org/2023.emnlp-main.281/

[16] J. Muguro, W. Njeri, K. Matsushita, and M. Sasaki, “Road traffic conditions in Kenya: Exploring the policies and traffic cultures from unstructured user-generated data using NLP,” *IATSS Res.*, vol. 46, no. 3, pp. 329–344, 2022, doi: 10.1016/j.iatssr.2022.03.003.

[17] H. Chang, L. Li, J. Huang, Q. Zhang, and K.-S. Chin, “Tracking traffic congestion and accidents using social media data: A case study of Shanghai,” *Accid. Anal. Prev.*, vol. 169, p. 106618, Apr. 2022, doi: 10.1016/j.aap.2022.106618.

[18] Z. Cheng *et al.*, “ERNIE-based Named Entity Recognition Method for Traffic Accident Cases,” *J. Phys. Conf. Ser.*, vol. 2589, no. 1, 2023, doi: 10.1088/1742-6596/2589/1/012020.

[19] N. A. Rakhmawati, Y. Awwab, A. C. Najib, and A. Irsyad, “Ontology-Based Traffic Accident Information Extraction on Twitter In Indonesia,” *Intel. Artif.*, vol. 25, no. 70, pp. 1–12, 2022, doi: 10.4114/intartif.vol25iss70pp1-12.

[20] K. Pahi and A. Shakya, “Road Accident News Information Extraction,” *Icaeic-2019*, vol. 2, no. 1, p. 65, 2019.

[21] Y. Ling, Z. Ma, X. Dong, and X. Weng, "A deep learning approach for robust traffic accident information extraction from online chinese news," *IET Intell. Transp. Syst.*, no. November 2022, pp. 1–16, 2024, doi: 10.1049/itr2.12493.

[22] N. Suat-Rojas, C. Gutierrez-Osorio, and C. Pedraza, "Extraction and Analysis of Social Networks Data to Detect Traffic Accidents," *Inf.*, vol. 13, no. 1, 2022, doi: 10.3390/info13010026.

[23] D. Shivakoti, "Automatic Detection and Extraction of Event Locations in News Report to locate in Map," 2016.

[24] T. S. Akinyetun, "Youth Political Participation, Good Governance and Social Inclusion in Nigeria: Evidence from Nairaland," *Can. J. Fam. Youth / Le J. Can. Fam. la Jeun.*, vol. 13, no. 2, pp. 1–13, 2021, doi: 10.29173/cjfy29648.

[25] A. Kapan, S. Kirmizialtin, R. Kukreja, and D. J. Wrisley, "Fine-tuning NER with spaCy for transliterated entities found in digital collections from the multilingual Persian Gulf," CEUR Workshop Proceedings, Apr. 2022. [Online]. Available: http://archive.nyu.edu/handle/2451/63943

[26] K. Satheesh, A. Jahnavi, L. Iswarya, K. Ayesha, G. Bhanusekhar, and K. Hanisha, "Resume ranking based on job description using SpaCy NER model," *Int. Res. J. Eng. Technol.*, vol. 7, no. 05, pp. 74–77, Apr. 2020, [Online]. Available: https://www.academia.edu/download/64550140/IRJET-V7I516.pdf

[27] B. Jehangir, S. Radhakrishnan, and R. Agarwal, "A survey on Named Entity Recognition — datasets, tools, and methodologies," *Nat. Lang. Process. J.*, vol. 3, p. 100017, Apr. 2023, doi: 10.1016/j.nlp.2023.100017.

[28] M. Kumar, "An Algorithm for Automatic Text Annotation for Named Entity Recognition using spaCy Framework," *Res. Sq.*, no. May, pp. 1–18, 2023, doi: 10.21203/rs.3.rs-2930333/v1.

[29] H. F. Akande, A. Muhsin, and R. L. Lawal, "Enhancing Data Extraction from Scanned Official Correspondences Using Named Entity Recognition : A Case Study at Kaduna Polytechnic," *Int. J. Sci. Eng. Res.*, no. June, pp. 1–6, 2023.

[30] C. Berragan, A. Singleton, A. Calafiore, and J. Morley, "Transformer based named entity recognition for place name extraction from unstructured text," *Int. J. Geogr. Inf. Sci.*, vol. 37, no. 4, pp. 747–766, Apr. 2023, doi: 10.1080/13658816.2022.2133125.

[31] S. Sharma and M. Mohania, "Comparative Analysis of Entity Identification and Classification of Indian Epics," in *ICMI '22: International Conference on Multimodal Interaction*, ACM, Apr. 2022, pp. 404–413. doi: 10.1145/3536221.3556573.

[32] D. Lemke, V. Mattauch, O. Heidinger, and H. W. Hense, "Who hits the mark? A comparative study of the free geocoding services of Google and OpenStreetMap," *Gesundheitswesen*, vol. 77, no. 8–9, pp. e160-5, Apr. 2015, [Online]. Available: https://europepmc.org/article/med/26154258

[33] J. Ibrahim, C. Loch, and K. Sengupta, "Two Express Road Rehabilitation Projects," in *How Megaprojects Are Damaging Nigeria and How to Fix It*, 2022, pp. 161–175. doi: 10.1007/978-3-030-96474-0_9.

[34] R. T. Ayinde, K. O., Omotosho, S. M., Amusa, N. A., Adisa, A. L., Abiola, O., & Omotope, "Heavy Metal Pollution From Vehicular Exhausts on Napier Grass (Pennisetum purpureum) along Lagos-Ibadan Expressway, Southwest, Nigeria," *Ethiop. J. Environ. Stud. Manag.*, vol. 14, no. 1, pp. 98–112, 2021, doi: https://ejesm.org/doi/v14i1.8.

[35] T. Goddard, K. Ralph, C. G. Thigpen, and E. Iacobucci, "Does news coverage of traffic

crashes affect perceived blame and preferred solutions? Evidence from an experiment," *Transp. Res. Interdiscip. Perspect.*, vol. 3, p. 100073, 2019, doi: 10.1016/j.trip.2019.100073.

[36] T. De Ceunynck, J. De Smedt, S. Daniels, R. Wouters, and M. Baets, "'Crashing the gates'–selection criteria for television news reporting of traffic crashes," *Accid. Anal. Prev.*, vol. 80, pp. 142–152, Apr. 2015, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0001457515001396

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: