GEO-PARSING AND GEO-VISUALIZATION OF ROAD TRAFFIC CRASH INCIDENT LOCATIONS FROM PRINT MEDIA FOR EMERGENCY RESPONSE AND PLANNING

Patricia Ojonoka Idakwo*, **Olubayo Adekanmbi & Anthony Soronnadi**
Data Science Nigeria
AI Hub, 33 Queens Street, Alagomeji
Yaba, Lagos, Nigeria
{patricia,olubayo,anthony}@datasciencenigeria.ai

Amos David
School of Computing and Information Technology
African University of Science and Technology
Abuja, Nigeria
{adavid}@aust.edu.ng

ABSTRACT

Road traffic crashes (RTC) are a major public health concern across the globe, particularly in Nigeria where road transport is the most common mode of transportation. In this paper, we present an approach to RTC related geographic information retrieval and visualization from news articles utilizing the geo-parsing natural language processing technique for emergency response and planning. To capture RTC-details with a high degree of accuracy and precision, we created a dataset from RTC related Nigerian news articles, and developed the RTC-NER Baseline and RTC-NER custom spaCy - based Named Entity Recognition (NER) models using the RTC dataset. We evaluated and compared their performance using standard metrics of precision, recall, and f1-score. The RTC-NER performed better than the RTC-NER baseline model for all three metrics with a precision rating of 93.63, recall of 93.61, and f1-score of 93.62. We further used the models for toponym recognition to extract RTC location details, toponym resolution to retrieve corresponding geographical coordinates, and finally, geo-visualization of the data to display the RTC incident environment for emergency response and planning. Our study showcases the potential of unstructured data for decision-making in RTC emergency responses and planning in Nigeria.

1 INTRODUCTION

Road traffic crashes (RTCs) are a developmental challenge and a major public health concern. In 2016, injuries from RTCs were the ninth leading cause of death among people of all ages and the leading cause of death among children and young adults with about 1.2 million deaths globally WHO (2015). There has been an increasing number of deaths from RTC injuries among young people in low and middle income countries over the years, with about 44 percent of these deaths occurring in lower middle-income countries such as Nigeria Ahmed et al. (2023); Awoniyi et al. (2022).

Nigeria has the highest proportion of injuries and deaths from RTCs in Africa, with RTC's being the leading cause of trauma-related deaths, the third-leading cause of deaths, and the most common cause of disability Onyemaechi & Ofoma (2016). There has been an influx of vehicles in Nigeria over the years, with the resulting effect of increased road traffic and RTCs Audu et al. (2021); Rembalovich et al. (2020). For Nigeria to meet the United Nations' Decade of Action for Road

*Additional Affiliation: School of Computing and Information Technology, African University of Science and Technology, Abuja, Nigeria {pojonoka}@aust.edu.ng

Safety 2021-2030 target of halving deaths from RTC injuries by 2030 through timely post-crash responses Rosen et al. (2022), it is imperative that emerging technologies such as machine learning be applied to RTC data for data-driven emergency responses and planning.

RTC incidents on Nigeria's roads are reported in detail in the print media; hence the information extracted from the textual data in the news articles is suitable for RTC emergency response and planning Shivakoti (2016). RTC location details can be extracted from such reports using a natural language processing (NLP) technique called Geo-parsing, which is a vital component of geographic information retrieval (GIR), and geographic information extraction (GIE) for geospatial analysis and visualization Wang et al. (2022).

Geo-parsing is the process of extracting toponyms (place names or location entities) from text, and linking it to corresponding geographic coordinates in two key steps: toponym resolution and toponym recognition Liu et al. (2022). Toponym recognition (location entity recognition) is a subset of named entity recognition (NER) involving the identification of toponyms in text such as news articles, social media posts, and other forms of literature Ma et al. (2023). Toponym resolution (geocoding) entails linking toponyms with corresponding geographic coordinates such as latitude and longitude.

In this study, we employed an integrated approach for geo-parsing through custom NER in order to implement a domain-specific NER model for RTC news articles, geocoded the extracted geographic information, and performed geo-visualization. We developed the RTC-NER Baseline and RTC-NER models, and compared their performances in recognizing toponyms in RTC related news articles.

The contribution of this study to research entails creation of RTC dataset, RTC NER models for toponym recognition in an integrated approach to geo-parsing and geo-visualization of RTC incident locations for speedy emergency response and planning.

2 METHODOLOGY

The methodological framework in Figure 1 employed for teh study. It comprises of 4 key stages namely: Data Collection; Data Pre-processing and Custom NER Model Training; Geo-parsing; and Geo-Visualization.

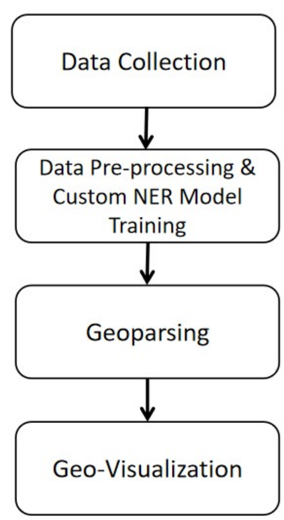

Figure 1: Methodological Framework

2.1 DATA COLLECTION

To assemble a comprehensive dataset of road accident news in Nigeria, a query combining RTC terms "Road accidents + crash" was used to search Nigerian online print media: DailyPost, The Sun

Table 1: Dataset Words Distribution

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

and the Punch. The query yielded 856 news articles published within a seven-year period between October 2, 2015 and October 9, 2023.

This approach to dataset curation from media reporting of RTC is prone to biases which could lead to non-representative dataset Tim et al. (2015) and possibly reduce the accuracy of our models. Some of these biases are: victim bias(RTCs involving celebrities, politicians or high profile entities receiving more coverage than crashes involving ordinary individuals); severity bias (focus of media coverage on crashes with high casualty figures); place bias (media report of RTCs in urban centers as opposed to rural areas due to concentration of journalists and high population in urban centers; vehicle type bias (media report of RTCs involving buses and cars as opposed to those involving pedestrians and motorcycles); and time bias (media reports are usually of more recent incidents thus overlooking older RTC incidents).

2.2 Data Pre-processing and Model Training

Data was cleaned by using python regular expressions to remove non-ascii characters, extra spaces, quotation marks and other punctuation marks such as questions marks from each news article. For increased detail and precision, further data pre-processing was conducted in a two-phased approach shown in Figure 2.

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

Phase 2 entailed scaling to the entire dataset of 856 news articles. Sentence tokenization was then done using python nlp library to split the entire corpus into 9584 sentences. Based on the sentence structure of Nigerian RTC news reporting, sentences which 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 split into training (3374) and test (844) data using the 80-20 rule for annotation.

2.2.1 Datasets Words Distribution

To develop a robust and efficient NER and de-bias it in order to avoid shortcut learning in the RTC-NER model, the distribution of words that translate to entities in the RTC-NER was examined Ma et al. (2023). Words which represent roads, places, and hospitals were grouped accordingly for the training and test datasets in each iteration, as follows: Roads [Road, Highway, Expressway]; Place [Village, Town, Community, Local Government Area (LGA), State]; and Hospital [center, centre, hospital, clinic]. The percentage distribution for RTC-related word occurrences in Table 1 shows an almost equal distribution of word groups across the training and test datasets in both iterations, with training having approximately 80 percent and test having approximately 20 percent in accordance with the 80-20 rule for splitting data into training and test data. The balanced performance of both RTC-NER models is therefore assured.

2.2.2 Corpus Annotation with spaCy NER Annotator

An annotation tool called the "NER-Annotator", a user-friendly web interface for manual annotation of entities for spaCy model training was utilized Kapan et al. (2022). We defined a set of custom tags/labels of relevance to RTC incidents as shown in Table 2. The Training and test data were converted to individual .txt files as input and JSON files were produced as output of the NER-Annotator. Find a screenshot of the NER-annotator web interface in Figure 3.

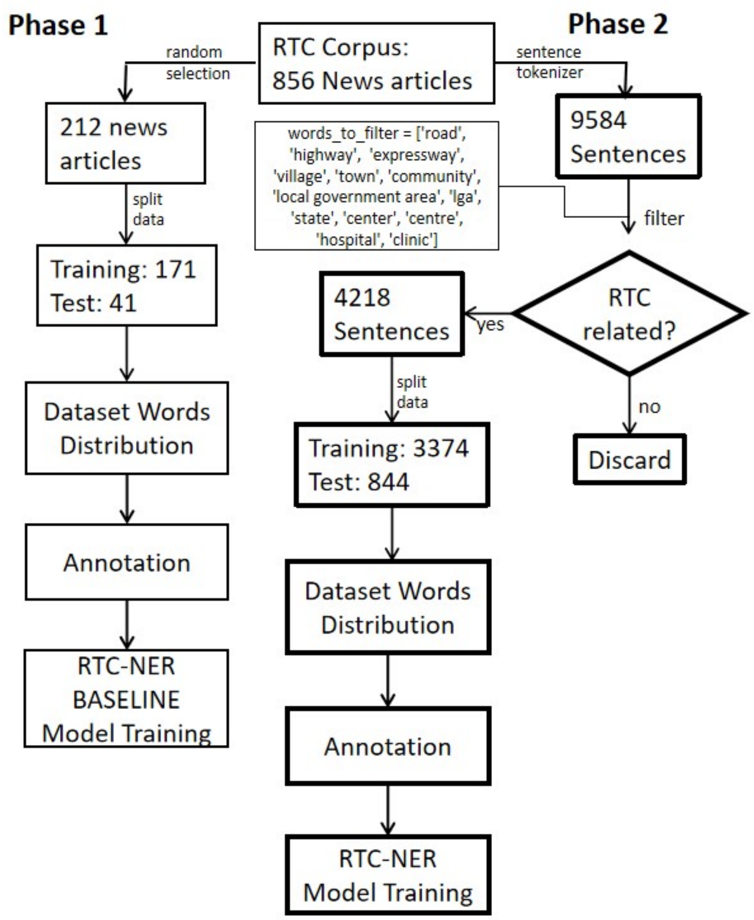

Figure 2: Two-Phased Approach to Data Pre-processing and Model Training

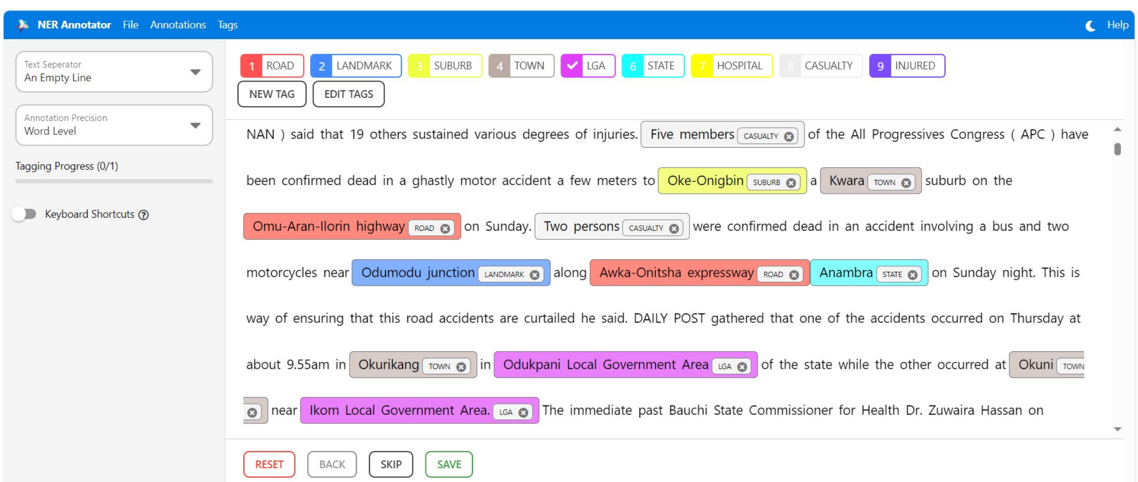

Figure 3: SpaCy NER-Annotator interface

2.2.3 NER TRAINING WITH SPACY 3.61

SpaCy 3.61 is a Python-based open-source library for NLP with several multi-lingual pre-trained models such as '`en_core_web_sm`' which can identify up to 18 entities ranging from people, dates,

Table 2: Custom Tags used for our RTC Corpus

Tag Name	Description	Example
ROAD	The name of the road where RTC event occurred	Lagos-Ibadan Express Way
LANDMARK	Place names with spatial relation to RTC event location	near Foursquare camp
SUBURB	Settlement in a town where RTC event occurred	Ogbere
TOWN	Name of the town of RTC event or where suburb is located	Ajebo
LGA	Local Government Area where town is Located	Remo North
STATE	The state where LGA / Town are located in	Ogun State
HOSPITAL	Hospital name where RTC victims where taken to	Victory Hospital, Ogbere
CASUALTY	Number of people who died due to the RTC event	5 deaths
INJURED	Number of people Injured in the RTC event	3 Injured

city to organizations Satheesh et al. (2020); Kapan et al. (2022). Additionally, spaCy provides features for developing and re-training custom NER models on domain specific entities since its pre-trained models fail to identify these entities in text Berragan et al. (2022); Sharma & Mohania (2022). The choice of SpaCy 3.61 for the NER model training was therefore based on its ease of use, cost-effectiveness, flexibility and efficiency enabling us to focus on data preparation and model development in light of the complexity of the RTC corpus.

The training pipeline entailed fine-tuning the blank SpaCy NER model with our annotated training datasets to develop the RTC-NER Baseline and RTC-NER models. The RTC-NER Baseline model was trained in 4200 steps (epochs) while RTC-NER model was trained in 4600 steps (epocs). Both 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” tokenizers; ”Adam.v1” optimizer; batch size of 1000; dropout rate of 0.1; learn_rate of 0.001 and evaluation frequency of 200.

2.3 GEO-PARSING

Toponym recognition was carried out using the RTC-NER Baseline and RTC-NER models for GIR; while toponym resolution was performed using the Google Geocoding API, The choice of Google API for geocoding was based on its cost-effectiveness, high accuracy, ease of use and extensive global coverage =for accurate geocoding Lemke D (2015). The output of the geo-parsing on a sample test RTC news articles is displayed in Figure 4.

2.4 GEO-VISUALIZATION

The results of geo-parsing of the sample toponyms seen in Figure 4 were displayed on an interactive leaflet map as shown in Figure 5. This was done using Folium, a Python library for geographic data visualization in Jupyter notebook Kurada et al. (2021); Aghav et al. (2022).

3 RESULTS AND DISCUSSION

The results from the study are grouped into performance metrics for the RTC-NER baseline and RTC-NER models, and for the entities in each models. Research output in the form of the extracted geographic information and maps from the geo-parsing and geo-visualization stages are also discussed below.

Table 3: Comparison of RTCNER Baseline and RTCNER Models

Model	Precision	Recall	F1-Score	No. of Test Samples
RTC_NER Baseline	92.37	90.15	91.25	44
RTC_NER	93.63	93.61	93.62	844

3.1 Performance Evaluation

Performance of the custom NER models was measured using standard metrics such as precision, f1-score, and recall Kapan et al. (2022). The test data used for the model evaluation were 44 news articles for the RTC-NER Baseline model and 844 sentences for the RTC-NER model.

Table 3 shows that the overall the RTC-NER model is a better performing model than the RTC-NER baseline for all three metrics with Precision of 93.63, recall of 93.61 and F-Score of 93.62. Going by the precision and recall values, the RTC-NER model makes fewer false positive predictions (identifying non-entities as entities); and misses fewer true positive predictions (failing to identify actual entities).

The performance of the models at the entities level shown in Table 4 confirms that the RTC-NER model with a larger training data achieved a better overall performance than the RTC-NER Baseline model. All RTC-NER entities performed better than those of the RTC-NER Baseline with ROADS having the highest F1-Score of 97.71 and LANDMARK having the lowest F1-score of 88.82.

For recall and precision, LANDMARK achieved the lowest recall at 85.56; SUBURB earned the lowest precision rating of 86.84 among entities in the RTC-NER model. Thus, the likelihood of the RTC-NER model failing to identify actual 'LANDMARK' entities is highest while the likelihood of identifying non-entities as 'SUBURB' entities is highest. This is probably due to the fact that 'LANDMARK' and 'SUBURB' entities appeared the least number of times in the Training datasets as landmarks are not often used to describe RTC incident locations in news articles, while RTC incidents in suburbs are not often reported in news articles owing to place bias.

Table 4: Entity-Level Comparison of RTC-NER Baseline and RTC-NER Models

Entity	RTCNER Baseline			RTCNER		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score
ROAD	96.41	95.43	95.92	97.54	97.88	97.71
LGA	92.19	92.19	92.19	95.32	96.40	95.86
STATE	87.31	83.57	85.40	94.32	94.16	94.24
HOSPITAL	93.23	89.86	91.51	95.02	95.69	95.35
TOWN	91.52	90.42	90.96	91.15	92.57	91.85
LANDMARK	91.46	87.21	89.29	92.33	85.56	88.82
INJURED	0	0	0	90.41	90.59	90.50
CASUALTY	0	0	0	90.28	89.44	89.86
SUBURB	-	-	-	86.84	91.67	89.19

3.2 Geo-parsing and Geo-visualization

The geo-parsing phase produced toponyms from the sample RTC news article, and their corresponding latitude and longitude as shown in Figure 4. These are then visualized on an interactive map as shown in Figures 5 and 6.

After detecting three unique latitudes and longitudes in the sample RTC news article, the interactive web map shows the points representing these unique sets of coordinates. The name of the road: 'Lagos-Ibadan Expressway', can be seen on the red line feature in the map in Figure 5, while the three distinct points can be seen as blue markers in Figure 6.

In Figure 5, the coordinate points to the center of the identified road, and not necessarily at the exact point on the road where the RTC occurred. According to Wang et al. (2022), points at the centers of towns, cities, villages or polygon/line geospatial features are the usual output of geo-parsing toponyms from text leading to a distance offset between actual event locations and geo-

	Location	Latitude	Longitude
0	Lagos-Ibadan Expressway.	6.923588	3.636422
1	near the Foursquare Camp, Ajebo	7.109120	3.723304
2	Victory Hospital, Ogbere	6.739754	4.164174

Figure 4: Latitude and Longitude of some extracted RTC Toponyms

parsed locations. In our study, distance offsets can drastically reduce accuracy and efficiency of the RTC-NER models leading to delay in victim post-care, loss of lives, lower confidence of first responders in emergency response, as well as introduce noise into the RTC location data making it unreliable for further spatial analysis such as identification of RTC hotspots.

To overcome the challenge of distance offsets, we identified entities such as 'LANDMARK' and 'HOSPITAL' which represent points on the earth surface to determine the actual RTC incident sites. In our sample RTC news article, the recognized toponyms which are points on the Earth's surface are: Victory Hospital, Ogbere (HOSPITAL) and near the Foursquare Camp, Ajebo (LANDMARK).

In Nigeria, the choice of hospital by RTC rescue teams is based on the availability of medical personnel, equipment, or space in the emergency ward, not necessarily their proximity to the RTC incident site. As such, hospitals are not fit to be used as determinants of RTC incident locations. Landmarks on the other hand have spatial relations to the actual RTC sites and give more accurate insight on RTC incident locations; hence, they are more relevant for our study. In Figure 5, the actual point on the road where the RTC incident occurred falls within a buffer area of the landmark (that is the area covered by the light blue circle).

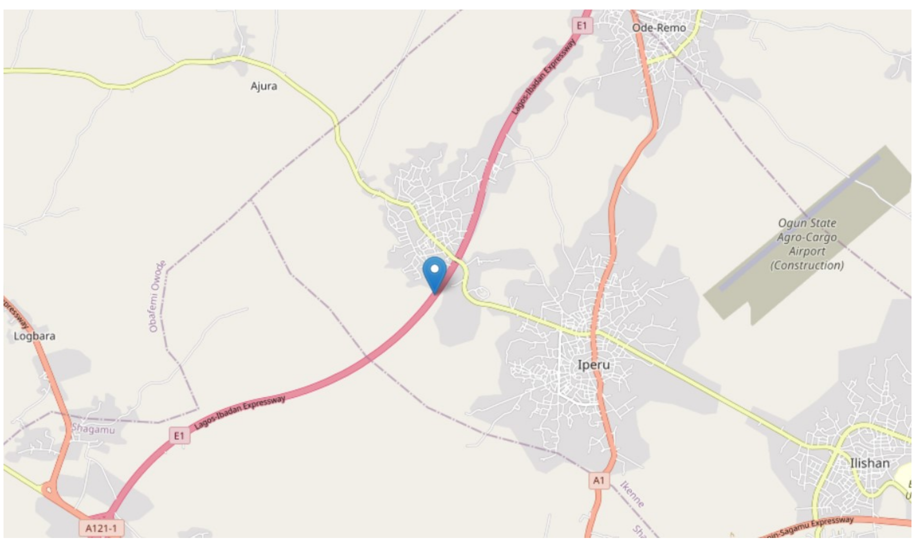

Figure 5: Map Showing the RTC Road (Lagos-Ibadan Expressway) Location

4 Conclusion and Future Work

In this paper, we developed an extensive methodological framework for RTC domain-specific NER models, namely RTC-NER Baseline and RTC-NER, which both have very high accuracy and precision scores and overall high performance. With fine-tuning on a larger corpus, the RTC-NER outperformed the RTC-NER Baseline model in achieving the initial goal of GIR of RTC entities from RTC-related news articles. The geographic coordinates of the toponyms were then extracted for geo-visualization of the RTC incident environment.

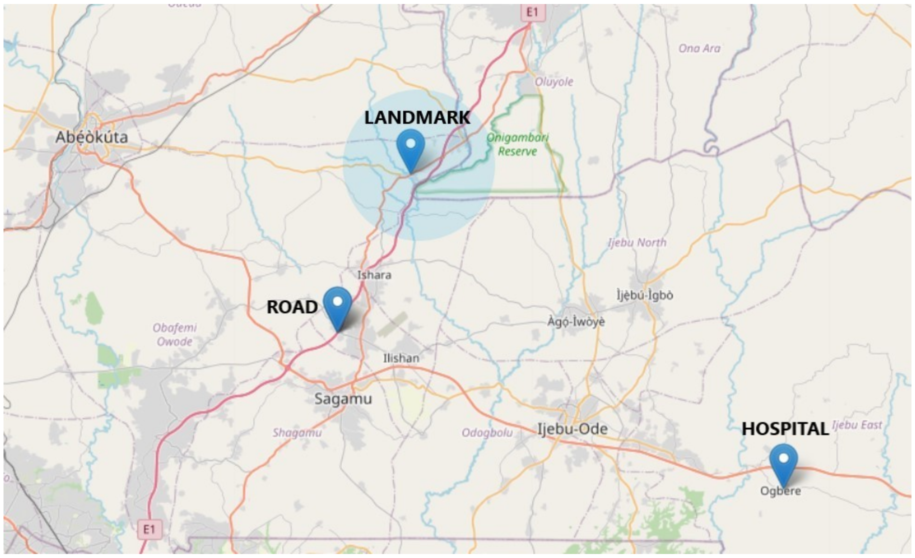

Figure 6: Map of the RTC Related Toponyms

A limitation of the study is non-representativeness of the dataset due to biases in media reporting of RTC which could affect the accuracy of our models. Another is the offset distance between the actual RTC incident location and the geo-parsed location which could affect emergency response and planning decisions. In out study, this challenge of distance offset was overcome by identifying 'LANDMARK' entity for more accurate and reliable geo-visualization of the RTC incident locations.

The output of this study including the custom RTC dataset and RTC-NER models for geo-parsing of RTC incident locations can be made openly available for further research in emergency response and planning. They also form a baseline for future work in reducing bias in the dataset by including other sources of data such as social media and official RTC reports; was well as reducing offset distances between RTC incident locations and geo-parsed locations. Furthermore, the RTC-NER model can be optimized for better performance through hyper-parameter tuning.

References

Shubham Aghav, Prashant Solanki, and Sushila Palwe. Gis for disaster management: A system to aid users with relevant information. In *Proceedings of the Third International Conference on Information Management and Machine Intelligence: ICIMMI 2021*, pp. 159–165. Springer, 2022.

Sirwan K Ahmed, Mona G Mohammed, Salar O Abdulqadir, Rabab G Abd El-Kader, Nahed A El-Shall, Deepak Chandran, Mohammad E Ur Rehman, and Kuldeep Dhama. Road traffic accidental injuries and deaths: A neglected global health issue. *Health science reports*, 6(5):e1240, 2023.

Akeem A Audu, Olufemi F Iyiola, Ayobami A Popoola, Bamiji M Adeleye, Samuel Medayese, Choene Mosima, and Nunyi Blamah. The application of geographic information system as an intelligent system towards emergency responses in road traffic accident in ibadan. *Journal of transport and supply chain management*, 15:17, 2021.

Oluwafunbi Awoniyi, Alexander Hart, Killiam Argote-Aramendiz, Amalia Voskanyan, Ritu Sarin, Michael S Molloy, and Gregory R Ciottone. Trend analysis on road traffic collision occurrence in nigeria. *Disaster medicine and public health preparedness*, 16(4):1517–1523, 2022.

Cillian Berragan, Alex Singleton, Alessia Calafiore, and Jeremy Morley. Transformer based named entity recognition for place name extraction from unstructured text. *International Journal of Geographical Information Science*, pp. 1–20, 2022. ISSN 1365-8816.

Almazhan Kapan, Suphan Kirmizialtin, Rhythm Kukreja, and David Joseph Wrisley. Fine-tuning ner with spacy for transliterated entities found in digital collections from the multilingual persian gulf. volume 3232, pp. 288–296. CEUR Workshop Proceedings, 2022.

Ramachandra Rao Kurada, Y Ramu, and Sunil Pattem. Lessoning geospatial visualizations on real-time data. In *2021 IEEE International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS)*, pp. 1–6. IEEE, 2021.

Heidinger O Hense HW Lemke D, Mattauch V. Who hits the mark? a comparative study of the free geocoding services of google and openstreetmap. *Proceedings of the 2022 International Conference on Multimodal Interaction*, 77(8-9), 2015.

Zilong Liu, Krzysztof Janowicz, Ling Cai, Rui Zhu, Gengchen Mai, and Meilin Shi. Geoparsing: Solved or biased? an evaluation of geographic biases in geoparsing. *AGILE: GIScience Series*, 3: 9, 2022.

Ruotian Ma, Xiaolei Wang, Xin Zhou, Qi Zhang, and Xuan-Jing Huang. Towards building more robust ner datasets: An empirical study on ner dataset bias from a dataset difficulty view. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 4616–4630, 2023.

NOC Onyemaechi and Uchenna R Ofoma. The public health threat of road traffic accidents in nigeria: A call to action. *Annals of medical and health sciences research*, 6(4):199–204, 2016.

Georgy Rembalovich, Vyacheslav Terentyev, Konstantin Andreev, Nikolay Anikin, and Vladimir Teterin. Improving the emergency system for a traffic accident. In *IOP Conference Series: Materials Science and Engineering*, volume 918, pp. 012072. IOP Publishing, 2020.

Heather E Rosen, Imran Bari, Nino Paichadze, Margaret Peden, Meleckidzedeck Khayesi, Jesús Monclús, and Adnan A Hyder. Global road safety 2010–18: an analysis of global status reports. *Injury*, 2022.

K Satheesh, A Jahnavi, L Iswarya, K Ayesha, G Bhanusekhar, and K Hanisha. Resume ranking based on job description using spacy ner model. *International Research Journal of Engineering and Technology*, 7(05):74–77, 2020.

Shreya Sharma and Mukesh K. Mohania. Comparative analysis of entity identification and classification of indian epics. pp. 404–413, 2022. doi: 10.1145/3536221.3556573.

Dinesh Shivakoti. Automatic detection and extraction of event locations in news report to locate in map. Master's thesis, University of Stavanger, Norway, 2016.

De Ceunynck Tim, De Smedt Julie, Daniels Stijn, Wouters Ruud, and Baets Michèle. "crashing the gates" - selection criteria for television news reporting of traffic crashes. volume 80, pp. 142–152. Elsevier, 2015.

Shu Wang, Xinrong Yan, Yunqiang Zhu, Jia Song, Kai Sun, Weirong Li, Lei Hu, Yanmin Qi, and Huiyao Xu. New era for geo-parsing to obtain actual locations: A novel toponym correction method based on remote sensing images. *Remote Sensing*, 14(19):4725, 2022.

World Health Organization WHO. *Global status report on road safety 2015*. World Health Organization, 2015.