International Journal of Digital Earth

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tjde20

TopoBERT: a plug and play toponym recognition module harnessing fine-tuned BERT

Bing Zhou, Lei Zou, Yingjie Hu, Yi Qiang & Daniel Goldberg

To cite this article: Bing Zhou, Lei Zou, Yingjie Hu, Yi Qiang & Daniel Goldberg (2023) TopoBERT: a plug and play toponym recognition module harnessing fine-tuned BERT, International Journal of Digital Earth, 16:1, 3045-3063, DOI: 10.1080/17538947.2023.2239794

To link to this article: https://doi.org/10.1080/17538947.2023.2239794

Published online: 10 Aug 2023.

INTERNATIONAL JOURNAL OF DIGITAL EARTH
2023, VOL. 16, NO. 1, 3045–3063
https://doi.org/10.1080/17538947.2023.2239794

TopoBERT: a plug and play toponym recognition module harnessing fine-tuned BERT

Bing Zhou[a], Lei Zou[a], Yingjie Hu[b], Yi Qiang[c] and Daniel Goldberg[a]

[a]Department of Geography, Texas A&M University, College Station, USA; [b]Department of Geography, University at Buffalo, Buffalo, USA; [c]School of Geoscience, University of South Florida, Tampa, USA

ABSTRACT

Extracting precise geographical information from the textual content, referred to as toponym recognition, is fundamental in geographical information retrieval and crucial in a plethora of spatial analyses, e.g. mining location-based information from social media, news reports, and surveys for various applications. However, the performance of existing toponym recognition methods and tools is deficient in supporting tasks that rely on extracting fine-grained geographic information from texts, e.g. locating people sending help requests with addresses through social media during disasters. The emerging pretrained language models have revolutionized natural language processing and understanding by machines, offering a promising pathway to optimize toponym recognition to underpin practical applications. In this paper, TopoBERT, a uniquely designed toponym recognition module based on a one-dimensional Convolutional Neural Network (CNN1D) and Bidirectional Encoder Representation from Transformers (BERT), is proposed and fine-tuned. Three datasets are leveraged to tune the hyperparameters and discover the best strategy to train the model. Another seven datasets are used to evaluate the performance. TopoBERT achieves state-of-the-art performance (average f1-score = 0.854) compared to the seven baseline models. It is encapsulated into easy-to-use python scripts and can be seamlessly applied to diverse toponym recognition tasks without additional training.

ARTICLE HISTORY
Received 22 December 2022
Accepted 18 July 2023

KEYWORDS
Natural language processing; geoparser; convolutional neural network; toponym recognition; BERT

1. Introduction

Since the emergence of social sensing, scholars have been endeavoring to sense the pulse of society with the help of satellite images, sensor networks from the Internet of Things (IoT), and various forms of textual information from the Internet. Extra attention has been paid to mining knowledge from social media because people nowadays are consciously or unconsciously sharing their views towards ongoing events online, which propels social media to become one of the few agents that reflects the real-time societal awareness, reactions, and impacts of particular events. This trait is a rare feature seldom shared by other forms of data sources.

In light of this feature, social media have been extensively associated with locations and leveraged in spatial analysis and modeling for various applications. Avvenuti et al. (2014) presented an early earthquake detecting and warning system using geotagged Twitter data, which offers prompt detection of events. Several case studies processed social media data with geocoding and

CONTACT Lei Zou lzou@tamu.edu Department of Geography, Texas A&M University, College Station, TX, 77843, USA

sentiment analysis tools to analyze the spatial patterns of changing public awareness and emotions toward hurricanes in different phases of the disaster management cycle (Zou et al. 2018, 2019). Huang et al. scrutinized the human mobility patterns during the COVID-19 pandemic at multiple scales based on geotagged Twitter data (Huang et al. 2020). A recent work proposed VictimFinder, a deep learning-based framework for harvesting help requests from social media during hurricanes (Zhou et al. 2022). The combination of social media and remote sensing data was used to assess damage during emergencies (Cervone et al. 2017). Social media data were also demonstrated to be valuable for informing sustainable urban planning (Abdul-Rahman et al. 2021; Milusheva et al. 2021).

The aforementioned studies and other social media-based spatial analysis and modeling investigations depend highly on extracting the location information of social media data. However, social media users have started to pay more attention to privacy, which resulted in a significant drop in the number of geotagged posts (Lin et al. 2022). Simultaneously, social media companies such as Twitter have published policies forbidding users to attach precise longitudes and latitudes to messages (Zou et al. 2023). Moreover, the geographical information bound up with social media posts might not necessarily be equivalent to the place names described in the textual content of the post, which plays critical roles in specific scenarios, e.g. when people request rescue for others through social media and mention victims' addresses in their messages. Thus, extracting location information from the textual content of social media data has inevitably become an issue that needs to be addressed. This breeds the process of geoparsing, a two-step approach that includes toponym recognition (identifying place names from texts) and toponym resolution (transforming location names to geographical coordinates). This paper focuses on the first component of geoparsing.

Existing studies on toponym recognition reveal that models built with deep learning architecture generally outperform other types of models, such as rule-based models and statistical-based models (Hu et al. 2022a; Hu et al. 2022b; Hu et al. 2022c). Most recent deep learning toponym recognition models are constructed with Recurrent Neural Network (RNN) or Long Short-Term Memory (LSTM), which are once considered as the silver bullet to process sequence data like natural language (Qi et al. 2020; Wang, Hu, and Joseph 2020). However, RNN, LSTM, and their variants, suffer from information vanishing problems when the sequence gets longer (Vaswani et al. 2017). The prevailing pretrained language models based on attention mechanism provide cure to this problem and are becoming the novel game changer in tasks relevant to natural language processing. However, most recent toponym recognition studies leveraging pretrained language models frequently used simple classifiers such as linear classifier or Conditional Random Field (CRF), limiting the performance of the constructed toponym recognition tools (Hu et al. 2022a; Ma et al. 2022). The Convolutional Neural Network (CNN) has been proven to be effective in signal processing (Zhao, Mao, and Chen 2019) and prominent in recognizing spatial, hierarchical local features because it is beneficial for incorporating Tobler's first law in the modeling process. Therefore, it has been applied in remote sensing object detection tasks (Li, Hsu, and Hu 2021). Since toponyms usually come in groups in texts, they can be considered spatially clustered in the textual space where Tobler's first law may apply. How the word embedding vectors concatenate in the vector space resembles how remote sensing images are scanned column-wise and row-wise. Therefore, combining CNN with pretrained language models is promising for designing toponym recognition models that can consider the clustering nature of place names in texts and yields optimal performance.

Inspired by the above idea, this paper proposes to develop a novel toponym recognition module, TopoBERT, based on a one-dimensional CNN and the Bidirectional Encoder Representation from Transformers (BERT), the landmark pretrained language model. We utilized three datasets to tune the hyperparameters to discover the optimal model training strategy and compared three classifiers to validate the advancement of using CNN as the classifier. Seven additional datasets were used to evaluate the performance of TopoBERT and compare it with seven baseline models. The novelty and major contributions of this work are listed below:

(1) A novel toponym recognition module architecture based on BERT and one-dimensional CNN, which has not been studied by existing work is proposed and tested.
(2) The study evaluates the performance of TopoBERT across seven baseline models and seven datasets, which is more inclusive than the majority of existing literature, to test the robustness and generalizability of the proposed method.
(3) A ready-to-use Python package is developed and made available to the public to service toponym recognition tasks and to benefit related studies in the future.

The remainder of this paper is structured as follows. Section 2 presents a literature review of related work and how this paper addresses the existing gaps. Section 3 concisely introduces the holistic design and implementation of the TopoBERT framework, as well as the datasets and parameters used in fine-tuning and testing the framework. The results of the experiments are documented in Section 4. Section 5 illustrates the potential limitations of this work and lists several future research directions. Finally, Section 6 epitomizes the findings and implications of this study.

2. Related work

2.1. Deep learning-based toponym recognition

Existing studies on toponym recognition can be categorized into four parties based on the character of the solutions, namely rule-based, gazetteer-based, statistical learning-based, and hybrid approaches (Hu et al. 2022a). Growing evidence shows that statistical learning and hybrid methods that incorporate deep learning techniques render better performance than methods that solely rely on rules or gazetteers (Dutt et al. 2018; Hu et al. 2022a; Qi et al. 2020; Wang, Hu, and Joseph 2020). For example, Wang et al. (2018) leveraged a Skip-Gram model for word representation and Deep Belief Network model to recognize toponyms among Chinese texts which outperformed CRF based approaches. Based on Bidirectional Long Short-Term Memory (BiLSTM), Wang, Hu, and Joseph (2020) introduced NeuroTPR to extract place names from social media messages. Qi et al. (2020) brought about an open-sourced named entity recognition python toolkit called Stanza, which is able to detect place names and supports multiple languages. SAVITR is a system that combines natural language processing (NLP) techniques and gazetteers for real-time location extraction (Dutt et al. 2018). GazPNE addressed the incompleteness of gazetteers and fused gazetteers, rules, and LSTM structure to form a reliable place name extractor (Hu et al. 2022a).

However, most existing models are based on the RNN, which might suffer from information vanishing problems in understanding textual content when the input sequence gets larger and the network deeper. Moreover, deep neural networks frequently require large, annotated datasets and are time-consuming to train to achieve feasible results. The emerging pretrained language models offer a promising pathway to optimize toponym recognition with a limited amount of training data. First, the multi-head attention mechanism in most pretrained language models (Devlin et al. 2018; Radford et al. 2019) enable the vectorized representations of the tokenized text to capture more contextual information of the entire input sentences (Vaswani et al. 2017). This mechanism helps better identify the role of a single token, e.g. a place name word, in a given sentence. Second, the pretrained language models are developed in a semi-supervised manner with large datasets such as Wikipedia (Devlin et al. 2018), such that fewer labeled data are required to fine-tune the model for specific downstream tasks. In this work, we leveraged BERT to address this limitation.

Most recently, the popularity of incorporating pretrained language models in named entity recognition (NER) tasks has increased. For example, BERT-base-NER is a Python package developed by the HuggingFace.[1] Such NER models based on BERT has been compared with other embedding methods like ELMO (Cardoso, Martins, and Estima 2022). Experiments were also conducted to test the performance of BERT in identifying named entities across different languages (Labusch et al.

2019; Luoma and Pyysalo 2020; Souza, Nogueira, and Lotufo 2020). Another package, GazPNE2, fused BERT into its workflow for place name disambiguation (Hu et al. 2022c). However, most existing models were designed to recognize named entities in general, including but not only place names, so their prediction of location name extraction might be disturbed. Moreover, the aforementioned studies leveraged pretrained language models merely to obtain contextual embedding. The models were attached to simple classifiers such as linear classifier or CRF (Hu et al. 2022a; Hu et al. 2022c; Ma et al. 2022; Souza, Nogueira, and Lotufo 2020).

2.2. Toponym recognition performance evaluation

Advancement in toponym recognition can benefit numerous fields that rely on analysis of unstructured texts with geographical information. Scholars have spared no efforts in improving the performance of toponym recognition. However, the performances of existing models and tools were not evaluated under identical criteria or with the same collection of datasets. A recent attempt is a comprehensive review comparing the performance and computational efficiency of 27 most widely used approaches for toponym recognition based on 26 public datasets (Hu et al. 2022b). The results showed that deep learning is by far the most promising technique in location recognition but improvements can still be made. Another study conducted a spatially explicit evaluation of the performance of the existing methods across different geographic space (Liu et al. 2022). The study identified that none of the methods tested performs equally well across geographic space, and some are biased towards some regions but against others (Liu et al. 2022). A lack of generalizability and spatial bias are the two key gaps identified by the two studies. In this work, we avoid designing overly complicated classifiers to avoid overfitting, thereby ensuring the generalizability of the model. Seven representative models are chosen as competitive baseline models and seven datasets are selected from existing literature as the testing data to thoroughly evaluate the robustness of the model performance.

3. Methodology

3.1. Pretrained model selection

Identifying location names from input sentences is a token classification task (Figure 1), which contains two parts: a language model and a classifier. It behaves similarly to how human beings analyze whether or not the given words are place names. First, the input sentence is tokenized to tokens, as exemplified in Figure 1. The 101 token is a predefined token that is intentionally appended to the starting point of a sentence. The 1030 token is the corresponding token of the given text in the dictionary. The tokenized sequence ends up with all '0' values because the maximum sequence length is set to 512 in this study. If the length of the original text is shorter than 512, the tokenizer will pad the sequence to 512 with '0' values. Second, the language model attempts to understand the language by transforming the tokenized input data into higher dimensional space as vectors that capture the meaning of words in a given sentence. Then the classifier makes predictions based on the transformed vectors and determines whether the input word belongs to the location entity. Finally, each word of the input sentence is categorized into one of three classes, namely others (O), the beginning of the location (B-LOC), and falling inside of the location (I-LOC).

In this study, BERT is selected as the core of TopoBERT. Numerous BERT variants have been proposed since BERT built its prestigious influence in the NLP domain. For example, ALBERT is lite version of BERT with fewer parameters but renders similar performance (Lan et al. 2020). DistilBERT is lighter version of BERT achieved with knowledge distillation (Sanh et al. 2020). RoBERTa is a robust language model optimized from BERT that achieved better performance across several NLP tasks (Liu et al. 2019). However, a recent study shows that leveraging such models do not necessarily boost the performance of sequence-to-sequence classification tasks

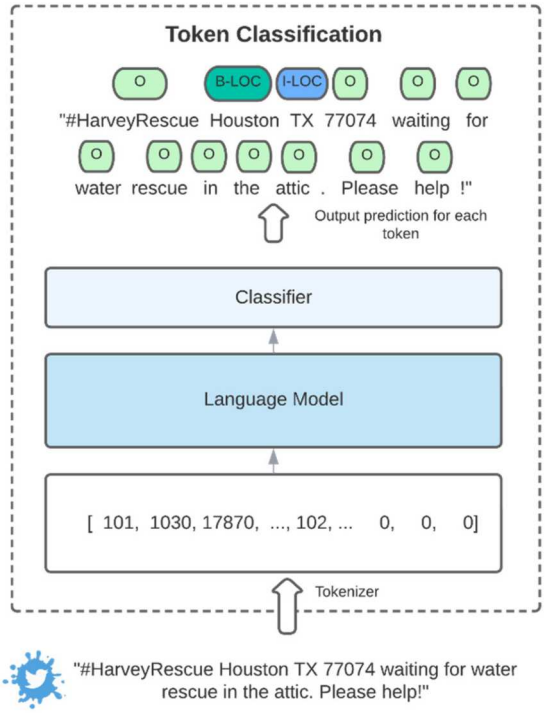

Figure 1. Demonstration of token classification workflow.

(Zhou et al. 2022). Therefore, the landmark model, BERT, can be used as a reasonable starting point.

BERT is structured by stacking the encoder components of the Transformer architecture and is designed to be pretrained in an unsupervised manner. It takes advantage of the Attention mechanism (Vaswani et al. 2017), which resolves the information vanishing issue that often upsets recurrent neural networks such as Long Short-Term Memory (Hochreiter and Schmidhuber 1997) and Gated Recurrent Neural Network (Bahdanau, Cho, and Bengio 2014) when the input sequence gets longer. Moreover, distinguished from many other bidirectional language models, such as ELMo (Peters et al. 2018), in which the contextual representation of every word is the concatenation or summation of the forward and backward representations, BERT reads the entire sequence of words at once and is trained using a Masked Language Model (MLM) approach and a Next Sentence Prediction (NSP) approach which genuinely has implemented the bidirectional or unidirectional concept. These two features combined facilitate better language understanding and bring the trophy to BERT throughout a number of NLP tasks under the General Language Understanding Evaluation (GLUE) benchmark (Devlin et al. 2018).

Off-the-shelf pretrained BERT model weights can be separated into several categories based on the size of the model, whether upper and lower cases are taken into consideration, the targeted language, and unique training strategies (https://huggingface.co/transformers/v3.3.1/pretrained_models.html). Since place names are case sensitive and only the English language is involved in this study, 'bert-base-cased' and 'bert-large-cased' are selected as the candidate pretrained models to be evaluated. The 'bert-base-cased' model comprises 12 layers, and each hidden layer has 768 nodes, with 12 self-attention heads and a total of 110 million parameters. The 'bert-large-cased' model consists of 24 layers, and each hidden layer has 1024 nodes, with 16 self-attention heads and 340 million parameters. The parameters are pretrained with case-sensitive English text from BooksCorpus (800 million words) and English Wikipedia (2500 million words).

3.2. Framework design and implementation

As mentioned earlier, there is an acute conflict between the need for sufficient geolocated social and news media to conduct robust spatial analysis and the diminishing availability of geotagged messages. Additionally, the geotagged and content-mentioned addresses might differ but are helpful in distinct scenarios. For instance, geotags indicate users' locations when posting on social media, which link users with places and are essential for investigating users' spatial perceptions and behaviors. On the other hand, the content-mentioned addresses are associated with the topics being discussed in messages, which bridge semantics with places and are valuable in discovering phenomena in different locations. Both locational data should be kept for further analyses and applications. A reliable and ready-to-use geoparser reconciling geotagged and text-mentioned locations to geolocate social and news media messages can resolve such challenges. Therefore, we present a general location extractor that can be used upon social media and news media. The workflow is shown in Figure 2.

The existing geotagged points, bounding boxes, and place names of the social media data are retained. The textual content goes through a rule-based data preprocessing module before they are fed to a place name extractor consisting of a zip code extractor and toponym recognition module. The data preprocessor first eliminates URLs, non-ASCII characters, '@' mentions, emojis and punctuations in the text with regular expression. Then the stop words provided by the NLTK toolkit[2] are removed to reduce the number of false positive predictions. Once the place names are pulled out from texts or geotags, a geocoding service is applied to transform the place names into precise coordinates. The toponym recognition module is marked with an orange dashed rectangle in Figure 2 and serves as the crucial backbone of the entire workflow.

By stacking a classifier on top of BERT, the combo can be fine-tuned to accomplish place name extraction from texts. A recent study shows that model performance can be enhanced by applying classifiers more complex than the simple linear classifier or CRF (Zhou et al. 2022). The CNN models are competent in detecting underlying features (Lee and Dernoncourt 2016), and one-dimensional Convolutional Neural Network (CNN1D) has been proven to be effective in signal processing (Kiranyaz et al. 2021; Zhao, Mao, and Chen 2019). In the toponym recognition case, the vector derived from the pretrained language models can be regarded as one-dimensional signals, the pattern of which can be properly captured by using CNN1D. Inspired by these ideas, we proposed TopoBERT, which leverages CNN1D to capture the pattern of the word embedding vectors extracted from the hidden layers of BERT. The CNN layer is followed by a simple multi-

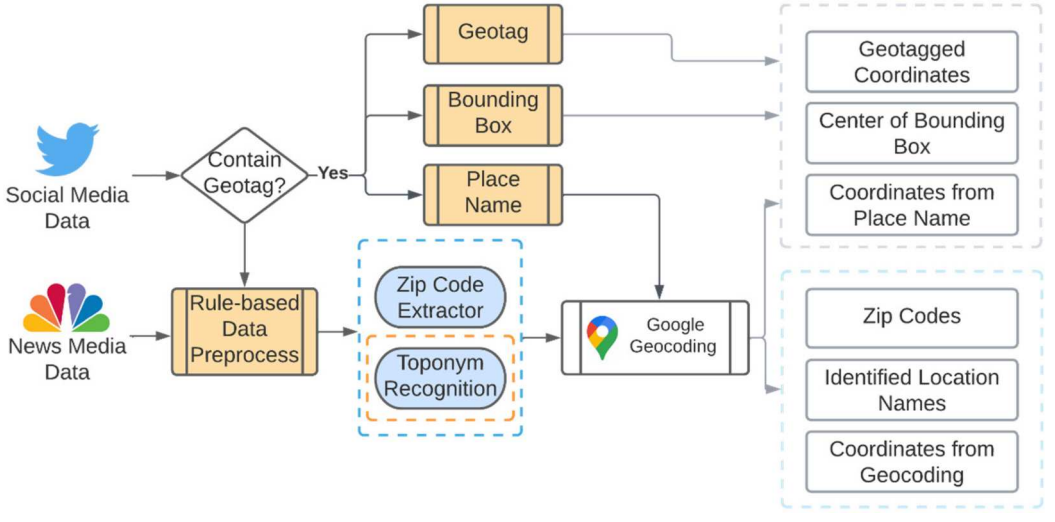

Figure 2. Holistic design of location extraction framework for textual content.

layer fully connected layer to perform token classification (Figure 3). To validate its advantage against other commonly used classifiers, i.e. linear classifier and multi-layer perceptron (MLP), two additional models are constructed. These three architectures are referred as BERT-CNN1D, BERT-Linear and BERT-MLP. CNN1D with kernel size 3 is applied in the BERT-CNN1D model (Figure 3). The output channel of the convolution is 16, followed by a max pooling layer of size 2, which further generalizes the features and reduces model complexity. All channels of the max pooling layer output are concatenated into a single vector and fed to a fully connected MLP with hidden layer size equals to 128. In the BERT-Linear model, the simple linear classifier connects the output of BERT to the final prediction results with the softmax activation function (Figure 3). The MLP classifier applied in the BERT-MLP model contains three fully connected layers (Figure 3). The input layer size is equivalent to the BERT output vector size. The number of hidden layer nodes is 256, and the output layer size equals the number of distinct tokens from the training dataset.

All model combinations were implemented using Python and pertinent packages. The dataset splitting took advantage of the ScikitLearn library,[3] and the BERT models were implemented based on the huggingface Transformer library.[4] The model fine-tuning pipeline was built using PyTorch[5] functions.

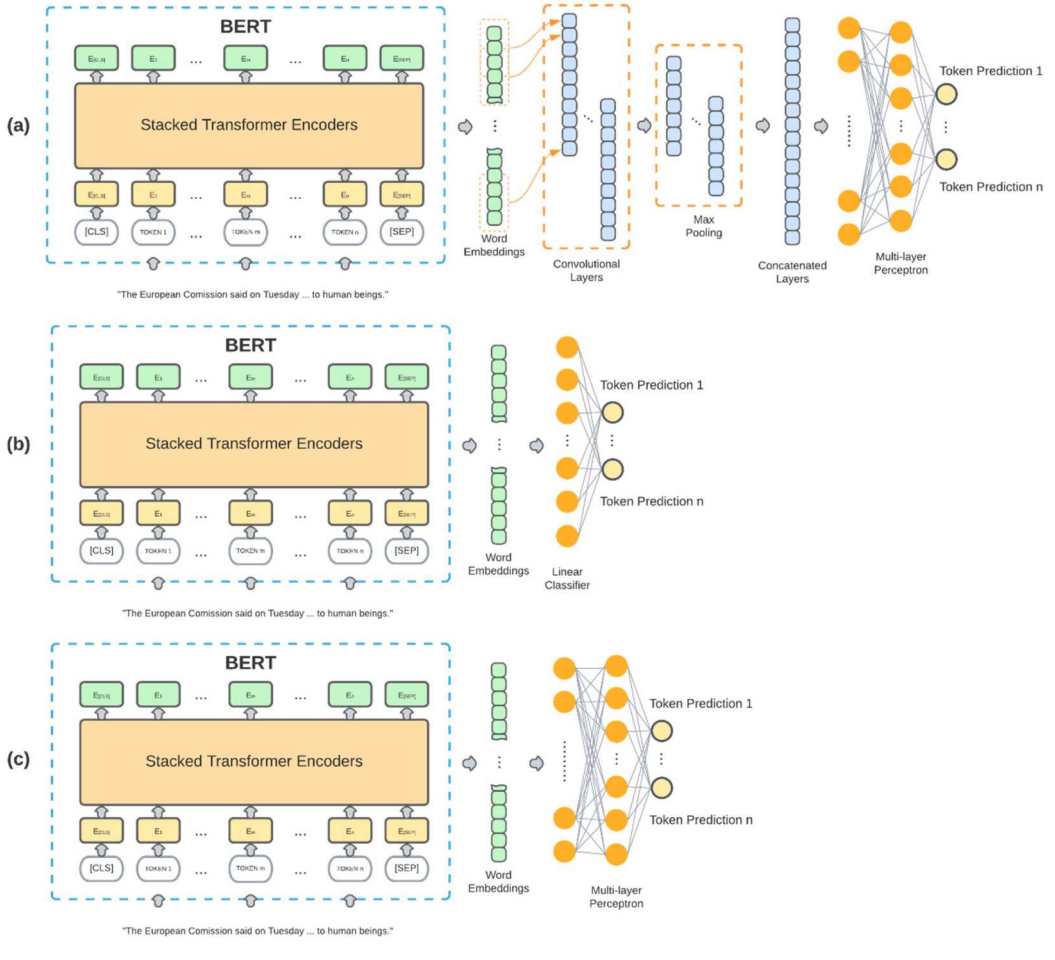

Figure 3. TopoBERT architectures. (a) Architecture with CNN1D as classifier. (b) Architecture with linear classifier. (c) Architecture with MLP as classifier.

Table 1. Summary of the datasets used in training and evaluation.

Dataset	Purpose	Entity Count	Place Name Count	Description
CoNLL2003	Training	203,621	8297	A dataset created for general-purpose NER task and benchmarking
WNUT2017	Training	11,813	1140	Dataset of shared task on NER in Twitter at the Workshop on Noisy User-generated Text in 2017
Wiki_Gen	Training	56,466	21,000	Automatically generated dataset from Wikipedia articles and location names
Harvey2017	Testing	19,268	3973	A dataset originally collected from the University of North Texas repository
LouFlood2016	Testing	30,220	219	Data collected from the 2016 Louisiana flood
HouFlood2015	Testing	28,761	220	Data collected from the 2015 Houston flood
NzEq2013	Testing	30,630	550	Data collected from the 2013 New Zealand earthquake
GeoCorpora	Testing	130,306	2546	Dataset generated from multiple hazardous events globally
Ritter's dataset	Testing	48,862	566	A dataset created for general-purpose NER task
HumAID-1000	Testing	23,429	1130	Manually annotated dataset

3.3. Datasets

Totally three different datasets were utilized to train the TopoBERT module and seven datasets were included to evaluate the performance (Table 1).

The data distribution of each label type in the three datasets is depicted in Figure 4(a)-(c), respectively. CoNLL2003 is a shared task that concerns NER and has been widely applied for training deep learning models (Sang, Kim, and Meulder 2003). The data contain entities of five types: persons (PER), organizations (ORG), locations (LOC) and miscellaneous names (MISC) and other words that are irrelevant to named entities of the aforementioned four groups (O). The prefix 'B-' and 'I-' are used to tag the beginning of a named entity and words that fall inside a named entity (Sang, Kim, and Meulder 2003). The dataset is originally divided into training, validation, and test data which are noted as CoNLL2003-Train, CoNLL2003-Validation and CoNLL2003-Test. Training data are used to train a deep learning model, validation data are used to tune the hyperparameters of the model, and test data are used to evaluate the performance of the trained model. The dataset was modified to suit the purpose of this study by labeling all the named entities as 'O' except for the location entities. Around 4.1% of the tags are location entities in this dataset.

WNUT2017 is a relatively smaller dataset collected from Twitter and manually annotated, the objective of which is to tackle the issues caused by novel, emerging, singleton named entities in noisy texts (Derczynski et al. 2017). It aims to offer support to sustainable named entity recognition

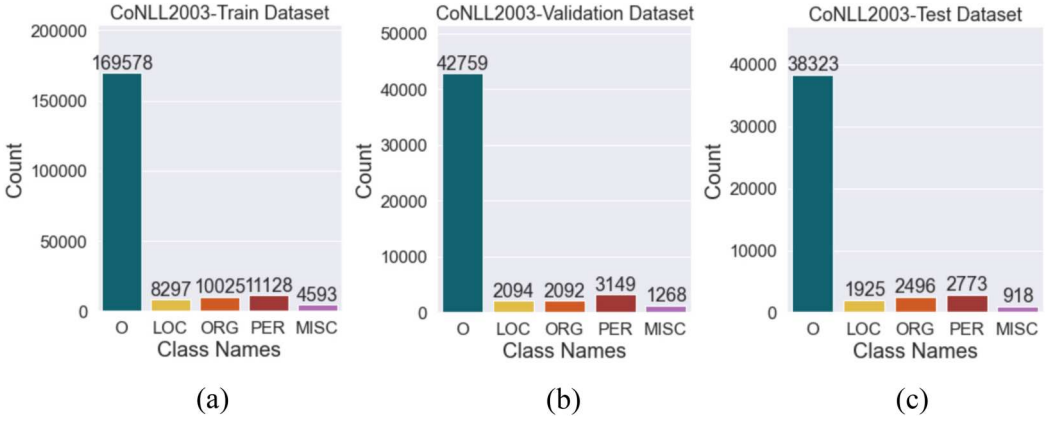

Figure 4. Data Distribution of CoNLL2003 Dataset.

systems. This dataset contains seven different groups: person, location, corporation, product, creative work, group, and none of the above. Considering the main focus of this paper, this dataset was preprocessed to retain only the location entities tag and to unify the tag symbols based on CoNLL2003. Location entities in WNUT2017 were tagged with 'B-LOC' or 'I-LOC' while the rest were tagged with 'O'. The total number of location names in this dataset is 1140 (9.7%).

Wiki_Gen is dataset generated to provide deep learning models a boarder understanding of the patterns of location names to enhance recall performance. The dataset fuses automatically generated content from Wikipedia articles by a data producing workflow proposed by Wang, Hu, and Joseph (2020) and commonly used place names from tweets. The proposed auto-annotation approach utilizes the first paragraph of Wikipedia articles, which usually encompasses various entities presented with hyperlinks. These hyperlinks are checked if they are associated with a geographical location. If so, the hyperlinked word is labeled as a toponym. Then the Wikipedia article is divided into multiple short sentences within 280 characters with additional strategies such as random flipping to mimic the general patterns of Twitter posts (Wang, Hu, and Joseph 2020). Generative deep learning models are used to generate additional tweet-like sentences containing location names. Around 37.2% entities in the dataset are location names.

Harvey2017 is a dataset originally collected from the University of North Texas repository, which contains 7,041,866 tweets collected based on hashtag query. It has been pruned, randomly subsampled and manually annotated to form a new dataset with 1000 tweets aiming to evaluate toponym recognition tools (Wang, Hu, and Joseph 2020). This dataset is adopted by this paper to test the performance of TopoBERT. A total of 3973 (20.6%) entities are labeled as locations in this dataset.

LouFlood2016 and HouFlood2015 are two datasets originally created in the work of Al-Olimat et al. (2018). The datasets are in JSON format and the place names and other texts are tagged with inLOC and outLOC. The tags are processed and transformed to the same data format as the aforementioned datasets. Location entities take up around 0.01% in both datasets.

NzEq2013 refers to data collected from the 2013 earthquake took place in New Zealand. The dataset was created by Middleton et al. (2018). The dataset is in JSON format and the place names are tagged with 'inLOC' and 'outLOC'. The tags are processed and transformed to the same data format as the aforementioned datasets. Location entities take up around 0.02% in this dataset.

GeoCorpora was created to evaluate geoparsing. The dataset was generated by Wallgrün et al. (2018) by harvesting data from a collection of events including earthquake, Ebola, wildfire, flood, rebel and so on from 2014 to 2015. The place names in the dataset cover continents, countries, states, cities, and admin units. The original labels are processed to match other datasets. Around 0.02% of the entities in this dataset represent locations.

Ritter's dataset obtained from Ritter et al. (2011) is a general-purpose NER dataset. The annotated named entities in this dataset include location, facility, person, and organization. The label regime for this dataset is in line with CoNLL2003. Only entities tagged with place names (0.01%) are used for evaluation in this study.

HumAID-1000 is large-scale dataset with around 77,000 manually annotated tweets sampled from a reservoir of 24 million tweets collected from 19 disastrous events that took place from 2016 to 2019 (Alam et al. 2021). The location entities are marked 'B-LOC' and 'I-LOC' and takes up approximately 0.05% of the entire dataset.

3.4. Model training

TopoBERT is envisioned to be a ready-to-use module that renders optimal performance in toponym recognition. Models with different architectures were trained and evaluated with the four datasets specified in Section 3.3 to determine the best model architecture and training strategy. The training process utilized CoNLL2003-Train as the training dataset by default and compared it to

another two larger datasets fusing CoNLL2003 with WNUT2017 or Wiki_Gen. The original dataset is labeled at the word level. It cannot be input to BERT directly due to BERT's word-piece encoding, which will lead to large numbers of out-of-vocabulary words. To tackle this issue, we split the input data at the word level and applied the BERT word-piece tokenizer to each word. The same label was assigned to each word-piece of a single word. The labeled word pieces were then merged to form the new input data which could be processed by BERT. This experiment aimed at measuring the performance fluctuations caused by training data size and heterogeneity. CoNLL2003-Validation was used during the training process to tune several fundamental hyperparameters such as training epochs and learning rate.

In the TopoBERT module, the parameters of the classifier component were initialized with random non-zero numbers, and the BERT component was initialized with pre-trained parameters. The entire module was trained with the fine-tuning approach (Devlin et al. 2018), and the parameters were updated using a mini-batch gradient descent approach with early stopping. The maximum length of the input sequence was limited to 128 in this experiment. The maximum number of training epochs was set to 50. As recommended by the original BERT paper, the initial learning rate and the training batch size were set to 2e−5 and 32, respectively (Devlin et al. 2018). The most commonly used cross-entropy loss was employed as the loss function for this multi-class classification task. AdamW was selected as the optimizer during training which adjusts the learning rate dynamically to accelerate parameter convergence and implements weight decay to lower the chance of overfitting. Warm up steps, which use a low learning rate for the first several weight updating iterations, were also introduced during training to reduce the impact of deviating the model drastically from sudden exposure to unseen datasets.

3.5. Evaluation and metrics

Once the training process is accomplished, TopoBERT is encapsulated as a ready to use package that requires no additional tuning. We select seven competitive baseline models based on the results from Hu et al. (2022b). These models are Stanford NLP (Finkel, Grenager, and Manning 2005), spaCy,[6] FlairNER,[7] nLORE (Fernández-Martínez and Periñán-Pascual 2021), Stanza (Qi et al. 2020), NeuroTPR (Wang, Hu, and Joseph 2020), and GazPNE2 (Hu et al. 2022c).

Three commonly used evaluation metrics, precision, recall, and F1-score (Equations (1)–(3)), were applied to gauge the performance and bias of the models. Precision calculates the percentage of correctly identified location names (noted as True Positives, TP) among all the location names predicted by the model, which combines both TP and False Positives (FP). Recall measures the percentage of correctly identified ones amongst all ground truth, which is the combination of TP and False Negatives (FN). F1-score is the harmonic mean of precision and recall, providing a comprehensive metric to evaluate model performance.

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$F1\text{–}score = 2* \frac{Precision* Recall}{Precision + Recall} \tag{3}$$

The outputs of BERT models are at the word-piece level, and they are concatenated using the special prefix '##'. The word-level labels are assigned based on the starting word-piece of the word. The evaluation metrics are based on 'per-token' scores. Additionally, the location name entity consists of two types of labels (B-LOC and I-LOC). For example, the correct label for 'Larmar St Houston Texas Tomorrow' should be 'B-LOC I-LOC I-LOC I-LOC 0'. If the model predicts 'B-

LOC I-LOC 0 0 I-LOC', then TP = 2, FP = 1, FN = 2. In order to gauge the comprehensive performance of the model on toponym recognition, the evaluation metrics were calculated using a micro average approach, which computes a global average of precision, recall, and F1-score. It calculates the TP, FP, and FN by counting the total number of TP, FP, and FN under each class, namely, 'B-LOC' and 'I-LOC'.

4. Analysis and results

4.1. Choosing pretrained parameters

The first step of the experiment targeted determining the optimal pretrained parameters for the BERT model. We hypothesize that larger models outperform smaller models. To verify this hypothesis, the performance of TopoBERT models initialized with 'bert-base-cased' and 'bert-large-cased' with a linear classifier stacked on top were tested. The results are displayed in Table 2.The ones with the best performance of each metrics are highlighted in bold.

These two models were trained with CoNLL2003-Train and evaluated with CoNLL2003-Test. Compared to 'bert-base-cased', the precision of the prediction increases from 0.900 to 0.934 by using 'bert-large-cased' while the recall almost remains static. The F1-scores show that 'bert-large-cased' renders better results which are in conformity with the original BERT paper (Devlin et al. 2018) and validates our hypothesis. Therefore, 'bert-large-cased' was harnessed in all the follow-up experiments.

4.2. Influence of training data

The second step of the experiments aims to determine the optimal classifier and measure the influence of the training data. The model performances were evaluated using two different datasets, CoNLL2003-Test and the testing data combining the seven databases listed in Table 1 (Harvey2017, LouFlood2016, HouFlood2015, NzEq2013, GeoCorpora, Ritter's Dataset, and HumAID-1000). We hypothesize that (a) the model with CNN1D classifier yields better results and (b) models trained with larger datasets perform better in place name recognition. Tables 3 lists the result of model performance trained with CoNLL-2003-Train and evaluated with CoNLL-2003-Test. Table 4 lists the evaluation metrics of all models with different training strategies with the seven datasets as testing datasets. The ones with the best performance of each metrics are highlighted in bold.

In Table 3, when models were trained with CoNLL2003-Train, the one with a simple linear classifier produced the best precision (0.934), and the one with CNN1D produced the best recall (0.920) and F1-score (0.921). MLP performed the worst among the three classifiers. In Table 4, the average performance of the trained models with different training data variations on seven datasets are presented. When models were trained with CoNLL2003-Train, the one with the linear classifier outperformed the rest with a precision equal to 0.545, recall of 0.684, and F1-score of 0.598. When

Table 2. Evaluation results with CoNLL2003-Test dataset for testing on different pretrained parameters.

BERT Model	Classifier	Precision	Recall	F1-score
bert-base-cased	Linear	0.900	**0.904**	0.902
bert-large-cased	Linear	**0.934**	0.901	**0.917**

Table 3. Evaluation results with CoNLL2003-Test dataset for testing on different classifier types.

Training Data	Classifier	Precision	Recall	F1-score
CoNLL2003-Train	Linear	**0.934**	0.901	0.917
CoNLL2003-Train	MLP	0.904	0.910	0.907
CoNLL2003-Train	CNN1D	0.923	**0.920**	**0.921**

Table 4. Evaluation results with seven datasets for testing on training data variation and classifier types.

Training Data	Classifier	Precision	Recall	F1-score
CoNLL2003-Train	Linear	**0.545**	0.684	**0.598**
CoNLL2003-Train	MLP	0.456	**0.696**	0.539
CoNLL2003-Train	CNN1D	0.519	0.646	0.576
CoNLL2003 + WNUT2017	Linear	0.518	**0.599**	0.551
CoNLL2003 + WNUT2017	MLP	0.581	0.456	0.506
CoNLL2003 + WNUT2017	CNN1D	**0.696**	0.492	**0.568**
CoNLL2003 + Wiki_Gen	Linear	0.802	0.758	0.778
CoNLL2003 + Wiki_Gen	MLP	0.790	0.768	0.777
CoNLL2003 + Wiki_Gen	CNN1D	**0.827**	**0.886**	**0.854**

models were trained with CoNLL2003-Train and WNUT2017, the model with CNN1D successfully defended its trophy by rendering a precision of 0.696, recall of 0.492, and F1-score of 0.568. The models with MLP worked slightly worse than the ones with linear classifiers. When the models were trained with CoNLL2003-Train and Wiki_Gen, the model with CNN1D rendered the best performance with precision of 0.827, recall of 0.886 and F1-score of 0.854. This can be explained by the attribute that CNN is proficient in capturing the underlying features of the input data which leads to more generalized models and when the models are evaluated with a plethora of datasets, model generalizability plays a major role. The above elucidation certifies the hypothesis that models with CNN1D generate optimal performance. It also shows that more complicated classifiers like multi-layer perceptron do not necessarily render better results than linear classifiers.

However, when viewing Table 4 regarding the results from training with different datasets, the metrics indicate that the model trained with the CoNLL2003-Train and WNUT2017 dataset generally performed worse than the ones trained with CoNLL2003-Train. This phenomenon contradicts the hypothesis that models trained with larger datasets perform better. However, when the models were trained with CoNLL2003-Train and Wiki_Gen, a significant performance boost can be observed. This is in alignment with the hypothesis and fulfills the purpose of using the additional dataset to enhance the model recall. After scrutinizing the dataset used for training, we noticed some inconsistencies in the labeling criteria of the datasets. Some examples are listed in Table 5 and the unexpected phenomenon can be interpreted by the heterogeneity of the datasets. For instance, the word 'Canadian' is labeled as 'B-MISC' (beginning of a miscellaneous name) in the CoNLL2003 dataset but is identified as 'B-LOC' (beginning of a location) in the WNUT2017 dataset. The words 'Planet' and 'east' are categorized as 'Others' in the CoNLL2003 and Wiki_Gen datasets but misclassified as locations in the WNUT2017 dataset. The phrase 'orchard academy,' regarded as an organization under the CoNLL2003 criteria, is labeled as 'Others' in Wiki_Gen and a location entity in WNUT2017. Consequently, combining several heterogeneous datasets can be considered as adding some helpful unseen samples to the original training data while introducing a substantial amount of noise.

Rolnick et al. (2017) experimented on several deep learning models when trained with noisy data and claimed that the CNN model is more resilient to noise than MLP and linear models. The trend of performance change shown in Table 4 when trained with different datasets is in accordance with this previous finding. It is noticeable that the models experience an increase in precision and a

Table 5. Examples of disparities amongst labels across the datasets used for training the model.

	Dataset		
Example Entity	CoNLL2003	Wiki_Gen	WNUT2017
'Canadian'	B-MISC	O	B-LOC
'Planet'	O	O	B-LOC
'east'	O	O	B-LOC
'orchard' 'academy'	B-ORG/I-ORG	O	B-LOC/I-LOC
'earth'	O	N/A	B-LOC

drastic decrease in recall when trained with a combined dataset. This incident can as well be triggered by noisy data. Since deep learning models attempt to learn the underlying patterns of the training data, the existing noise will confuse the model, resulting in a fewer number of positive predictions. This might result in an increase in precision and a decrease in recall.

4.3. Comparison with baseline models

Based on the observation and interpretation above, the BERT model initialized with 'bert-large-cased', stacked with a CNN1D classifier, and fine-tuned with CoNLL2003-Train and Wiki_Gen was selected as the finalized TopoBERT module. Table 6 shows a comparison between the optimal TopoBERT and seven other models and tools based on seven testing datasets. The ones with best performance of each metrics are highlighted in bold. The SpaCy version v3.2.1 is used with model 'en_core_web_sm' loaded. Broad location indicates that we include entities in both LOCATION and ORGANIZATION for Stanford NER 4.3.1, and we include entities in the types of LOC, ORG, FACILITY, and GPE for spaCy NER.

Evaluation results show that each model has its unique character and performs differently on different dataset. Stanford NER tend to render results with high precision but low recall, which means the majority of its positive predictions (place names) are correct. GazPNE2 performs stably with similar precision and recall scores. TopoBERT presents the highest recall scores on six out of the seven datasets tested. This indicates that TopoBERT is good at identifying the majority of positive instances. The average scores are computed among all datasets to compare the general performance of the models. Stanford NER has the highest precision score of 0.872 and TopoBERT prevail in both recall and F1-score with scores equal to 0.886 and 0.854, respectively. This result confirms that the proposed and trained TopoBERT model outperforms other baseline models by at least 9.1%.

TopoBERT has been developed as a ready-to-use module. The output data of TopoBERT include word labels and confidence in the prediction. It complies with JSON file format for ease of use. The source code has been uploaded to GitHub and can be accessed with the link: https://github.com/SPGBarrett/gearlab_topobert.

5. Discussion

This paper presents a geoparsing framework and breeds a plug-and-play toponym recognition module that can facilitate spatial analysis based on social media or news media data. Figure 5 shows a practical application of this framework in locating Twitter posts under fine-grained topics during hazardous events. The study area is the State of Florida, and the dots in multiple colors displayed on the map are tweets posted during Hurricane Irma harvested by Twitter developer Application Programming Interface (API). The red pins indicate humanitarian help requests. The orange dots represent animal related help requests. The green dots represent reports of infrastructure malfunction and the blue dots mean people offer help or provide shelter information. A total number of 6325 tweets are visualized on the map and the location information of 3269 (51.6%) of them are retrieved by TopoBERT and google geocoding service. The module also enjoys the potential of being used for location name detection for news media to pinpoint the discussed topics (Quezada, Peña-Araya, and Poblete 2015; Zhou and Luo 2012) and help identify fake news (Shu et al. 2019).

This paper concentrates mainly on designing a novel architecture of a reliable and versatile module for toponym recognition. However, the performance enhancement can continue by addressing the following issues.

First, the models are trained and evaluated based on well prepared datasets. This can be regarded as a best-case scenario compared to real life situations. Place name usage can be highly ambiguous and random, especially within social media platforms. Typos are common and might cause out-of-vocabulary words in language models. Place name abbreviations such as 'Boulevard' and 'blvd', 'Drive' and 'Dr.', and 'Street' and 'St.' are frequently utilized interchangeably on social media

Table 6. Evaluation results with seven datasets for comparing TopoBERT with seven baseline models.

		Harvey2017	LouFlood2016	HouFlood2015	NzEq2013	GeoCorpora	Ritter's dataset	HumAID-1000	Average
Stanford NER 4.3.1	Precision	**0.861**	**0.942**	0.927	0.852	0.797	0.767	**0.893**	**0.863**
	Recall	0.455	0.335	0.210	0.499	0.551	0.424	0.409	0.412
	F1-score	0.595	0.495	0.342	0.629	0.652	0.546	0.562	0.546
spaCy 3.2.1	Precision	0.840	0.917	0.896	0.790	0.732	0.609	0.857	0.806
	Recall	0.510	0.360	0.221	0.657	0.564	0.497	0.434	0.463
	F1-score	0.635	0.517	0.355	0.717	0.637	0.547	0.576	0.569
FlairNER	Precision	0.777	0.933	0.847	0.603	0.731	0.705	0.815	0.773
	Recall	0.643	0.638	0.248	0.689	0.692	0.505	0.546	0.566
	F1-score	0.704	0.757	0.384	0.643	0.711	0.588	0.654	0.634
nLORE	Precision	0.762	0.848	0.823	0.599	0.650	0.491	0.827	0.714
	Recall	0.717	0.779	0.640	0.653	0.673	0.501	0.690	0.665
	F1-score	0.739	0.812	0.720	0.625	0.661	0.496	0.752	0.686
NeuroTPR	Precision	0.801	0.866	0.817	0.602	0.581	0.558	0.736	0.709
	Recall	0.708	0.749	0.312	0.754	0.726	**0.969**	0.612	0.690
	F1-score	0.752	0.803	0.451	0.669	0.646	0.708	0.668	0.671
GazPNE2	Precision	0.817	**0.942**	**0.942**	0.690	0.724	0.637	0.872	0.803
	Recall	**0.835**	0.808	0.710	0.828	0.827	0.638	0.825	0.782
	F1-score	**0.826**	0.870	**0.810**	0.753	0.772	0.637	0.848	0.788
Stanza 1.2	Precision	0.728	0.937	0.845	0.769	0.737	0.657	0.809	0.783
	Recall	0.628	0.623	0.272	0.674	0.673	0.506	0.554	0.561
	F1-score	0.675	0.749	0.412	0.719	0.704	0.572	0.658	0.641
TopoBERT	Precision	0.843	0.823	0.763	**0.867**	**0.808**	**0.812**	0.872	0.827
	Recall	0.794	**0.924**	**0.863**	**0.922**	**0.944**	0.804	**0.948**	**0.886**
	F1-score	0.818	**0.871**	**0.810**	**0.894**	**0.871**	**0.808**	**0.908**	**0.854**

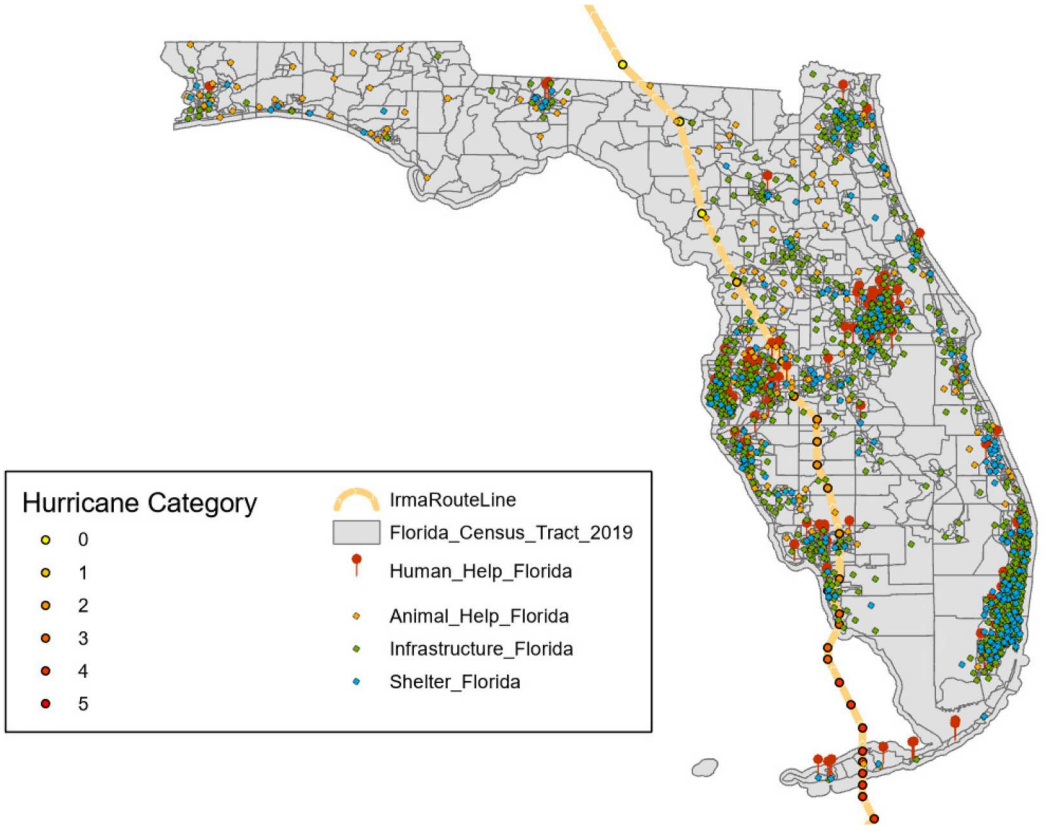

Figure 5. Toponym recognition applied to locate Twitter posts during 2017 Hurricane Irma.

and news media. People might unconsciously ignore the correct upper-case and lower-case usage, such as 'college station' and 'College Station', 'mexico' and 'MEXICO', etc. Meticulous data preprocessing methods can be incorporated to tackle this problem to achieve better overall performance. Second, several rule-base approaches can be leveraged to further boost the performance. Based on the results of the evaluation, TopoBERT renders a higher recall than precision, which means the algorithm tend to produce more False Positive predictions. Scrutinizing the predictions, we notice that texts such as 'CHICAGO VIOLENCE' and 'HOUSTON FLOOD' are both predicted as 'B-LOC I-LOC', which produce False Positives. Enlightened by the success of hybrid models (Hu et al. 2022a; Hu et al. 2022c), sets of grammar rules based on the composition of nouns, determiners, adjectives, conjunctions, numbers, and possessive endings can be designed (Giridhar et al. 2015). Additionally, commonly used gazetteers such as OpenStreetMap and GeoNames can be used as extra named entity matching criteria which will enhance the True Positives of the model. Regional criteria can be appended to the model while identifying place names by making country names, state names, county names, or bounding boxes as input variables of the model. This will allow the model to add constraints during the location recognition processing. The top-N words from word embedding models (Hu et al. 2022c; Mikolov et al. 2013), which are not place names, can be applied to filter words during data preprocessing. This will lessen the False Positives of the prediction.

Third, due to the data-hungry nature of deep learning, data availability and quality are topics being inevitably discussed when large complicated deep learning models are involved. Although larger datasets normally lead to better generalizability and performance in training deep learning models, this statement does not always hold in this research because the larger datasets are derived

from several distinguished smaller datasets labeled under unique regimes. Therefore, there is an urgent need to define criteria and build unified datasets for toponym recognition model training, evaluating, and benchmarking. The dataset can be manually modified based on existing datasets and augmented using rule-based methods, gazetteers or Generative Adversarial Network (Cao and Lee 2020; Feng et al. 2021; Shorten, Khoshgoftaar, and Furht 2021).

Fourth, fine-tuned language models can be few-shot or zero-shot learners, which means that the models can be applied directly to certain downstream tasks with very little or even no further training (Pushp and Srivastava 2017; Wei et al. 2021; Wortsman et al. 2021). This is because advanced language models can better capture the meaning of the text. This claim is also underpinned by the result of this paper which leverages BERT to boost the module capability. Therefore, incorporating gigantic models such as GPT-3 (Brown et al. 2020) might lead to another round of performance enhancement.

6. Conclusion

To further enhance the performance of toponym recognition by better understanding natural language, TopoBERT, a uniquely designed model incorporating the pretrained language model, BERT, and CNN1D is introduced. Experiments on the pretrained parameters, training dataset combinations, and model architectures reveal the following findings. First, the performance of toponym recognition models empowered by pre-trained language models is sensitive to the architecture of language models and classifiers. The TopoBERT models initialized with a larger-structured BERT model ('bert-large-cased') show an advantage over the models initialized with a basic BERT model ('bert-base-cased'). More complicated classifiers like MLP do not necessarily win over simple linear classifiers. Second, increasing training data size may produce worse results, especially for the recall, due to data heterogeneity. The model trained with the combination of CoNLL2003-Train and Wiki_Gen, and stacked on top with a CNN1D classifier renders the optimum results on both testing datasets (CoNLL2003-Test and the testing dataset combining seven databases). Finally, the developed TopoBERT module outperforms existing models in recognizing place names in texts. The clinched TopoBERT with the optimal model architecture and training strategy produces reliable toponym prediction and achieves an average F1-score of 0.854 on seven datasets, which surpasses other prevailing models or tools by at least 9.1%.

In a nutshell, the discoveries of this paper contribute to the subsequent directions and facilitate future studies. First, the model verifies that by incorporating advanced language models, the performance of toponym recognition can be boosted, which will lead to additional geolocated social and news media data for spatiotemporal analysis and generate more reliable and representative results. Second, the model is experimented with different architecture and training strategies to determine the optimal model structure and evaluated on unseen datasets to validate its generalizability. This investigation also urges a large, standardized dataset labeled with a unified regime to support toponym recognition model training and benchmarking. Third, by pinpoint locations from textual content of the social media, extra information that geotags fail to offer can be mined, e.g. people who post messages to request help for their family members or friends. Finally, a plug-and-play module is implemented and open-sourced to support pertinent applications and similar research. Other studies can leverage the developed TopoBERT module to harvest location-based information from textual data with few or no training datasets.

Notes

1. https://huggingface.co/dslim/bert-base-NER.
2. https://www.nltk.org/.
3. https://scikit-learn.org/.
4. https://huggingface.co/transformers/.

5. https://pytorch.org/.
6. https://spacy.io/.
7. https://huggingface.co/flair/ner-english.

Acknowledgements

We thank all anonymous reviewers for their insightful comments that greatly improved the manuscript.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The research is supported by two projects funded by the National Science Foundation in the U.S.: (1) Reducing the Human Impacts of Flash Floods – Development of Microdata and Causal Model to Inform Mitigation and Preparedness (Award No. 1931301) and (2) Geospatial Artificial Intelligence Approaches for Understanding Location Descriptions in Natural Disasters and Their Spatial Biases (Award No. 2117771).

Data availability statement

The data that support the findings of this study are openly available in github at https://github.com/SPGBarrett/gearlab_topobert.

References

Abdul-Rahman, M., E. H. Chan, M. S. Wong, V. E. Irekponor, and M. O. Abdul-Rahman. 2021. "A Framework to Simplify pre-Processing Location-Based Social Media big Data for Sustainable Urban Planning and Management." *Cities* 109:102986. https://doi.org/10.1016/j.cities.2020.102986.

Al-Olimat, H., K. Thirunarayan, V. Shalin, and A. Sheth. 2018. "Location Name Extraction from Targeted Text Streams using Gazetteer-based Statistical Language Models." In *Proceedings of the 27th International Conference on Computational Linguistics, 1986–1997*. https://aclanthology.org/C18-1169.

Alam, F., U. Qazi, M. Imran, and F. Ofli. 2021. "HumAID: Human-Annotated Disaster Incidents Data from Twitter with Deep Learning Benchmarks." *Proceedings of the International AAAI Conference on Web and Social Media* 15:933–942. https://doi.org/10.1609/icwsm.v15i1.18116.

Avvenuti, M., S. Cresci, M. N. La Polla, A. Marchetti, and M. Tesconi. 2014. "Earthquake Emergency Management by Social Sensing." In *2014 IEEE International Conference on Pervasive Computing and Communication Workshops (PERCOM WORKSHOPS)*, 587–592. Budapest, Hungary: IEEE.

Bahdanau, D., K. Cho, and Y. Bengio. 2014. *Neural Machine Translation by Jointly Learning to Align and Translate*. Preprint, arXiv:1409.0473.

Brown, T., B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, … D. Amodei. 2020. "Language Models are few-Shot Learners." *Advances in Neural Information Processing Systems* 33:1877–1901. https://doi.org/10.48550/arXiv.2005.14165.

Cao, R., and R. K. W. Lee. 2020. "Hategan: Adversarial Generative-Based Data Augmentation for Hate Speech Detection." In *Proceedings of the 28th International Conference on Computational Linguistics*, 6327–6338.

Cardoso, A. B., B. Martins, and J. Estima. 2022. "A Novel Deep Learning Approach Using Contextual Embeddings for Toponym Resolution." *ISPRS International Journal of Geo-Information* 11 (1): Article 1. https://doi.org/10.3390/ijgi11010028.

Cervone, G., E. Schnebele, N. Waters, M. Moccaldi, and R. Sicignano. 2017. "Using Social Media and Satellite Data for Damage Assessment in Urban Areas During Emergencies." In *Seeing Cities Through big Data*, edited by Piyushimita (Vonu) Thakuriah, Nebiyou Tilahun, and Moira Zellner, 443–457. Tilahun: Springer.

Derczynski, L., E. Nichols, M. van Erp, and N. Limsopatham. 2017. "Results of the WNUT2017 Shared Task on Novel and Emerging Entity Recognition." In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, 140–147.

Devlin, J., M. W. Chang, K. Lee, and K. Toutanova. 2018. *Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding*. Preprint, arXiv:1810.04805.

Dutt, R., K. Hiware, A. Ghosh, and R. Bhaskaran. 2018. "Savitr: A System for Real-Time Location Extraction from Microblogs During Emergencies." In *Companion Proceedings of the Web Conference 2018*, 1643–1649. https://doi.org/10.1145/3184558.3191623.

Feng, S. Y., V. Gangal, J. Wei, S. Chandar, S. Vosoughi, T. Mitamura, and E. Hovy. 2021. *A Survey of Data Augmentation Approaches for NLP*. Preprint, arXiv:2105.03075.

Fernández-Martínez, N. J., and C. Periñán-Pascual. 2021. "nLORE: A Linguistically Rich Deep-Learning System for Locative-Reference Extraction in Tweets." In *Intelligent environments 2021: Workshop proceedings of the 17th international conference on intelligent environments*. Vol. 29, edited by Engie Bashir and Mitja Luštrek, 243. Dubai: IOS Press. https://doi.org/10.3233/AISE210103.

Finkel, J. R., T. Grenager, and C. D. Manning. 2005. "Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling." In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, 363–370.

Giridhar, P., T. Abdelzaher, J. George, and L. Kaplan. 2015. "On Quality of Event Localization from Social Network Feeds." In *2015 IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops)*, 75–80. St. Louis, MO: IEEE.

Hochreiter, S., and J. Schmidhuber. 1997. "Long Short-Term Memory." *Neural Computation* 9 (8): 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.

Hu, X., H. S. Al-Olimat, J. Kersten, M. Wiegmann, F. Klan, Y. Sun, and H. Fan. 2022a. "GazPNE: Annotation-Free Deep Learning for Place Name Extraction from Microblogs Leveraging Gazetteer and Synthetic Data by Rules." *International Journal of Geographical Information Science* 36 (2): 310–337. https://doi.org/10.1080/13658816.2021.1947507.

Hu, X., Z. Zhou, H. Li, Y. Hu, F. Gu, J. Kersten, H. Fan, and F. Klan. 2022b. *Location Reference Recognition from Texts: A Survey and Comparison*. arXiv:2207.01683. https://doi.org/10.48550/arXiv.2207.01683

Hu, X., Z. Zhou, Y. Sun, J. Kersten, F. Klan, H. Fan, and M. Wiegmann. 2022c. "GazPNE2: A General Place Name Extractor for Microblogs Fusing Gazetteers and Pretrained Transformer Models." *IEEE Internet of Things Journal* 9 (17): 16259–16271. https://doi.org/10.1109/JIOT.2022.3150967.

Huang, X., Z. Li, Y. Jiang, X. Li, and D. Porter. 2020. "Twitter Reveals Human Mobility Dynamics During the COVID-19 Pandemic." *PloS one* 15 (11): e0241957. https://doi.org/10.1371/journal.pone.0241957.

Kiranyaz, S., O. Avci, O. Abdeljaber, T. Ince, M. Gabbouj, and D. J. Inman. 2021. "1D Convolutional Neural Networks and Applications: A Survey." *Mechanical Systems and Signal Processing* 151:107398. https://doi.org/10.1016/j.ymssp.2020.107398.

Labusch, K., P. Kulturbesitz, C. Neudecker, and D. Zellhöfer. 2019. "BERT for named entity recognition in contemporary and historical German." In *Proceedings of the 15th Conference on Natural Language Processing, Erlangen, Germany*, 8–11.

Lan, Z., M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut. 2020. *ALBERT: A Lite BERT for Self-supervised Learning of Language Representations*. arXiv:1909.11942. https://doi.org/10.48550/arXiv.1909.11942

Lee, J. Y., and F. Dernoncourt. 2016. *Sequential Short-text Classification with Recurrent and Convolutional Neural Networks*. Preprint, arXiv:1603.03827.

Li, W., C.-Y. Hsu, and M. Hu. 2021. "Tobler's First Law in GeoAI: A Spatially Explicit Deep Learning Model for Terrain Feature Detection Under Weak Supervision." *Annals of the American Association of Geographers* 111 (7): 1887–1905. https://doi.org/10.1080/24694452.2021.1877527.

Lin, B., L. Zou, N. Duffield, A. Mostafavi, H. Cai, B. Zhou, J. Tao, M. Yang, D. Mandal, and J. Abedin. 2022. "Revealing the Linguistic and Geographical Disparities of Public Awareness to Covid-19 Outbreak Through Social Media." *International Journal of Digital Earth* 15 (1): 868–889. https://doi.org/10.1080/17538947.2022.2070677.

Liu, Z., K. Janowicz, L. Cai, R. Zhu, G. Mai, and M. Shi. 2022. "Geoparsing: Solved or Biased? An Evaluation of Geographic Biases in Geoparsing." *AGILE: GIScience Series* 3:1–13. https://doi.org/10.5194/agile-giss-3-9-2022.

Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. 2019. *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. arXiv:1907.11692. https://doi.org/10.48550/arXiv.1907.11692.

Luoma, J., and S. Pyysalo. 2020. *Exploring Cross-sentence Contexts for Named Entity Recognition with BERT*. arXiv:2006.01563. https://doi.org/10.48550/arXiv.2006.01563.

Ma, K., Y. Tan, Z. Xie, Q. Qiu, and S. Chen. 2022. "Chinese toponym recognition with variant neural structures from social media messages based on BERT methods." *Journal of Geographical Systems* 24 (2): 143–169. https://doi.org/10.1007/s10109-022-00375-9.

Middleton, S. E., G. Kordopatis-Zilos, S. Papadopoulos, and Y. Kompatsiaris. 2018. "Location Extraction from Social Media: Geoparsing, Location Disambiguation, and Geotagging." *ACM Transactions on Information Systems* 36 (4): 1–27. https://doi.org/10.1145/3202662.

Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. 2013. "Distributed Representations of Words and Phrases and Their Compositionality." *Advances in Neural Information Processing Systems* 26. https://doi.org/10.48550/arXiv.1310.4546.

Milusheva, S., R. Marty, G. Bedoya, S. Williams, E. Resor, and A. Legovini. 2021. "Applying Machine Learning and Geolocation Techniques to Social Media Data (Twitter) to Develop a Resource for Urban Planning." *PloS One* 16 (2): e0244317. https://doi.org/10.1371/journal.pone.0244317.

Peters, M., M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. 2018. "Deep Contextualized Word Representations." In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2227–2237. New Orleans, Louisiana: Association for Computational Linguistics.

Pushp, P. K., and M. M. Srivastava. 2017. *Train Once, Test Anywhere: Zero-shot Learning for Text Classification*. Preprint, arXiv:1712.05972.

Qi, P., Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning. 2020. *Stanza: A Python Natural Language Processing Toolkit for Many Human Languages*. Preprint, arXiv:2003.07082.

Quezada, M., V. Peña-Araya, and B. Poblete. 2015. "Location-Aware Model for News Events in Social Media." In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 935–938.

Radford, A., J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. 2019. "Language Models are Unsupervised Multitask Learners." *OpenAI Blog* 1 (8): 9. https://doi.org/10.48550/arXiv.2005.14165.

Ritter, A., S. Clark, Mausam, and O. Etzioni. 2011. "Named Entity Recognition in Tweets: An Experimental Study." In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 1524–1534. https://aclanthology.org/D11-1141.

Rolnick, D., A. Veit, S. Belongie, and N. Shavit. 2017. *Deep Learning Is Robust to Massive Label Noise*. Preprint, arXiv:1705.10694.

Sang, Erik F., Tjong Kim, and F. Meulder. 2003. *Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition*. https://doi.org/10.48550/arXiv.cs/0306050.

Sanh, V., L. Debut, J. Chaumond, and T. Wolf. 2020. *DistilBERT, A Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter*. arXiv:1910.01108. https://doi.org/10.48550/arXiv.1910.01108.

Shorten, C., T. M. Khoshgoftaar, and B. Furht. 2021. "Text Data Augmentation for Deep Learning." *Journal of Big Data* 8 (1): 101. https://doi.org/10.1186/s40537-021-00492-0.

Shu, K., X. Zhou, S. Wang, R. Zafarani, and H. Liu. 2019. "The Role of User Profiles for Fake News Detection." In *Proceedings of the 2019 IEEE/ACM international conference on advances in social networks analysis and mining*, 436–439.

Souza, F., R. Nogueira, and R. Lotufo. 2020. *Portuguese Named Entity Recognition using BERT-CRF*. arXiv:1909.10649. https://doi.org/10.48550/arXiv.1909.10649.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, … I. Polosukhin. 2017. "Attention is all you Need." *Advances in Neural Information Processing Systems* 30. https://doi.org/10.48550/arXiv.1706.03762.

Wallgrün, J. O., M. Karimzadeh, A. M. MacEachren, and S. Pezanowski. 2018. "GeoCorpora: Building a Corpus to Test and Train Microblog Geoparsers." *International Journal of Geographical Information Science* 32 (1): 1–29. https://doi.org/10.1080/13658816.2017.1368523.

Wang, J., Y. Hu, and K. Joseph. 2020. "NeuroTPR: A Neuro-net Toponym Recognition Model for Extracting Locations from Social Media Messages." *Transactions in GIS* 24 (3): 719–735. https://doi.org/10.1111/tgis.12627.

Wang, S., X. Zhang, P. Ye, and M. Du. 2018. "Deep Belief Networks Based Toponym Recognition for Chinese Text." *ISPRS International Journal of Geo-Information* 7 (6): Article 6. https://doi.org/10.3390/ijgi7060217.

Wei, J., M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, … Q. V. Le. 2021. *Finetuned Language Models Are Zero-shot Learners*. Preprint, arXiv:2109.01652.

Wortsman, M., G. Ilharco, J. W. Kim, M. Li, S. Kornblith, R. Roelofs, R. G. Lopes, et al. 2021. *Robust Fine-Tuning of Zero-Shot Models*. https://arxiv.org/pdf/2109.01903.

Zhao, J., X. Mao, and L. Chen. 2019. "Speech Emotion Recognition Using Deep 1D & 2D CNN LSTM Networks." *Biomedical Signal Processing and Control* 47:312–323. https://doi.org/10.1016/j.bspc.2018.08.035.

Zhou, Y., and J. Luo. 2012. "Geo-location Inference on News Articles via Multimodal pLSA." In *MM'12: The Proceedings of the 20th ACM International Conference on Multimedia, co-Located with ACM Multimedia 2012, October 29-November 2, 2012, Nara, Japan*, edited by N. Babaguchi, K. Aizawa, J. Smith, S. Satoh, T. Plagemann, X.-S. Hua, and R. Yan, 741. Nara, Japan: Association for Computer Machinery. https://doi.org/10.1145/2393347.2396301.

Zhou, B., L. Zou, A. Mostafavi, B. Lin, M. Yang, N. Gharaibeh, H. Cai, J. Abedin, and D. Mandal. 2022. "VictimFinder: Harvesting Rescue Requests in Disaster Response from Social Media with BERT." *Computers, Environment and Urban Systems* 95:101824. https://doi.org/10.1016/j.compenvurbsys.2022.101824.

Zou, L., N. S. N. Lam, H. Cai, and Y. Qiang. 2018. "Mining Twitter Data for Improved Understanding of Disaster Resilience." *Annals of the American Association of Geographers* 108 (5): 1422–1441. https://doi.org/10.1080/24694452.2017.1421897.

Zou, L., N. S. N. Lam, S. Shams, H. Cai, M. A. Meyer, S. Yang, K. Lee, S.-J. Park, and M. A. Reams. 2019. "Social and Geographical Disparities in Twitter use During Hurricane Harvey." *International Journal of Digital Earth* 12 (11): 1300–1318. https://doi.org/10.1080/17538947.2018.1545878.

Zou, L., D. Liao, N. S. Lam, M. Meyer, N. G. Gharaibeh, H. Cai, B. Zhou, and D. Li. 2023. "Social Media for Emergency Rescue: An Analysis of Rescue Requests on Twitter During Hurricane Harvey." *International Journal of Disaster Risk Reduction* 85:103513. https://doi.org/10.1016/j.ijdrr.2022.103513.