Leave no Place Behind: Improved Geolocation in Humanitarian Documents

Enrico M. Belliardo
enrico.m.belliardo@gmail.com
ISI Foundation
Turin, Italy

Kyriaki Kalimeri
kyriaki.kalimeri@isi.it
ISI Foundation
Turin, Italy

Yelena Mejova
yelenamejova@acm.org
ISI Foundation
Turin, Italy

ABSTRACT

Geographical location is a crucial element of humanitarian response, outlining vulnerable populations, ongoing events, and available resources. Latest developments in Natural Language Processing may help in extracting vital information from the deluge of reports and documents produced by the humanitarian sector. However, the performance and biases of existing state-of-the-art information extraction tools are unknown. In this work, we develop annotated resources to fine-tune the popular Named Entity Recognition (NER) tools Spacy and roBERTa to perform geotagging of humanitarian texts. We then propose a geocoding method FeatureRank which links the candidate locations to the GeoNames database. We find that not only does the humanitarian-domain data improves the performance of the classifiers (up to F1 = 0.92), but it also alleviates some of the bias of the existing tools, which erroneously favor locations in the Western countries. Thus, we conclude that more resources from non-Western documents are necessary to ensure that off-the-shelf NER systems are suitable for the deployment in the humanitarian sector.

CCS CONCEPTS

• **Social and professional topics** → **Geographic characteristics**; • **Information systems** → *Data mining*; *Information systems applications*.

KEYWORDS

Geolocation, humanitarian, location bias, geotagging, geographic bias, name entity recognition, geocoding

1 INTRODUCTION

The vast network of humanitarian organizations performing a myriad of tasks around the world produces a significant amount of data and reports. For instance, the operations of the International Federation of Red Cross and Red Crescent Societies (IFRC) spans 192 countries and includes almost 14 million volunteers [1]. To manage the resulting reports, projects such as the Data Entry and Exploration Platform (DEEP)[1] have been developed to help humanitarian organizations to compile and store documentation, and to help structure the qualitative information inside. It is maintained by Data Friendly Space, a U.S.-based international non-governmental organization (INGO)[2], in collaboration with numerous international organizations.

In the light of the information overload generated by the data-rich operation of humanitarian efforts, computer-aided information extraction promises to increase the usefulness of the produced documentation. Latest developments in Deep Learning (DL) and Natural Language Processing (NLP) allow for the identification of potentially important information in text, and its classification in standard hierarchies, which can then be used to aggregate knowledge and share experience across time and projects. In particular, geolocation is an important aspect of humanitarian activity, spanning both operations in entire countries and regions, to specific villages, refugee camps, and civil infrastructure. The accuracy of location information is especially important in the light of the Sustainable Development Goals (SDGs) principle of "Leave No One Behind" that strives to reduce inequalities and vulnerabilities amongst all countries. Unfortunately, the data sources available for training competent models are often biased towards the West, even when secondary data is used to fill the "datasphere" gaps [6]. Location databases (or "gazetteers") have a well-known bias towards the US and other Western countries [9, 10], whereas popular alternative data sources such as Twitter and Wikipedia struggle with penetration in global South [31, 33]. Thus, it is imperative to create resources and tools for the accurate processing of diverse humanitarian data in order to ensure no country is disadvantaged when the information is collated.

In this study, we work closely with humanitarian partners to create a Geolocation Extraction tool tuned specifically to process documents generated by humanitarian projects. The tool performs two sub-tasks: **geotagging** – the extraction of text fragments that may be a location (or "toponyms"), and **geocoding** – the disambiguation of the toponym to a specific geographic location (with accompanying information on GPS coordinates, type of location, etc.). We contribute two datasets – one for each of these steps – wherein humanitarian reports are annotated by specialists in the field for candidate toponyms, and these toponyms are mapped to

---

[1]https://thedeep.io/
[2]https://datafriendlyspace.org/

the appropriate entry in GeoNames, an extensive geographical database[3]. Using these datasets, we tune popular existing Named Entity Extraction (NER) models Spacy[4] and roBERTa[5], improving their performance up to F1 = 0.92 on the target dataset. We then propose a geocoding algorithm FeatureRank tailored to the humanitarian domain, and compare it to two baselines from the literature. Finally, we show that the tuned model not only improves in accuracy, but that it may alleviate the geographic bias that favors locations in the Western countries.

2 RELATED LITERATURE

Named Entity Recognition (NER) is one of the basic Natural Language Processing (NLP) tasks involving the identification of entities of interest in texts, commonly constrained to Person, Organisation or Location. Whereas the early models used classic machine learning algorithms including support vector machines, random forests, and decision trees using as features capitalization, word endings, and regular expressions [22], from about 2011 neural networks have been used to create more generalizable models [5, 35]. Combining word context and its individual characters has allowed the training of systems that required little domain-specific training data. Word and sentence level embeddings have been used to build models using several long short-term memory (LSTM) layers to achieve performance of up to 90% F score in English [7] (this model was also published as a NeuroNER package[6]). Graph embedding algorithms that produce embeddings of vertices that preserve their proximity in a graph, including DeepWalk [23] and TransE [2], have been widely used in entity linking (EL) task, the task of matching a piece of text (a potential entity) to an entry in a knowledge base such as Freebase, DBpedia, and Wikidata. Since 2015, large pre-trained models like Bidirectional Encoder Representations from Transformers (BERT) [15] have given researchers the access to representations trained on huge amounts of text data without the need to access this data directly, while allowing strong expressive power suitable to many NLP tasks. These models can then be trained using the distant labels, improving task-specific recall and precision, and further enhanced using self-training [18]. Neural networks and pre-trained large models have also been used for entity linking [28]. For instance, BERT-based cross-encoders have been used on a concatenation of text segment context with a candidate entity description to produce a score for each entity candidate [19, 34]. However, few of these efforts have been directed specifically to geographical NER.

Natural language processing approaches have been applied to understand the context in the extraction and geocoding of historical floods, storms, and adaptation measures [16]. In the context of mass emergencies, Imran et al. [13] highlighted the existing approaches and challenges in processing social media messages during mass emergencies. This survey sheds light on the importance of real-time analysis of social media data for effective emergency response. Several techniques were explored to extract and refine location mentions in text such as geoparsing, location disambiguation, and geotagging [21]. For instance, geotagging text data on the web has been approached through geometrical methods [25], offering an alternative to relying solely on explicit geotags. The potential of Twitter data were highlighted both in determining the geographic origin of user-generated content [11], and developing predictive models that can estimate the location of Twitter users based on their posted content [36]. Moreover, mining Twitter data offered a better understanding of disaster resilience [37], while was also proven effective in event classification and location prediction during disasters [30]. Importantly, efforts have been made to develop annotated datasets and resources [14, 24] contributing to the development and evaluation of algorithms for the identification and monitoring of internal displacement, but also sentiment analysis, named entity recognition, and geolocation from textual data. In this work, we contribute annotated resources concerning the humanitarian reports, instead of general web data.

Despite these advancements there are still many open issues to accurate named entity recognition for the humanitarian sector. For instance, known biases of the geographic information and modes of communication are directly impacting the representation of populations unevenly [10]. Population bias in geotagged tweets has been a topic of concern in geolocation research [20], highlighting the potential biases that can arise from relying solely on geotagged tweets for location inference. Understanding and addressing these biases is crucial to ensure accurate geolocation results.

3 DATA COLLECTION

We employ data from the HumSet database [8], originated from the multi-organizational platform (DEEP) introduced earlier. Each entry consists of a relevant excerpt from a document, annotated with the humanitarian analysis framework categories. It is associated to a *lead_id* that identifies the original document (*lead*) from which the excerpt is extracted. Original documents come from different sources, including reports by humanitarian organizations and media articles. We downloaded them using *deepex* package[7] from the links available in HumSet. The documents come either in PDF format or HTML pages, and text extraction from both sources is performed with the same package. This parser also splits text in pages and paragraphs, according to the basic structure of the document. In this project, we use the entire text parsed from HumSet *leads*, and not only the selection of excerpts of the database.

3.1 Data description

The dataset contains 15 661 documents from 45 different projects. It is multilingual, with English (67.1%), Spanish (18.6%) and French (14.1%) representing almost the totality of the documents' languages (language information is available in HumSet). Each project typically focuses on one or more countries, but some focus on none. There are 33 countries represented by country-specific projects. The content of the documents may vary from text to images and tables. The parser is able to fetch plain text and discard figures and non textual elements. Some documents contain mainly non textual elements, therefore the resulting extracted text is very short. Table 1 shows a distribution of metrics of the text content extracted with the parser.

---

[3]https://www.geonames.org/
[4]https://spacy.io/models/en#en_core_web_md
[5]https://huggingface.co/Davlan/xlm-roberta-base-wikiann-ner
[6]http://neuroner.com/

[7]available at https://github.com/the-deep/deepex

Table 1: Summary statistics on text files parsed from documents.

	size(KB)	pages	paragraphs	words
mean	11.6	5.0	38.0	1621
std	40.5	14.5	130.9	5487
min	0.0	1	0	0
25%	1.2	1	2	170
50%	3.2	1	8	459
75%	7.6	3	26	1085
max	1354.2	358	4254	207 030

4 DATA ANNOTATION

This work contributes two annotated datasets, one for each step of the geolocation extraction process: (1) **geotagging**, which is a special case of NER, and (2) **geocoding**, where the found toponyms are disambiguated and linked to specific geographic coordinates [12]. For both tasks, we describe how the data was sampled, baseline models were applied to extract candidates, how these candidates were annotated, guided by coding schemas. Finally, we present the statistics about the resulting annotated datasets.

4.1 Annotation: geotagging

4.1.1 Sampling. We begin by selecting 500 English-language documents from HumSet, satisfying several metrics, as described below. As we want our sample to include as many different locations as possible, while respecting the country distribution ranking of the dataset, we treat the documents belonging to the 6 most popular project countries differently from the rest (the "tail" of the distribution). We impose different restrictions on the popular and tail countries in terms of on the number of pages (minimum of 3 or 2, respectively), average number of paragraphs per page (minimum 5 or 3) and average number of words per paragraph (minimum 500 or 300). This filter results in 1897 documents. We further sample up to 100 documents from each project country, to undersample popular countries, resulting in 1290 documents, which we then sample randomly to get the final 500 document sample.

4.1.2 Pre-annotation using baseline models. Labelling task is performed with Label Studio app[8]. With it, is possible to upload pre-annotated documents, reducing the effort of the user to a simple revision of the pre-annotations. In order to insure a variety of locations in our dataset, we truncate the documents at 4000 characters, thus encouraging the annotators to label more distinct documents. To further help annotators find the candidate toponyms, we considered pre-annotations which may ease the annotation process, letting the user to focus on difficult cases without the need to tag all the locations in the text. To explore this, we run a small test experiment prior the full annotation process involving 4 users and 25 pilot documents. Two users were asked to tag pre-annotated documents, the other 2 – documents without pre-annotations. Annotation agreement does not vary among users with pre-annotated documents and users without pre-annotated documents (remaining at around 0.90), however the speed of their work improves by 7%

(and this was during an adjustment period where some aspects of the process were unclear to the labelers). Thus, we proceed to create candidate toponyms using baseline NER models.

Pre-annotations are provided by two off-the-shelf NER models: Spacy *en_core_web_md*[9], a pre trained pipeline for English that includes NER components; and the roBERTa *xlm-roberta-base-wikiann-ner*[10], a multilingual roBERTa based NER model finetuned on 20 annotated Wikipedia datasets[11]. The pre-annotations are the results of the union of the predictions of the 2 models. In particular, in case of multi-word entities, if the predictions of the 2 models overlap, the union of words of the 2 models is used as pre-annotation (for instance, model 1 predicts *"the Mediterranean"* and model 2 *"Mediterranean sea"*, the resulting union is *"the Mediterranean sea"*).

4.1.3 Annotation schema. Defining clearly the meaning of location is crucial for the annotation task. As shown in many other works, a location-related term could take different meaning depending on the context [17, 27]. We follow the main distinction by Gritta et al. [12] between *literal* and *associative* toponyms. A *literal* toponym refers directly to a physical location and an *associative* toponym is only associated with a place. For example, "Syria" is a literal toponym in "latest events in central Syria", but an associative in "Syria Red Cross aided border regions". The annotation task, then, comprises of reading the text, and annotating candidate geotags, as well as creating new ones, and labeling them as either of these 2 categories: literal or associative. Instructions also include rules on handling lists of locations, which are split in post-processing, as well as how to annotate humanitarian-specific text such as names of reports, organizations, and other adjectives that may include toponyms. The instructions are available at shorturl.at/cxJN1.

4.1.4 Annotated geotagging dataset. A total of four 469 documents were annotated by annotators associated with the DEEP platform, specializing in humanitarian data analysis. The median number of toponyms per document is 25, and the total number of toponyms annotated is 11 025. Recall that each document has been truncated at 4000 characters to allow a broader inclusion of different documents. The most common toponyms in the data were *Libya, Syria, Tripoli, Afghanistan, Yemen, Niger, Sudan, Venezuela, Nigeria, Somalia*. The annotated dataset includes *lead_id* of the document, *source* url link, *text* of the document (first 4000 characters) and *annotations* and is available at https://github.com/embelliardo/HumSet_geolocation_annotations.

4.2 Annotation: geocoding

The second annotated dataset concerns the second task of this study: the association of toponyms to a disambiguated location, including the GPS coordinates. For this purpose, we have created a dataset of toponyms in context mapped to an ID of GeoNames entry. GeoNames is a geographical database that contains over 25 million geographical names, of which 4.8 million are populated places, which also come with 13 million alternate names in local alphabets[12].

---

[8]https://labelstud.io/

[9]https://spacy.io/models/en#en_core_web_md

[10]https://huggingface.co/Davlan/xlm-roberta-base-wikiann-ner

[11]https://huggingface.co/datasets/wikiann

[12]https://www.geonames.org/about.html

The toponyms were selected from a random selection of documents annotated in the previous section. To prepare the toponyms for annotation, regular expressions were used to split the lists of toponyms into individual locations (i.e. "Al Hudaydah and Taizz governorates" to "Al Hudaydah governorate" and "Taizz governorate"), associate noun modifiers to each location (i.e. "North and South Italy" to "North Italy" and "South Italy"), rephrase possessive pronouns (i.e. "City of New York'" to "New York City"), standardize special characters, and remove dashes and apostrophes. These cleaned toponyms were then matched with the database of GeoNames fields of *name* and *alternate name* (which provides additional ways of writing the location name), which were similarly preprocessed by standardizing special characters, and removing dashes and apostrophes. The match was done in two ways. First match was performed using longest consecutive word matching. The matches were then filtered through a custom list which includes cardinals alone, common words, and generic location signifiers. Second match was performed on the same GeoNames fields using the Whoosh search engine that uses the Okapi BM25F ranking function to retrieve strings[13]. This way of matching allowed for more flexible positioning of keywords (i.e. "Tripoli Airport" can be matched to "Tripoli International Airport", which consecutive word matching would miss). The search engine matches were then filtered based on the "feature code" of the GeoNames entity, which describes what kind of location it is. We keep locations having this code among the following (hand-crafted based on relevance): administrative division (AD), populated place (PPL), mountain (MT), sea (SEA), lake (LK), island (ISL) and airport (AIR). Note that we use the same pipeline for the preparation of the toponyms for geocoding using our automated algorithm below.

The annotation was performed using a custom-built tool that displays the toponym, the document as its context, and a list of candidate GeoNames entities, which are located on a map. The annotators were instructed to select a GeoNames entity most likely to be referenced by the toponym, and if the correct one cannot be found, to select "none". The main author of this paper performed the annotation, resulting in 561 unique document/toponym match pairs from 39 documents, with 474 having non-empty matches, spanning 78 countries. The annotated dataset includes *lead_id* of the document, *toponym*, *match* and the *GeoNames ID* of the correct match, and is available at https://github.com/embelliardo/HumSet_geolocation_annotations.

5 CUSTOMIZING GEO-LOCATION FOR HUMANITARIAN TEXTS

In the next sections, we describe the evaluation of state-of-the-art geotagging methods and tuning of these methods with the annotated data.

5.1 Geotagging

Using the annotated geotagging dataset, we were now able to evaluate the off-the-shelf NER models – Spacy and roBERTa – used in pre-selection of the toponyms. We compute two versions of the metrics: exact agreement considers only exact string matches to the ground-truth labels (subscripts e in Table 2), whereas partial

---

agreement considers any substring overlap (subscript p). Note that partial scoring includes exact matches. We define agreement between the output of the model and the ground-truth as the number of toponyms predicted by the model which are in the annotated set, divided by the union of toponyms of the two sets. We further use the standard definitions of precision and recall. In the case of multiple correct matches, we count such matches as one (by weighting each match proportional to the number of sub-matches), such that we do not inflate the number of matched toponyms. Table 2 reports the performance of the baseline models (as well as the tuned ones, described below). Both algorithms were evaluated on the entire annotated set using 10-fold cross-validation. The cross-validation is run in two ways: without stratification, and with stratification by the country of the project, such that the training data does not have information about the test project country, making it a more difficult task.

For both algorithms, and both evaluation methods, we find the tuning of the model with additional data to improve performance, both in terms of precision and recall. For roBERTa, the improvement in F1 is from 0.72 to 0.79 (not stratified) for exact matches and from 0.85 to 0.93 (not stratified) for partial ones. The improvement is even more for Spacy, a 15 percentage point improvement in F1 to 0.79 (not stratified) for exact matches, and a 13 point improvement in F1 to 0.91 (not stratified) for partial. Interestingly, the variance of performance of the Spacy model decreases with additional tuning, however in the stratified testing it remains higher. As expected, the performance is lower when the data is stratified by project country, but there are still some improvements. We also combine the two models by taking a union of the matches produced by both algorithms (and a largest span of those that partially overlap). This approach favors the recall, achieving 0.97 in partial matching, while still having a precision of 0.89. The training brings the performance of the two algorithms closer (with Spacy having largest improvement), but other considerations may be important in the deployment of the models. Whereas Spacy model takes about 6.5MB, roBERTa is much larger at 1.1GB. Similarly, Spacy runs faster, about 3.2 times faster than roBERTa. To conclude, if one favors recall, a combined model should be used; if a low variance in performance is desired, roBERTa may be a better choice; but if a lightweight model is needed for fast processing, Spacy may be a better selection.

Figure 1 shows the improvement in performance as the amount of training data is increased. We find a rise in performance within the first 10% (about 50 documents), which gradually increases especially for the partial matches, which is true for both models, and for both testing using perfect match and partial match. roBERTa benefits especially from the new data, achieving F1 score above 0.9 with additional data (for partial scoring).

Next, we examine the disagreements in the extracted toponyms by the different algorithms. In Table 3 we show several examples of a text fragment and the associated toponyms extracted by the two algorithms: in their baseline form and finetuned. The first example talks about a boat ("VOS vessel") named Theia, which is extracted as a location by the roBERTa baseline model, but is ignored by the finetuned one (Spacy performs correctly even as a baseline). The second example has the phrase "Syrian Government", which is an associative toponym that we do not select for the present task. The tuned roBERTa model learn to ignore such cases. Third

Table 2: Agreement, precision, recall, and F1 for geotagging methods: mean (standard deviation) over 10 folds. Exact matches with subscript e and partial with subscript p. Tested with and without stratification ("ST") by country of the project.

Method	A_e	P_e	R_e	$F1_e$	A_p	P_p	R_p	$F1_p$
roBERTa$_{baseline}$	.63(.03)	.73(.02)	.72(.03)	.72(.02)	.74(.03)	.85(.02)	.85(.03)	.85(.02)
roBERTa$_{tuned}$	**.75(.03)**	.78(.02)	**.81(.02)**	**.79(.02)**	**.87(.03)**	.92(.02)	.95(.01)	**.93(.01)**
roBERTa$_{tuned\_ST}$	.72(.03)	.79(.04)	.76(.03)	.78(.03)	.85(.04)	**.94(.02)**	.90(.04)	.92(.02)
Spacy$_{baseline}$	.53(.08)	.74(.07)	.57(.08)	.64(.07)	.64(.07)	.90(.04)	.68(.08)	.78(.05)
Spacy$_{tuned}$	.74(.03)	**.79(.03)**	.80(.03)	**.79(.02)**	.85(.02)	.92(.03)	.92(.03)	.91(.01)
Spacy$_{tuned\_ST}$	.65(.06)	.77(.06)	.70(.07)	.74(.05)	.77(.06)	.92(.04)	.82(.07)	.87(.04)
Combined	.72(.03)	.74(.04)	.80(.02)	.78(.03)	**.87(.03)**	.89(.03)	**.97(.00)**	.93(.02)
Combined$_{\_ST}$	.70(.05)	.73(.05)	.77(.03)	.75(.04)	.86(.04)	.90(.04)	.95(.01)	.92(.03)
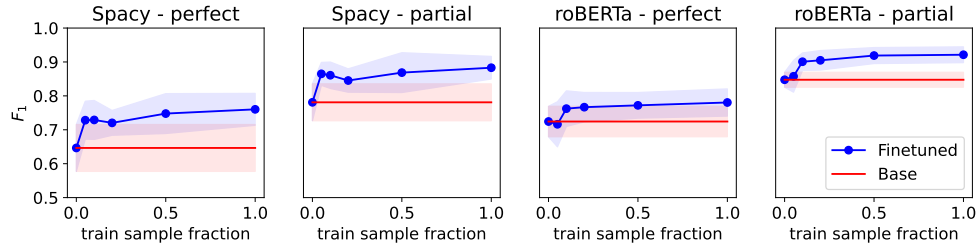
Figure 1: Average F_1 score and standard deviation on 10 folds cross-validation at different sample size. F_1 score is computed for Spacy and roBERTa models, baseline and finetuned, and for perfect and partial matches.

example is similar, but we find that the training benefits Spacy, such that the associative toponym is not selected. In the fourth example the location string is mixed with numeric characters. Whereas the baseline roBERTa model is not able to identify it, Spacy identifies the word, but with additional numerical characters. Training improves roBERTa, but makes Spacy ignore the string (note that a combined model would still find the toponym). The fifth example shows how the trained model (in this case roBERTa) is capable of recognising long lists of noun modifiers associated with "districts". Finally, the last example shows how the trained models are able to identify location at a very granular level, in this case an hospital building. Although not specifically focusing on particular kinds of toponyms, we find that smaller locations such as hospitals, airports, refugee camps, and others may especially necessitate training data to be caught by the named entity extractors.

5.2 Geocoding

5.2.1 Baseline Approaches. Considering the geocoding task – the mapping of toponyms to a unique set of geo-coordinates, and potentially to other meta-data associated with a location – the literature is sparse in this area. We consider two approaches to geocoding from the previous literature. Buscaldi & Magnini [3] propose an approach based on an iterative resolution of references by favoring "unambiguous" ones, with an assumption that the references appearing at greater frequency (such as countries) are less ambiguous. Intuitively, for each candidate geolocation for a toponym, the algorithm considers resolved references in the context around the

candidate, and computes inverse distance to the unambiguous geolocations. The candidate geolocation having the shortest distance to the unambiguous known references in the text is chosen as the best. Following the findings of the paper, we do not implement the text distance between the toponym and references in the text. We considered as unambiguous toponyms which had only one GeoNames match and had to be in one of the countries mentioned in the document, as well as the countries themselves. The second approach by Chen, Vasardani, and Winter [4] also uses the contextual information around a toponym. It involves the clustering of the candidate geolocations for all toponyms found in a document, finding the best clustering radius, and choosing the candidate geolocations from the biggest cluster. The model assumes that the toponyms found in a document are located near each other. It further does not provide a tie-breaking mechanism for choosing the best geolocation within the selected cluster (so we break ties randomly).

5.2.2 FeatureRank. As an alternative to the general-purpose geocoding algorithms, in this work we propose a custom feature-based geocoding method tailored for the humanitarian domain. This method, FeatureRank, considers geopolitical and population features of candidate locations extracted from GeoNames, as well as a document-wide distribution of the candidate locations for all toponyms, thus capturing both local and global knowledge.

First, the toponyms are preprocessed as in the Section 4.2. Recall that the preprocessing involves several steps in cleaning and splitting toponyms, searching GeoNames database using string matching and search engine, and filtering the candidate geolocation set using a set of codes. This results in a set of candidate

Table 3: Sample of toponyms with their context predicted by baseline and finetuned models. Correct annotation in bold.

roBERTa$_{tuned}$	roBERTa$_{baseline}$	Spacy$_{tuned}$	Spacy$_{baseline}$
1. two VOS vessels. Two will be located on board the VOS Theia, two will be located at Aden port, and one will ...			
None ✓	VOS Theia	None ✓	None ✓
2. ... and volatile in June, with tensions between the Syrian Government and 'reconciled' non-state armed groups reported ...			
None✓	Syrian Government	None ✓	None ✓
3. The US Government Congratulates Buhari in Spite of Violent and Corrupt Election			
None✓	None✓	None✓	US
*4. Plateau(5), Taraba(3), Gombe(1), Kaduna(1), **Kwara**(1), FCT(1), Benue(2), Rivers(1) Kogi(1) ...*			
Kwara✓	None	None	Kwara(1
*5. However, clashes intensified in **At Tuhayat and Zabid districts** of Hudaydah city ...*			
At Tuhayat and Zabid districts ✓	At Tuhayat	Tuhayat	None
*6. sources report 17 dead and eight wounded, currently in treatment at **Am-Timan hospital***			
Am-Timan hospital✓	None	Am-Timan hospital✓	None

geolocations for each toponym in a document. A variant of our algorithm SearchFeatureRank also uses a search engine to query for candidate locations (instead of substring match). Similar to the baseline models, we summarize the distribution of the geolocation candidates in the documents by computing a document-wide country distribution. To compute this distribution, for each toponym in the document we first create a country distribution of its geolocation candidates, and finally average them for a document-wide distribution. Note that we are assuming the document is largely concerning one country, which we find to be the case for most of the documents in our dataset, but is not necessarily true for other settings. We then geocode each toponym.

To do this, we compute a set of features for each geolocation candidate for a toponym:

- `IsCapital`: it is 1 if the GeoNames feature code is PPLC, and only if the candidate comes from the string match.
- `IsCountry`: it is 1 if the GeoNames feature code is PCLI or PCLS, and only if the candidate comes from the string match.
- `AdminLevel`: the numeric level (1-5) of the administrative unit (ADM or PPL) in the GeoNames feature code; 0 for capital cities or countries; 6 for geolocations without a level.
- `IsCity`: it is 1 if the GeoNames feature code begins with PPL (some populated location).
- `Population`: population provided by GeoNames.
- `DocCountryMatch`: the probability of the country of the candidate geolocation in the document country distribution.

Finally, the candidates are ranked using the following order: `IsCapital` descending, `IsCountry` descending, `DocCoun- tryMatch` descending, `AdminLevel` ascending, `IsCity` descending, `Population` descending. The algorithm selects the best results according to the ranking strategy, therefore a candidate is always chosen even if none of the candidates matches the correct location. A threshold rule has proven effective to filter out unlikely solutions: if

`AdminLevel > 5` and the candidate Country is not the country with highest `DocCountryMatch`, the location is discarded.

5.2.3 Evaluation. We evaluate these algorithms on the Geocoding dataset described above. In particular, we consider several metrics:

- For correct prediction, we consider two cases:
  - True Positive: match exists and correctly predicted
  - True Negative: match does not exist and is correctly ignored
  - Correct: sum of the two above
- For incorrect prediction, we consider three cases:
  - False Positive (new selection): match does not exist, but some value is wrongly predicted
  - False Negative: match exists but is not selected (i.e excluded by threshold)
  - False Positive (wrong selection): match exists but wrong candidate is selected
  - Incorrect: sum of the three above
- For False Positive (wrong selection) we compute the average distance of the incorrect guess to the correct one.
- For the toponyms which were tagged as non-locations, how many were identified as locations.

We can see from Table 4 that an approach based on features is better suited than other methods based on geographical distances. Both FeatureRank and SearchFeatureRank reach similar performances, with correct prediction in 79% and 80% of the cases, respectively. The only notable difference is in the median distance error, which decreases in the search engine version. Note that during the geocoding annotation process, it was possible to indicate whether a toponym is correct, and in 18 cases we found that the toponym was selected incorrectly during the first round of annotations. Using this information, we are able to test the performance of our models on such cases, and we find that FeatureRank correctly does not

match these cases to anything 50% of the time (compared to 0% by BM and 22% by Chen baselines).

When we consider the 29 locations known to be in US & Europe separately from all others, we find that our algorithm attains 100% accuracy for locations in US & Europe and 84% for those outside (and for those toponyms which are locations, but which are not in GeoNames, the algorithm performs the worst, at 48% accuracy). However, in terms of guessed locations, only 69% of the locations our algorithm guesses to be in US & Europe are correct, compared to 80% of those guessed in other locales. In particular, our algorithm works well for locations in *Libya*, *Colombia*, *Afghanistan* and *Ecuador*, where 100% of locations are correctly predicted. We note that the worst performing countries are *Panama*, *Mexico* and *Paraguay*, for which the algorithm never predicts correctly (however, we also do not have any projects about *Panama* and *Paraguay* in the training data, though locations from those countries could be mentioned in other projects). We hypothesize it is due to the similar naming and language conventions among these countries.

6 APPLICATION STUDY

Finally, we apply the tuned toponym extraction and the custom FeatureRank geocoding algorithm to the HumSet dataset. Although we do not have the ground truth for these labels, we examine potential biases in the locations that the baseline models find, compared to our tuned ones. In total, we annotate 6733 documents, extracting 13 967 distinct locations. Figure 2 shows the number of locations identified by the two baseline models, and two tuned models, by country (left) and by country grouped by their human development index (right). We find that, generally, the number of matched locations increases for both models. However, for roBERTa, the matches in US decrease dramatically, which is also reflected in the lower matches in countries with very high HDI. Note that this effect was not intended by the tuning of the algorithm, and is an interesting side effect. We can also notice that the number of locations detected increases even in countries that were not in the training set, like Democratic Republic Of the Congo (CD) and Mozambique (MZ).

Figure 3 shows a map with areas highlighted in blue where the baseline roBERTa model finds more locations, and highlighted in red where the tuned roBERTa model finds more. We observe that the baseline model favors United States and Europe, and Australia – countries with high and very high development index. In fact, the tuned model predicts 55% fewer matches in US. On the other hand, the countries highlighted in yellow – those for which there are projects in the HumSet data – show an increase in matches, especially Trinidad and Tobago has more than twice as many matches (from 92 by baseline to 191 by tuned), Domenican Republic increases by 17%, Bahamas by 16%, Sudan 14%, and Lybia by 12%. However, even countries which do not have projects in the HumSet documents have more matches, including Morocco at 102% more, South Korea 90%, North Korea 81%, and Myanmar 64%. As we do not know the ground truth of these matches, more work needs to be done to understand whether this is due to increase in accuracy, or whether this is an evidence of a new geographic bias. If these new matches are indeed correct, this would point to a generalizability of these models beyond the countries of the tuned data.

7 DISCUSSION

As we attempted to improve the performance of state-of-the-art NER for geolocation of entities in humanitarian texts, we discovered that training data from the humanitarian domain has not only improved accuracy of the tools, but also possibly helped alleviate the geographic bias of these tools toward the Western countries. Our findings during the application of the baseline and tuned algorithm to the HumSet data suggest that data extraction tools should be tested for bias in a systematic way, as proposed by algorithmic auditing proponents [32]. The social justice concerns auditors of economic and governmental systems extend to the use of AI tools in regards to possibly vulnerable populations, such as those needing humanitarian assistance. Similar concerns have been raised in the use of AI for medical diagnosis [29] and hiring [26]. Auditing is especially necessary for black-box algorithms wherein coverage and completeness are impossible to ascertain, unlike for algorithms that use gazetteers or databases.

In humanitarian context, geographic accuracy is imperative for the accurate assessment of needs and their development over time. During the development of the annotated resources, we encountered several peculiarities of geolocations present in humanitarian texts. First, we often encounter geo-relevant organization names, such as "Syrian Arab Red Crescent" or "European Commissioner for Crisis Management" that indicate their affiliation or source, but not necessarily the location of their current involvement. We observed that this often would decrease the performance of distance-based algorithms that average locations of found entity candidates. A specialized module to handle such information may benefit the mapping of the entities involved in a project. Second, locations often have qualifiers such as "western Yemen" or "north of Baghdad", necessitating for additional semantics in location representation that could capture both the directionality and uncertainty of the location. Third, rapid developments on the ground, such as the building or dismantling refugee camps, should be added to the geographic resources used for humanitarian aid on an ongoing basis.

More research is needed to improve these tools and evaluate them on a larger set of documents. Further, thus far only documents in English language were considered (although many humanitarian organizations publish their reports in English for international consumption). This paper does not use any personal data, instead it uses only published reports and news articles, available under the Apache 2.0 license. We hope the annotated resources and guidelines provided in this paper will spur further work on tuning NER tools to handle data around vulnerable populations.

REFERENCES

[1] American Red Cross. 2022. What is the Red Cross Global Network? https://www.redcross.org/about-us/news-and-events/news/2022/what-is-the-red-cross-global-network.html. September 16, 2022..
[2] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems* 26 (2013).
[3] Davide Buscaldi and Bernardo Magnini. 2010. Grounding toponyms in an italian local news corpus. In *Proceedings of the 6th workshop on geographic information retrieval*. 1–5.
[4] Hao Chen, Maria Vasardani, and Stephan Winter. 2018. Disambiguating fine-grained place names from descriptions by clustering. *arXiv preprint arXiv:1808.05946* (2018).

Table 4: proportion of correct and incorrect matches, and median distance from prediction to ground truth for wrong matches for geocoding methods. We propose FeatureRank and SearchFeatureRank.

	BM	Chen	FeatureRank	SearchFeatureRank
Locations (count)	543	543	543	543
True Positive	0.58	0.21	0.71	0.72
True Negative	0.00	0.01	0.08	0.08
Correct	0.58	0.22	0.79	0.80
False Positive (new selection)	0.13	0.12	0.05	0.05
False Negative	0.00	0.03	0.04	0.03
False Positive (wrong selection)	0.29	0.63	0.12	0.12
Incorrect	0.42	0.78	0.21	0.20
Median dist. (Km)	219.36	2610.94	83.11	61.56
Non-locations (count)	18	18	18	18
Correct	0.00	0.22	0.50	0.44
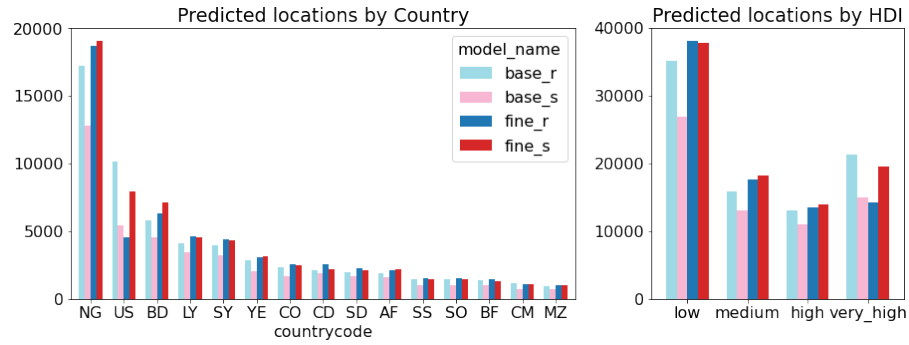
Figure 2: Number of locations detected in HumSet by baseline Spacy (s) and roBERTa (r) models, as well as the tuned ones by country (left, top 15 countries shown), and by the country's human development index (right).
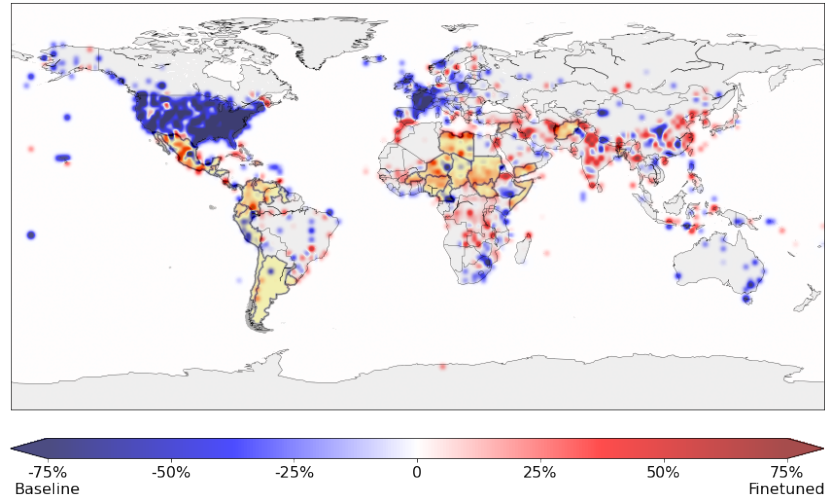
Figure 3: Map of the difference in locations identified by roBERTa: in red more by the tuned model, and in blue more by the baseline model. Countries in yellow are those for which HumSet contains projects.

[5] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of machine learning research* 12, ARTICLE (2011), 2493–2537.

[6] Craig M Dalton, Linnet Taylor, and Jim Thatcher. 2016. Critical data studies: A dialog on data and space. *Big Data & Society* 3, 1 (2016), 2053951716648346.

[7] Franck Dernoncourt, Ji Young Lee, and Peter Szolovits. 2017. NeuroNER: an easy-to-use program for named-entity recognition based on neural networks. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 97–102.

[8] Selim Fekih, Nicolo' Tamagnone, Benjamin Minixhofer, Ranjan Shrestha, Ximena Contla, Ewan Oglethorpe, and Navid Rekabsaz. 2022. HumSet: Dataset of Multilingual Information Extraction and Classification for Humanitarian Crises Response. In *Findings of the Association for Computational Linguistics: EMNLP 2022*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 4379–4389. https://aclanthology.org/2022.findings-emnlp.321

[9] Mark Graham. 2015. The hidden biases of Geodata. https://www.theguardian.com/news/datablog/2015/apr/28/the-hidden-biases-of-geodata. The Guardian. [Accessed April 10, 2023].

[10] Mark Graham and Stefano De Sabbata. 2015. Mapping information wealth and poverty: the geography of gazetteers. *Environment and Planning A* 47, 6 (2015), 1254–1264.

[11] Mark Graham, Scott A Hale, and Devin Gaffney. 2014. Where in the world are you? Geolocation and language identification in Twitter. *The Professional Geographer* 66, 4 (2014), 568–578.

[12] Milan Gritta, Mohammad Taher Pilevar, and Nigel Collier. 2019. A pragmatic guide to geoparsing evaluation: Toponyms, Named Entity Recognition and pragmatics. *Language Resources and Evaluation* 54 (09 2019). https://doi.org/10.1007/s10579-019-09475-3

[13] Muhammad Imran, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. 2015. Processing social media messages in mass emergency: A survey. *ACM Computing Surveys (CSUR)* 47, 4 (2015), 1–38.

[14] Muhammad Imran, Umair Qazi, and Ferda Ofli. 2022. Tbcov: two billion multilingual covid-19 tweets with sentiment, entity, geo, and gender labels. *Data* 7, 1 (2022), 8.

[15] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, Vol. 1. 2.

[16] Kelvin Lai, Jeremy R Porter, Mike Amodeo, David Miller, Michael Marston, and Saman Armal. 2022. A natural language processing approach to understanding context in the extraction and geocoding of historical floods, storms, and adaptation measures. *Information Processing & Management* 59, 1 (2022), 102735.

[17] Jochen L. Leidner. 2007. Toponym resolution in text: Annotation, evaluation and applications of spatial grounding. *SIGIR Forum* 41 (01 2007), 124–126.

[18] Chen Liang, Yue Yu, Haoming Jiang, Siawpeng Er, Ruijia Wang, Tuo Zhao, and Chao Zhang. 2020. Bond: Bert-assisted open-domain named entity recognition with distant supervision. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*. 1054–1064.

[19] Lajanugen Logeswaran, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, Jacob Devlin, and Honglak Lee. 2019. Zero-Shot Entity Linking by Reading Entity Descriptions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 3449–3460.

[20] Momin Malik, Hemank Lamba, Constantine Nakos, and Jürgen Pfeffer. 2015. Population bias in geotagged tweets. In *Proceedings of the international AAAI conference on web and social media*, Vol. 9. 18–27.

[21] Stuart E Middleton, Giorgos Kordopatis-Zilos, Symeon Papadopoulos, and Yiannis Kompatsiaris. 2018. Location extraction from social media: Geoparsing, location disambiguation, and geotagging. *ACM Transactions on Information Systems (TOIS)* 36, 4 (2018), 1–27.

[22] David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes* 30, 1 (2007), 3–26.

[23] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. 701–710.

[24] Fabio Poletto, Yunbai Zhang, André Panisson, Yelena Mejova, Daniela Paolotti, and Sylvain Ponserre. 2021. Developing Annotated Resources for Internal Displacement Monitoring. In *Companion Proceedings of the Web Conference 2021*. 136–144.

[25] Mansi A Radke, Nitin Gautam, Akhil Tambi, Umesh A Deshpande, and Zareen Syed. 2018. Geotagging Text Data on the Web—A Geometrical Approach. *IEEE Access* 6 (2018), 30086–30099.

[26] Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. 2020. Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 469–481.

[27] Satoshi Sekine, Kiyoshi Sudo, and Chikashi Nobata. 2002. Extended Named Entity Hierarchy. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*. European Language Resources Association (ELRA), Las Palmas, Canary Islands - Spain. http://www.lrec-conf.org/proceedings/lrec2002/pdf/120.pdf

[28] Özge Sevgili, Artem Shelmanov, Mikhail Arkhipov, Alexander Panchenko, and Chris Biemann. 2022. Neural entity linking: A survey of models based on deep learning. *Semantic Web* Preprint (2022), 1–44.

[29] Laleh Seyyed-Kalantari, Haoran Zhang, Matthew BA McDermott, Irene Y Chen, and Marzyeh Ghassemi. 2021. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature medicine* 27, 12 (2021), 2176–2182.

[30] Jyoti Prakash Singh, Yogesh K Dwivedi, Nripendra P Rana, Abhinav Kumar, and Kawaljeet Kaur Kapoor. 2019. Event classification and location prediction from tweets during disasters. *Annals of Operations Research* 283 (2019), 737–757.

[31] Statista. 2022. Leading countries based on number of Twitter users as of January 2022. https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/. [Accessed April 10, 2023].

[32] Briana Vecchione, Karen Levy, and Solon Barocas. 2021. Algorithmic auditing and social justice: Lessons from the history of audit studies. In *Equity and Access in Algorithms, Mechanisms, and Optimization*. 1–9.

[33] Wikipedia. 2022. List of Wikipedias. https://meta.wikimedia.org/wiki/List_of_Wikipedias. [Accessed April 10, 2023].

[34] Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2020. Scalable Zero-shot Entity Linking with Dense Entity Retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 6397–6407.

[35] Vikas Yadav and Steven Bethard. 2018. A Survey on Recent Advances in Named Entity Recognition from Deep Learning models. In *Proceedings of the 27th International Conference on Computational Linguistics*. 2145–2158.

[36] Xin Zheng, Jialong Han, and Aixin Sun. 2018. A survey of location prediction on twitter. *IEEE Transactions on Knowledge and Data Engineering* 30, 9 (2018), 1652–1671.

[37] Lei Zou, Nina SN Lam, Heng Cai, and Yi Qiang. 2018. Mining Twitter data for improved understanding of disaster resilience. *Annals of the American Association of Geographers* 108, 5 (2018), 1422–1441.