Abstract—Public conversations on Twitter comprise many pertinent topics including disasters, protests, politics, propaganda, sports, climate change, epidemics/pandemic outbreaks, etc., that can have both regional and global aspects. Spatial discourse analysis rely on geographical data. However, today less than 1% of tweets are geotagged; in both cases—point location or bounding place information. A major issue with tweets is that Twitter users can be at location A and exchange conversations specific to location B, which we call the Location A/B problem. The problem is considered solved if location entities can be classified as either origin locations (Location As) or non-origin locations (Location Bs). In this work, we propose a simple yet effective framework—the True Origin Model—to address the problem that uses machine-level natural language understanding to identify tweets that conceivably contain their origin location information. The model achieves promising accuracy at country (80%), state (67%), city (58%), county (56%) and district (64%) levels with support from a Location Extraction Model as basic as the CoNLL-2003-based RoBERTa. We employ a tweet contextualizer (locBERT) which is one of the core components of the proposed model, to investigate multiple tweets’ distributions for understanding Twitter users’ tweeting behavior in terms of mentioning origin and non-origin locations. We also highlight a major concern with the currently regarded gold standard test set (ground truth) methodology, introduce a new data set, and identify further research avenues for advancing the area.

Index Terms—location extraction, geotagging tweets, social media analytics, Twitter analytics, location transformer

I. INTRODUCTION

Microblogging platforms such as Twitter and Sina Weibo have a significant impact on public discourse; these platforms have raised the possibility of people participating in public conversations, and have become an active source of information during both day-to-day life and in unusual circumstances [14] such as earthquakes, floods, hurricanes, tsunamis, cyclones, wildfires, and pandemics. Especially during emergency events, the number of conversations generated on these platforms reaches hundreds of thousands and even millions in long-term events such as the ongoing COVID-19 pandemic. For instance, during an ongoing disastrous event, people tend to use these platforms excessively, as they share their safety status and exchange conversations to query the safety status of their friends and family. People also share what they have seen, felt, or heard from others. During such critical hours of a disaster, the use of social media peaks at unprecedented levels, and based on these public conversations, first responders and decision-makers can visualize a more comprehensive real-time picture of the situation to aid in formulating actionable plans. Collection, processing, and analysis of such an enormous amount of socially generated data is not practical to manually undertake and warrants the use of automated stream processing systems for a timely understanding of how an event is unfolding—situation awareness—concerning its temporal and spatial dimension [14].

Spatial analysis rely on geographic location data. Location-involved analysis concerning microblogging conversations include discourse-based forecasting, movement and travel pattern analysis, behavior-based interests and intent analysis, generation of task and region-specific heatmaps, environment monitoring, transportation planning, urban planning, mobility analysis, crisis management and so on [14]. However, today less than 1% of tweets are geotagged [12], [29], be it with point coordinates or bounding place information. For instance, the billion-scale COVID-19 tweets dataset at IEEE [11], curated since the inception of the pandemic, reports 480k tweets being geotagged with point coordinates [10] out of 2 billion tweets collected (the daily distribution of tweets for both the datasets is shown in Figure 1). Therefore, it is critical to extract the origin locations of as many tweets to decrease the possibility of bias during location-involved discourse analysis such as [15].

A. The geolocation extraction problem

Numerous tweet objects in the tweet data dictionary\(^1\) assist in the extraction of location information. If a tweet is geotagged, it either contains the coordinates object to represent the tweet’s point location or the place object to represent a place. The place information available in the place object might not be the origin location of the tweet. For non-geotagged tweets, tweet objects such as text, user:location, user:description, entities:media, and lang (for regional languages) are helpful for location inference. Country or region-specific contexts in profile_image_url https and profile_banner_url objects also can assist in location inference. The utc_offset and time_zone objects have been used in the past for inferring probable

----

\(^{1}\)https://developer.twitter.com/en/docs/twitter-api/v1/data-dictionary/

---

Rabindra Lamsal  
*The University of Melbourne*  
Melbourne, Australia  
rlamsal@student.unimelb.edu.au

Aaron Harwood\(^8\)  
*The University of Melbourne*  
Melbourne, Australia  
arhrwd@gmail.com

Maria Rodriguez Read  
*The University of Melbourne*  
Melbourne, Australia  
maria.read@unimelb.edu.au

\(^8\)Work done while the author was at the University of Melbourne.
locations; however, these objects have become deprecated in recent versions of Twitter’s APIs.

Fig. 1: Daily distribution of tweets in the COV19Tweets [11] and GeoCOV19Tweets [10] Datasets. Today less than 1% of tweets are geotagged.

**Location A/B problem.** The major issue with determining the origin locations of tweets is that a Twitter user can be at location A and participate in a public discourse specific to location B—which we call the Location A/B problem [13]. Tweets comprise two types of geolocations, i.e., origin and non-origin locations. Origin locations are the ones from where the tweet has been made; a tweet such as “I love the weather here in Melbourne” suggests that the user was at the mentioned geolocation at the time the tweet was made. However, tweets such as “Stop this war. #Russia” and “Melbourne is a beautiful city.” do not provide sufficient evidence to conclude that the mentioned geolocations are the origin locations. Also, tweets with point coordinates might not necessarily originate from the respective point sources, which eventually contributes to the Location A/B problem. We discuss this issue later in Section IV-B1 with supporting data.

The textual data contains the primary context for origin and non-origin locations, therefore, this study investigates if machine-level natural language understanding is effective in identifying tweets that possibly contain origin locations. We rely exclusively on the text object, since the other location-inferable tweet objects, discussed earlier, enclose location identities that might be independent to any particular tweet.

The paper is organized as follows: Related work is discussed in Section II. Section III discusses the fundamentals—geocoding, locations vectors and data. Section IV proposes a framework for addressing the Location A/B problem. Section V evaluates the proposed framework. Section VI investigates Twitter users’ behavior in terms of mentioning origin and non-origin locations in tweets, and Section VII concludes the paper.

**II. RELATED WORK**

Categorizing locations as origin or non-origin involves the extraction of location entities as an initial step. There have been numerous studies that aimed at inferring location entities from tweets. The majority of the studies involve using named-entity recognizers [6], [20] for extracting toponyms from the text tweet object. Additional tweet objects, such as user:location, user:description, user:url, user:timezone, have also been incorporated as possible spatial indicators [32].

In [18], Li et al. proposed a location identification method that combines multiple tweets of a user to identify top-k locations of the user and further refines the identified locations to come up with top-k locations for a tweet. In [17], Li and Sun extracted fine-grained locations from tweets based on mentioned point-of-interest (POI) while exploiting the Foursquare check-ins done by users from the same geolocation. Similarly, Mahmud et al. [22] proposed an ensemble approach that involved extracting features such as words, hashtags, and places, leveraging domain knowledge, and combining statistical and heuristics classifications to infer the home locations of Twitter users.

Davis Jr et al. [3] inferred locations of tweets based on analyzing the follow/follower associations amongst Twitter users; the idea involved obtaining reciprocal relationships by taking an intersection between the followers of a user and the people the user follows. In [28], Priedhorsky et al. proposed an inference method based on Gaussian mixture models (GMMs) trained on geotagged tweets. For inferring the origin location of a new tweet, the previously trained GMMs (for each unique $n$-gram, a 2D GMM was fitted to model it geographic distribution) were combined for the $n$-grams present in the tweet. In [36], Yuan et al. proposed a probabilistic model that incorporates the who (user), where (geo), when (time), and what (terms) contexts for inferring the location of a user given a tweet and a time. Zubiaga et al. [37] created eight different classifiers based on user:location, user:lang, user:timezone, lang, user:utc_offset, user:name, user:description, and text tweet objects to infer the country of origin of tweets. Similarly, Lau et al. [16] proposed an end-to-end neural network that uses text, created_at, user:utc_offset, user:timezone, user:location, user:created_at tweet objects to predict the geolocation of a tweet.

Content similarity has also been helpful in location inference problems. In [9], Kinsella et al. estimated distributions of terms for each location considered in the study and assigned a new tweet to the location with the highest probability by sampling from the distributions. In [27], Paraskevopoulos and Palpanas used similarity measure between a tweet and a set of geotagged tweets alongside their temporal characteristics to infer its origin at city level. Similarly, Li et al. [19] split users’ timelines into different clusters, each cluster implying a distinct location at city level. Next, they classified each cluster into its predefined class, such that future tweets are assigned city-level locations accordingly. In [7], Ikawa et al. proposed an association-based approach to identifying locations of tweets based on the relations between locations and their relevant keywords from past tweets.

Gazetteer-based searches for n-grams have been employed extensively to identify location candidates. In [25], Middleton et al. performed location token matching for geoparsing tweet content. Similarly, in [23], Malmasi and Dras proposed Noun Phrase extraction and n-gram-based matching to detect
mentions of locations in tweets. Similarly, in [29], Qazi et al. performed n-gram-based matching to identify location candidates. The authors extended their work [8] with the use of language-based NER models. The choice of gazetteer remains to be a trade-off between recall and efficiency; Dutt et al. [5] evaluated the performance of the most favored gazetteers, i.e., GeoNames and OpenStreetMaps.

**Contributions of this study**

- As per our extensive literature search, this study is the first to address the Location A/B problem; we introduce a tweet-processing framework—the True Origin Model—for the task. The framework has two components, Location Extraction Model (LEM) and locBERT. The location inference methods/models in the existing literature are the LEM candidates. The LEM candidates do not consider Location A/B context and tweets such as “Stop this war. #russia” also get geotagged and included, which introduces biases in spatial analysis or discourse-based models. To address this issue, the next component in the framework, locBERT, enriches geotagging by extracting presence/absence of origin location information.
- We introduce the concept of tweets origin location evidence dataset for training language models that aim to identify the origin locations of tweets.
- We investigate multiple distributions of tweets to explore Twitter users’ tweeting behavior in terms of mentioning origin and non-origin locations.
- We identify and present significant research avenues for future work advancing the area.

## III. The Fundamentals

### A. Forward/reverse geocoding

**Forward Geocoding**, or simply Geocoding, refers to the process of converting addresses (such as “700 Swanston Street, Carlton, Melbourne, VIC”) to their geographic coordinates (such as longitude 144.96449828 and latitude -37.80011159). The addresses may not necessarily be structured as the aforementioned example; for instance, “melbourne university” converts to longitude 144.96130134 and latitude -37.7970796, and “chapman avenue glenroy” converts to longitude 144.9121865 and latitude -37.80011159. **Reverse geocoding**, on the other hand, is the process of converting longitude/latitude pairs into human-readable addresses.

**Planet-level API endpoints.** There are services that provide forward and reverse geocoding endpoints—Google’s Geocoding API, Mapbox, positionstack, and Nominatim, to name a few. The existing services either limit the number of search hits to their endpoints or become expensive to pay for when experimenting with hundreds of thousands of queries. Therefore, we built planet-level forward (search endpoint) and reverse (reverse endpoint) geocoding endpoints for this study on a Linux machine with 24 VCPUs and 216 gigabytes memory. The endpoints are powered by OpenStreetMap data and elastic search², with a search index of 61 gigabytes³ (when compressed). Both the endpoints return a payload in JSON format with the following keys: coordinates, properties:{country, city, countrycode, postcode, type, street, district, name, state}.

### B. Location vectors

This study aims to extract geographic locations from tweets at five different levels—district, county, city, state, and country. Given a location, \( V_{LOC} \), a location vector, \( V_{LOC} \), is returned by the forward geocoding endpoint, with values for each component of \( V_{LOC} \). For instance, if \( LOC \) is a district, its location vector, \( V_{LOC} \), will have valid entries for each component; however, if \( LOC \) is a state, the granular components—city, county, district—will have NULL entries, and state and country components will have valid entries. However, during reverse geocoding, for a set of longitude/latitude pairs provided to the reverse geocoding endpoint, all the components of \( V_{LOC} \) are returned with valid entries.

### C. Twitter Data

The primary source of data for this study is the GeoCOVID19Tweets Dataset [10], which is the geo-version of the billion-scale COVID19Tweets Dataset [11] that is currently being maintained and updated daily at IEEE DataPort and is one of the longest-running tweet collections regarding the COVID-19 pandemic. The dataset is being curated while tracking Twitter’s near-real-time feed for coronavirus-related tweets using more than 90 keywords and hashtags that are closely related to the pandemic; experimenting on this dataset helps capture the overall conversational dynamics of the pandemic such that the proposed method can be implemented on a similar scale or smaller COVID-19 specific datasets. The dataset [10] contains tweet identifiers of geo-tagged (point-location) COVID-19 specific tweets; this study makes use of the dataset’s latest version, i.e., the tweets collected over the period March 20, 2020 – April 28, 2022. Twitter’s data re-distribution policy restricts the sharing of data except for tweet identifiers, user identifiers, and message identifiers; the identifiers need to be hydrated to re-create a raw tweet corpus locally [12]. The tweet identifiers in the GeoCOVID19Tweets Dataset were hydrated using Twitter’s Tweets lookup⁴ endpoint. Figure 2 shows the geographical coverage of the dataset.

**Issue with gold standard test set methodology.** The currently regarded gold standard test set methodology for evaluating the performance of tweet origin location identification models uses sets of geotagged tweets with point locations that are extracted using the coordinates:coordinates tweet object. However, during this study, we observed that around 71% of tweets with point locations on Twitter are in fact Instagram posts re-shared by their respective authors. This is based on inspection of the source tweet object on one of the longest-running (25+ months) COVID-19-specific tweets repositories [10], [11]. Instagram aims primarily at

---

²https://en.wikipedia.org/wiki/Elasticsearch  
³https://download1.graphhopper.com/public/  
⁴https://developer.twitter.com/en/docs/twitter-api
sharing photos and videos, with less of an emphasis on real-time posts compared to Twitter’s feed which has a focus on “as-they-happen” personal and news updates. When the tweets in [10] were scrutinized, the majority of tweets shared from Instagram were identified as stories/updates from the past (for example—#throwback to a visit from our hiring partner! #repost #thankyou #repost walgreensjobs. Pre-Covid, one of our pharmacy teams visited the Miami campus of Florida Vocational Institute to speak to future pharmacy...). Furthermore, third-party platforms that provide auto-share and scheduling features also give rise to the Location A/B problem. Tweets created through the native application can also exhibit Location A/B issues. These observations suggest that even the geotagged tweets (with point locations) might not necessarily originate from the respective point sources; concluding that the contexts of tweets should be critically considered when using the geotagged tweets as ground truth for any particular study.

IV. THE PROPOSED FRAMEWORK AND EXPERIMENTS

Our proposed framework—the True Origin Model—is valid for all possible cases where a set of tweet identifiers are available, primarily: (i) pre-collected sets of tweet identifiers, (ii) searching or streaming of raw twitter data. The first case requires the use of the Tweet lookup endpoint, while the second case requires requesting search or streaming endpoints with valid queries (combination of keywords, hashtags, filters/conditions).

Figure 3 illustrates the overall view of the proposed True Origin Model. The hydrated tweets or search/stream-returned tweets are provided as inputs to the core of the model that consists of three major components: (i) a Location Extraction Model (LEM), (ii) two planet-level geocoding endpoints, and (iii) locBERT (a tweet contextualizer). The LEM component, discussed in more detail in Section IV-A, extracts location mentions present within the text tweet object. The input tweets are checked for location mentions, and only the tweets with at least one location entity are considered for the next phase.

Each identified location mention is validated by a planet-level forward geocoding endpoint (discussed in Section III-A). If the geocoding endpoint returns a valid response for an identified location mention, we consider the location as valid. Tweets with at least one valid location mention are provided as input to locBERT (discussed in detail in Section IV-B)—a transformer-based [33] model that we propose for identifying tweets that conceivably contain their origin location information. Tweets identified as containing origin information are processed further to extract location vectors based on the identified valid location mentions. Finally, we take the majority of similar vector components for each tweet to extract the probable origin location.

Next, we discuss LEM and locBERT in detail.

A. Location Extraction Model

The Location Extraction Model (LEM), in general, is a location-entity extraction model that identifies location mentions in unstructured texts or documents. For instance, from the tweet “Love this bunch. Off to the Park for a morning walk in nature #NJ #NewJersey. First COVID Times visit to #Manhattan. HRH @MENTION & I, are in #NewYork for a few days.... #NYC here we are…”, a LEM candidate identifies the following tokens as location entities: Park, NJ, NewJersey, Manhattan, NewYork, NYC.

The location inference methods/models that we discussed in the related work section are LEM candidates. For experimentation purposes, and to avoid reinventing the wheel, we forked some of the most popular transformer-based named-entity extraction models from Hugging Face [34]. Note that, the LEM is an unrestricted component that can incorporate any geotagger for extracting location entities. We randomly sampled 60k tweets from the GeoCOV19Tweets Dataset using pandas’ DataFrame.sample procedure (for reproducibility: random_state=19). We experimented with the four most forked transformer-based [33] LARGE architecture pre-trained models on Hugging Face, namely XLM-RoBERTa [2], RoBERTa [21], BERT [4], ELECTRA [1], on the sampled tweets corpus. These models were fine-tuned on the CoNLL-2003 dataset [30] to extract location entities. The performance of each model on the sampled tweets corpus is summarized in Table I. We consider only the LARGE architectures since they seem to outperform the base variants; using LARGE architectures justify our use case since we aim to identify as many location mentions in the text tweet object. Algorithm 1 presents the pseudocode for performance comparison of the LEM candidates.

The results in Table I show that RoBERTa outperforms all models in all four cases: identifying the most number of (i) location mentions, (ii) tweets with at least one location mention, (iii) valid locations, and (iv) tweets with at least one valid location. XLM-RoBERTa benchmarks second and is followed by BERT and ELECTRA in all four cases. However, there is another dimension to these performance scores—the

5https://live.rlamsal.com.np/

6https://pandas.pydata.org/
### Algorithm 1: Pseudocode for performance comparison of LEM candidates

**Function** location_validity(location):
- Send request to the forward geocoding endpoint.
- if len(response[‘features’]) == 0 then validity = False
- else validity = True
- return validity

**Function** locations_extract(tweet):
- result = model(tweet)
- Maintain a list loc_details.
- Maintain a list locations.
- for each identified location in result do
  - Add the identified location to locations list
  - Maintain a count of identified location(s).
  - if location_validity(identified location) == True then
    - Maintain a count of valid identified location(s).
    - Append locations and maintained counts to loc_details
- return loc_details

---

**Example 2:** @MENTION and I traveled to #Miami at the beginning of June for his 30th Bday EMOJI and our 3 year anniversary EMOJI! Many of you were asking for travel tips amid COVID-19, so I finally put a blog together to help you…

**Example 3:** Continuing to support local businesses, I drove to Berchman’s Brewing Company this evening which, sadly, is closing their downtown Yakima taproom at the end of July due to the pandemic EMOJI and I purchased a grunt…

---

** TABLE I: Performance of the transformer-based location extraction models on COVID-19 English language tweets corpus. Notes:**
- a) number of locations identified in the tweet corpus,
- b) number of tweets identified with at least one location,
- c) number of valid locations amongst the identified ones,
- d) number of tweets identified with at least one valid location.

| Model      | #loc^a | #loc^b | #loc^c | #loc^d |
|------------|--------|--------|--------|--------|
| XLM-RoBERTa | 75,609 | 37,027 | 71,980 | 37,410 |
| RoBERTa    | 78,454 | 38,027 | 73,537 | 37,964 |
| BERT       | 72,474 | 37,045 | 71,601 | 37,492 |
| ELECTRA    | 67,887 | 34,227 | 67,078 | 34,122 |

**TABLE II: Invalids identified as valids by the transformer-based location extraction models. Notes:**
- a) location(s) for which the planet-level geocoding endpoint returned invalid response,
- b) tweet(s) without any valid location(s).

| Model      | invalid locations^a | invalid tweets^b |
|------------|----------------------|------------------|
| XLM-RoBERTa | 3,629                | 554              |
| RoBERTa    | 4,517                | 535              |
| BERT       | 873                  | 182              |
| ELECTRA    | 909                  | 105              |

---

B. locBERT

1) **Origin location evidence:** On Twitter, besides the tweets geotagged with point location, there are a set of tweets that can be assigned origin location based on how they have been composed. For instance, consider the following tweets:

**Example 1:** Fun 3 Mile Interval Run along the Riverwalk! Supporting our beautiful city of Chicago, our beloved Tamale Guy who has recently been hospitalized, fighting Covid-19 and yes, the USPS, I will always need stamps.
403.9k tweets, 71.10% of them were shared on Twitter from Instagram by the authors of the respective Instagram posts. Instagram aims primarily at sharing media (photos/videos), while Twitter is used for instantaneous personal and news updates and promotes real-time posts. Instagram-post-turned-tweets might not necessarily originate from the locations referenced in the tweets. Furthermore, services with auto share and scheduling features, such as dlvr.it and Hootsuite, also seem to generate a significant number of tweets on Twitter. We provide a list of the top 15 geotagged tweets’ sources in Table III. Therefore, even in the case of geotagged tweets, the point location provided in the tweet data dictionary might not necessarily be the origin location. Taking these factors into consideration, we design a transformer-based language model to implement machine-level natural language understanding in the identification of tweets that conceivably contain their origin location information.

**TABLE III: Top 15 geotagged tweets’ sources.**

| Source               | %     | Source               | %     |
|----------------------|-------|----------------------|-------|
| Instagram            | 71.1% | Twitter for iPhone   | 0.69% |
| dlvr.it              | 13.52%| Untappd              | 0.65% |
| Hootsuite Inc.       | 2.10% | heapevents.info      | 0.54% |
| Tweetbot for iOS     | 1.59% | TweetCaster (iOS)    | 0.37% |
| Foursquare           | 1.10% | Tweetlogix           | 0.32% |
| TweetCaster (Android)| 1.04% | Covid Genie          | 0.28% |
| Squarespace          | 1.01% | Foursquare Swarm     | 0.27% |
| FUNcation            | 0.77% |                      |       |

2) Training data: From the GeoCOVID19Tweets Dataset, we use two independent annotators to manually annotate 2,800 tweets (tweets origin location evidence dataset) into two classes based on the presence and absence of origin location information: (i) true origin, (ii) low evidence, analogous to the examples presented in Section IV-B1 and Appendix. A tweet was annotated as “true origin” only if it gives a clear context for the presence of origin location; otherwise, it was annotated as “low evidence”, even if the context was ambiguous. The annotated dataset had 887 “true origin” tweets and 1913 “low evidence” tweets—the dataset is imbalanced in the ratio [1.58, 0.73]. The Cohen’s kappa coefficient\(^7\) was computed for measuring inter-annotator agreement among the two independent annotators—the statistic was 0.83, which can be interpreted as an almost-perfect agreement [24].

3) Training locBERT candidates: We preprocess the annotated tweets as follows: (i) replace HTML entities with their character representation; such as “&amp;” to “&”, (ii) replace URLs with HTTPURL token, (iii) replace user mentions with USER token, (iv) replace emoji\(^8\) with EMOJI token, (v) clean unnecessary spaces, indentations, paragraph breaks. Next, the dataset is partitioned into train (80%), validation (10%), and test (10%) sets for reproducibility: random_state=19. As the candidate models for locBERT, we fine-tuned the BASE architectures of BERT [4], DistilBERT [31], ELECTRA [1], RoBERTa [21], XLM-RoBERTa [2], XLNet [35], and BERTweet [26] on the annotated dataset using the Hugging Face library. The following configurations were considered during fine-tuning of the locBERT candidates: Number of training epochs: 30, Learning rate: 1e−6, Training batch size: 8, Maximum sequence length: 70, Optimizer: AdamW\(^9\).

The results from the experiments are summarized in Table IV. The fine-tuned BERTweet outperformed all other models in terms of F\(_1\) for both classes. BERTweet is a language model pre-trained using RoBERTa pre-training procedure on a large-scale English language tweet corpus of 845 million tweets and an additional 23 million COVID-19 specific tweets. The results show that pre-training language models on domain-specific text corpora perform better than the models trained on “standard” text corpora. We use the fine-tuned BERTweet as locBERT of our True Origin Model. Refer to Appendix for a sample of tweets that were incorrectly classified by locBERT during its evaluation. Understanding the context is not straightforward as in sentences such as “I love the weather here in Melbourne”, “Stop this war. #russia” and “Melbourne is a beautiful city.”—some real-world examples are provided in Appendix.

**TABLE IV: Performance of locBERT candidates**

| Model      | F\(_1\) true origin | F\(_1\) low evidence |
|------------|----------------------|----------------------|
| BERT       | 0.65                 | 0.86                 |
| DistilBERT | 0.61                 | 0.82                 |
| ELECTRA    | 0.65                 | 0.82                 |
| RoBERTa    | 0.70                 | 0.87                 |
| XLM-RoBERTa| 0.66                 | 0.85                 |
| XLNet      | 0.70                 | 0.86                 |
| BERTweet   | 0.74                 | 0.87                 |

V. EVALUATION OF THE PROPOSED FRAMEWORK

In this section, we evaluate LEM, planet-level endpoints, and our proposed locBERT together as a pipeline, shown in Figure 3. The standard test data for this study is the GeoCOVID19Tweets Dataset. Since our method incorporates the context of tweets, it is, therefore, valid to use the tweets present in the GeoCOVID19Tweets Dataset for the evaluation.

A. Location vectors

Following the proposed workflow of the True Origin Model in Figure 3, the tweets that were identified by RoBERTa to have at least one valid location (in Section IV-A) were provided as inputs to locBERT for categorizing the tweets as: (i) true origin, or (ii) low evidence (as discussed in Section IV-B2). However, there were a set of invalid locations identified by the RoBERTa model that were marked as valid by the geocoding endpoint. Therefore, as a further pre-processing step, we filtered locations that had: (i) character length less than 1, (ii) only numeric digits, and (iii) solely a generic definition of a region, i.e., “city”, “earth”, “europe”, “asia”, “americas”, “africa”, “world”, “town”, “county”, “district”. After this step, 37.1k out of the 38k tweets were available for evaluation purposes. Those tweets were then input through

---

\(^7\)sklearn.metrics.cohen_kappa_score

\(^8\)https://pypi.org/project/emoji/

\(^9\)https://pytorch.org/docs/stable/generated/torch.optim.AdamW.html
out of the 37.1k tweets, locBERT classified 14.9k tweets as “true origin” and the remaining 22k as “low evidence”. There were instances where a longitude/latitude pair returned a NULL location vector. Also, there were cases where the location(s) identified by LEM returned a NULL location vector. In both of these cases combined, 410 “true origin” class tweets were filtered out with 14,574 tweets, hereafter called the Eval Set, available for evaluation. Table V gives the overview of the pre-processing, classification, and filtration steps discussed above.

| Total tweets | 60,000 |
| Tweets with at least one valid location | 38,027 |
| Tweets after pre-processing | 37,102 |
| locBERT categorization | true origin: 14,984, low evidence: 22,118 |
| Available tweets for evaluation | 14,574 |

B. Conclusive location vectors

For each tweet in Eval Set, we query the reverse geocoding endpoint with the longitude/latitude pair to compute its ground-truth location vector (a list for a single tweet). Similarly, for the location(s) identified in each tweet, we compute the location vector(s) (a list of lists for a single tweet) querying the forward geocoding endpoint. We then take the majority of similar vector components to compute a conclusive location vector. The conclusive location vector is compared with the ground-truth vector for evaluation. We provide an overview of the True Origin Model as pseudocode in Algorithm 2. The results from the evaluation are summarized in Table VI.

C. Comparison with existing studies

It is not justifiable to compare the performance of locBERT with the existing geotagging methods (which are the LEMs) since locBERT’s sole aim is to enrich the geotagging process by identifying the presence/absence of the origin location information. If we process N tweets through existing geotaggers, we get N geotagging results, with tweets such as “Stop this war. #russia” also getting geotagged. However, locBERT ignores such tweets and geotags only (O+N) number of tweets, where O is the proportion of tweets with origin location evidence.

Also, previous studies use text tweet object alongside objects such as created_at, user:location, user:description, user:url, user:timezone, user:verified, user:utc_offset, and user:name. The utc_offset and time_zone objects have been used extensively in the past as location indicators; however, these objects have become deprecated and return NULL values in the tweet payload today. In their absence, it has become critical to explore the possibilities within the text tweet object more than ever. The proposed model could have considered incorporating additional tweet objects for extracting more location vectors to increase its location prediction accuracy. However, the involvement of location indicators other than the text tweet object creates forecast biases when tweets originate from places different than suggested by the indicators. The next-generation true origin models should focus on critically utilizing the text tweet object; similar to user:timezone and user:utc_offset, the currently available location indicators might become deprecated in the future, but raw content will continue to be available. Models trained solely on the text object also have applications outside Twitterverse.

VI. TWEETS ANALYSIS WITH locBERT

We employed locBERT on the GeoCOV19Tweets Dataset (N = 403.9k) to understand the distribution of Twitter users’ behavior towards mentioning origin and non-origin locations, while tweeting, based on the following tweet objects: possibly_sensitive, media, user_verified, user:followers_count, user:friends_count, user:created_at, and country. Figure 4 presents multiple distributions from the experiments. We make numerous deductions based on these distributions. We report that 35.9% of the geotagged tweets on Twitter mention origin locations, and in all the experimented cases, the number of
tweets with origin locations is less than the number of tweets that mention non-origin locations. We observed that only 2.99% of tweets contain sensitive content in shared URLs and the proportion of true origin and low evidence tweets is approximately equal to what has been reported in the overall distribution. We noticed that out of 403.9k tweets 6.65% of them had media items (photos, videos, or GIFs), and amongst the media-attached tweets, 11.2% were true origin tweets, and, in the case of non-media, 37.68% of them were true origin tweets. However, in the case of verified and unverified Twitter accounts, we did not observe a significant difference in the proportion—unverified had 36% and verified had 30% tweets with origin locations.

Next, we created 18 bands to break down the number of followers and followings data. It is evident from Figure 4e that users with huge followers tend to tweet less, while users with 500-1k followers seem to be posting the highest number of tweets and are followed by users in the 0-50 and 200-300 bands. As the number of tweets decreases with the increase in the number of followers, the tweets with origin locations also seem to be following a similar trend. Also in terms of followings (Figure 4f), we observe a similar behavior.

Further, we created a distribution for account creation year versus the frequency of tweets. Figure 4g shows that the users who joined Twitter between 2008–2014 have significant participation in the public discourse. However, the same timeline demonstrates a presence of a notable difference in the proportion of true origin and low evidence tweets. Users who joined Twitter between 2006–2008, and 2014–present seem to be the ones mentioning origin locations comparatively higher in terms of proportion. Furthermore, we created a geolocation-based (country) distribution, and observed a relatively higher presence of low evidence tweets for almost all countries except Australia and the Philippines.

We investigated the influence of the tweets’ sources on the generation of the true origin and low evidence tweets. Sorted by tweet frequency, Instagram tops the list with a presence of 37.61% low evidence tweets (and 33.48% true origin
tweets), followed by dlvr.it, Hootsuite Inc., Tweetbot for iOS, TweetCaster for Android, and Squarespace, all generating low evidence tweets. After Instagram, Foursquare and Untappd are the second and third major sources of true origin tweets. These findings in terms of sources show that even the tweets geotagged with point coordinates cannot be considered ground truth data for evaluating models such as the one proposed in this study if the context for the presence and absence of origin location information is left out.

VII. Conclusion

In this paper, we introduced a tweet-processing framework—the True Origin Model—for addressing the Location A/B problem. The proposed approach involved a location extraction model (LEM), two planet-level geocoding endpoints, and a tweets contextualizer working together in a pipeline. Overall, we: (i) experimented with the four most forked transformer-based NER models as LEM candidates; (ii) created self-hosted instances of two planet-level geocoding endpoints; and (iii) designed a transformer-based model (locBERT) to identify tweets that possibly contain origin locations. locBERT achieved promising results across country (80%), state (67.5%), city (58%), county (56.6%), and district (64%) levels, with LEM as basic as the CoNLL-2003-based RoBERTa. We briefly discussed the issue with the area’s currently regarded gold standard test set methodology. We employed locBERT to investigate multiple tweets’ distributions for understanding Twitter users’ tweeting behavior in terms of mentioning origin and non-origin locations. Finally, we presented some significant research avenues to take on.

Future directions

There are multiple significant research avenues to take on. The best candidate for locBERT achieved an average F1 of 0.805 (F1 on “true origin” class was 0.74) and this suggests there remains room for improvement. Further, the issue with common places needs to be addressed. For instance, the place name “Victoria” is a state in Australia as well as a city in British Columbia, Canada. Such commonality issues can be addressed (to an extent) by scrutinizing the location entities mentioned in chronologically earlier “true origin” tweets in the user’s timeline. Twitter provides three timeline endpoints to fetch tweets from a specific Twitter account. Furthermore, there is an need for an effective processing of ubiquitous location entities such as Hall, Ring Road, Central park, etc. We scrutinized a sample of incorrect identifications and observed that the presence of generic/ubiquitous location entities significantly decreases the model’s performance since the location vector of each identified location entity contributes equally for generating the conclusive location vector. Table VII presents a sample of location entities associated with incorrect identifications.

| TABLE VII: Some example generic/ubiquitous location entities associated with the incorrect identifications |
|---------------------------------------------------------------|
| Saint, PeelsSt, Manship, aroc, ari, The Bush, East Lynn, Sola, AVATAR, Kings Inn, The Dungeon, Willows, West End, ola, Brandon, Central Park, Steam Pub, Bay, blackrock, states, East Cobb, Soka, North City, Beaches, Ring Road, The Oak, hall, The Junction, Port, lakeview East, Ambassador Park, Poolside, Cafe, Home, International Airport, Gym, coast, paradise, South Bank |

ACKNOWLEDGEMENT

This study is supported by the Melbourne Research Scholarship from the University of Melbourne, Australia. This study was also supported by use of the Nectar Research Cloud, a collaborative Australian research platform supported by the NCRIS-funded Australian Research Data Commons (ARDC). Also, thanks to Digital Ocean11 for providing infrastructures needed for the curation of the COV19Tweets Dataset12.

REFERENCES

[1] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, “Electra: Pre-training text encoders as discriminators rather than generators,” arXiv preprint arXiv:2003.10555, 2020.
[2] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, “Unsupervised cross-lingual representation learning at scale,” arXiv preprint arXiv:1911.02116, 2019.
[3] C. A. Davis Jr, G. L. Pappa, D. R. R. De Oliveira, and F. de L. Arcanjo, “Inferring the location of twitter messages based on user relationships,” Transactions in GIS, vol. 15, no. 6, pp. 735–751, 2011.
[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” arXiv preprint arXiv:1810.04805, 2018.
[5] R. Dutt, K. Hiware, A. Ghosh, and R. Bhaskaran, “Savir: A system for real-time location extraction from microblogs during emergencies,” in Companion Proceedings of the The Web Conference 2018, 2018, pp. 1643–1649.
[6] J. Gelernter and N. Mushegian, “Geo-parsing messages from microtext,” Transactions in GIS, vol. 15, no. 6, pp. 753–773, 2011.
[7] Y. Ikawa, M. Enoki, and M. Tatsubori, “Location inference using microblog messages,” in Proceedings of the 21st international conference on world wide web, 2012, pp. 687–690.
[8] S. Kinsella, V. Murdock, and N. O’Hare, “i’m eating a sandwich in glasgow” modeling locations with tweets,” in Proceedings of the 3rd international workshop on Search and mining user-generated contents, 2011, pp. 61–68.
[9] R. Lamsal, “Coronavirus (covid-19) geo-tagged tweets dataset,” 2020. [Online]. Available: https://dx.doi.org/10.21227/fpsb-jz61
[10] R. Lamsal, “Coronavirus (covid-19) tweets dataset,” 2020. [Online]. Available: https://dx.doi.org/10.21227/781w-e4f2
[11] Lamsal, A. Harwood, and M. R. Read, “Addressing the location a/b problem on twitter: The next generation location inference research,” in Proceedings of the 6th ACM SIGSPATIAL LocalRec, 2022, pp. 1–4.
[12] Lamsal, A. Harwood, and M. R. Read, “Socially enhanced situation awareness from microblogs using artificial intelligence: A survey,” ACM Computing Surveys (CSUR), 2022.
[13] Lamsal, A. Harwood, and M. R. Read, “Socially enhanced situation awareness from microblogs using artificial intelligence: A survey,” ACM Computing Surveys (CSUR), 2022.
[14] Lamsal, A. Harwood, and M. R. Read, “Socially enhanced situation awareness from microblogs using artificial intelligence: A survey,” ACM Computing Surveys (CSUR), 2022.
[15] ———, “Twitter conversations predict the daily confirmed covid-19 cases,” Applied Soft Computing, vol. 129, p. 109603, 2022.
[16] J. H. Lau, L. Chi, K.-N. Tran, and T. Cohn, “End-to-end network for twitter geolocation prediction and hashing,” arXiv preprint arXiv:1710.04802, 2017.

11https://www.digitalocean.com
12https://www.digitalocean.com
