




EventMapping: Geoparsing and Geocoding of Twitter Messages in the Greek Language

Gerasimos Razis¹ , Ioannis Maroufidis², and Ioannis Anagnostopoulos¹

¹ Computer Science and Biomedical Informatics Department, University of Thessaly,
35131 Lamia, Greece
{razis, janag}@uth.gr

² School of Electrical and Computer Engineering, National Technical University of Athens,
15780 Athens, Greece

Abstract. The rapid development of technology has changed the way people get informed from traditional media to online ones. Governmental accounts emerged on Twitter for rapid broadcasting of information related to real-time events and incidents. However, the accurate definition of locations of those events is a difficult task given the fact that their description is provided in free text with no predefined format for the efficient resolving of geolocation. This paper proposes the EventMapping framework that aims at identifying and extracting geographic information directly from official Greek governmental Twitter accounts and visualizing it on an interactive map. For the purposes of this work, two methodologies are implemented and evaluated. The best approach reached the level of 0.93 in terms of F1-score as far as the geographical term detection is concerned, as well as 93.5% in terms of accuracy on the resolved coordinates. The source-code of our framework is publicly available in open-source format on GitHub.

Keywords: Geographic Information · Language Processing · Text Analysis · Event Extraction · Geocoding · Geoparsing · Map Visualization · Twitter · Greek Language · Emergency Mapping

1 Introduction

The rapid development of technology in the last 20 years has greatly changed the way the Greek public is informed. In particular, the widespread use of the internet and the adoption of smart mobile phones have contributed to the transition from the traditional mass media, such as television and newspapers, to the modern ones, such as online blogs, and social media. According to Kepios¹, an organization investigating people's digital behaviors, at the start of 2023, 84% of the total population in Greece had access to the internet. It is estimated that the average user spends 6 h a day on the internet, 3 h watching TV (broadcasting and streaming), and 2 h on Online Social Networks (OSNs). Searching for information is the main reason that drives Greek users online (86.1%),

¹ <https://kepios.com/reports>.

filling up spare time is second (67.7%), whereas being updated with the latest events and news comes third (66.5%) [1].

Contrary to other popular OSNs, Twitter focuses almost exclusively on the area of instant broadcasting of information about recent events. Its popularity has led to the creation of governmental accounts, both for individuals and agencies, such as @Potus, the U.S. President, and @DeptofDefense, the U.S. Department of Defense, respectively. The OSNs are often used for managing and reporting emergencies or crises, as the extracted information can be very informative in such events [2, 3] and can be used for a plethora of scenarios, including mapping.

Two important Greek civil protection Twitter accounts are of the Fire Brigade (@Pyrosvestiki) and Hellenic Police (@HellenicPolice). These public agencies post continuous updates on incidents and events regarding their actions and their results. Usually, a textual description of the geographic information where these events took place is present. However, the locations are provided in free text as part of the social message, as either there is no predefined format for posting this type of information in the OSN messages, or the process of embedding this information is complex and requires third-party tools. Moreover, the level of geographic detail is not consistent, as the descriptions of the locations can range from a broad regional unit level to an intersection of roads.

According to Twitter², less than 2% of the total tweets are geotagged, namely containing geographical coordinates. However, simply relying on these coordinates is not efficient, as it would be ambiguous whether they intended to describe the location of the event mentioned in the tweet, or simply the author's location at the time of posting the message. These inherent problems make the visualization task challenging, as it requires the development of an automated process for extracting geospatial information from the text of the OSN messages.

As discussed in [4], a subset of information retrieval is information extraction, instances of which are the geocoding, geoparsing, and geotagging. Geocoding involves the transformation of a well-defined textual description of an address into a spatial coordinate. Geoparsing is often applied prior to geocoding and involves the identification of the terms describing a geographic location in a text, prior to their transformation into spatial coordinates. Finally, geotagging is the direct assignment of spatial coordinates into a content item, typically describing its location.

There are two aims of this study. Firstly, we propose a framework for the identification and extraction of terms describing geographic information related to events from the Twitter messages written in Greek of the Fire Brigade and Hellenic Police, while in parallel visualizing these social posts on an interactive map. To this end, two methodologies are implemented and evaluated in detail. Secondly, the source-code of our EventMapping framework, including the methodologies, collected OSN data, extracted geospatial and geocoding information, along with the evaluation results are open-source and available on our GitHub repository³. Considering the efforts of the Greek governmental agencies to inform the citizens constantly and accurately about the latest updates and actions, we

² <https://developer.twitter.com/en/docs/tutorials/advanced-filtering-for-geo-data>.

³ <https://github.com/jmarou/EventMapping>.

believe that our open-source Geographic Information System (GIS) can further improve information and public awareness.

The remainder of this work is organized as follows. In the next section, an overview of the related studies is presented, describing text-based geoparsing and geocoding approaches. In Sect. 3, the architecture of our service is presented along with the data acquisition process. In Sect. 4, we describe in detail the methodologies and processes towards the identification and extraction of terms describing geographic information related to events and their visualization on a map. In Sect. 5, we analytically present and discuss the results of the proposed framework, by providing the evaluation metrics. Finally, in Sect. 6, we present the conclusions of our study by summarizing the outcomes and proposing future directions.

2 Related Work

The research area of extracting terms that describe geographic information from long documents or short OSN posts and inferring their coordinates has attracted high interest. The authors of [4] developed a geoparsing and geotagging framework using the OpenStreetMap database, along with a language model based on tags and multiple gazetteers. Their approach was evaluated against the publicly available DBpedia Spotlight, the GeoNames geographical gazetteer, and the Google Geocoder API using a manually labelled dataset with Twitter content. The proposed algorithm identifies named entities using a combination of the Stanford Part-of-Speech (POS) tagger and a Regular Expression (RegEx) pattern. The derived geolocation is at a high level (city or country), based on lower-level details. The authors experimented with content in various languages; however, the results were better for text written in English.

Another work relying on the DBpedia Spotlight entity recognizer is [5]. The identified parts are then linked to the DBpedia RDF resources, and the geographic information is extracted. Using a machine learning classifier, the erroneous geographic coordinates found in DBpedia are discarded from the pool of candidate results.

The Work in [6] describes a semi-supervised approach for geolocating Twitter posts and classifying them towards US regions, relying on sparse coding and dictionary learning methodologies. If the coordinates cannot be inferred, a multiclass classification takes place to infer the US region and state of the examined tweet.

A methodology for identifying and geolocating emergency-related Twitter messages and representing them on a map is presented in [7]. The study reports the lack of native coordinates in the OSN posts but highlights the existence of textual geographical references. The posts are enhanced with indirect information from their retweets or replies. This work also relies on a POS tagger for extracting candidate geographic terms, and a rule-based RegEx for further increasing the Named Entity Recognition (NER) recall including typical location names of streets, roads, and so on.

Another framework relying on NER for geoparsing unstructured text and returning geographic information is [8]. It is based on the open-source spaCy⁴ library for identifying the toponyms and the GeoNames gazetteer for inferring the coordinates. A neural network classifier is applied to the candidate coordinates to derive the final ones.

⁴ <https://spacy.io/>.

The authors in [9] relied on Natural Language Processing (NLP) libraries for identifying terms in travel blogs describing geographic areas or places of interest and their spatial relations. GeoNames was used for the geocoding process, whereas the Levenshtein distance metric for associating those terms with the gazetteer's entries.

A study focused exclusively on Greek content is [10], describing a methodology for the semi-automatic geocoding of information found on web pages. Grammar rules were used for identifying the areas in the text where the geographic information is mentioned, whereas external lexicons were employed for correcting spelling mistakes and converting important geographic-related terms into their normalized form. The derived terms were transformed into a standardized format for the geocoding phase, and approximate string matching was used against the predefined database entries. The authors opted to decrease the precision of that process, namely the derived terms to contain more noise, in favor of the recall. Finally, the results were displayed on a map and could be manually corrected.

Another study analyzing Twitter messages mainly in the Greek language with the aim of visualizing on a map the related information is [11]. For the geoparsing task, the grammatical and syntactic rules of the Greek language were considered for the efficient identification of the terms, whereas for the geolocation task a database was used containing descriptive geolocations and their corresponding coordinates.

Similar to many works in the literature, NLP techniques, such as POS tagging, and RegEx, are extensively used in our EventMapping framework as well, and we also relied on a manually annotated dataset for evaluating both our geoparsing, and geocoding tasks. Our research is focused exclusively on content written in Greek and in identifying fine-grained locations (e.g., at a road or crossroad level). Therefore, we cannot use datasets mentioned in the literature that contain geotagged posts due to the complete absence of such detailed information, especially for areas of Greece. Finally, as in [7], our goal is to avoid using pre-loaded data, gazetteer lookups, and training tasks for further improving the accuracy of our framework. The EventMapping service is designed to be lightweight and with minimal dependencies.

Compared to the related literature, our study differs in two important aspects. Firstly, our research is focused exclusively on OSN content written in Greek, a language generally not well-supported by the established NLP tools. Due to that, we opted to release our EventMapping and the described methodologies as an open-source project. Secondly, where possible, our framework can identify fine-grained textual locations (e.g., at a road or crossroad level) in the OSN messages and their resulting coordinates, rather than providing broader-level information (e.g., on a county or region level). Moreover, we do not rely on any offline gazetteers, given that the Greek areas are under-represented and information on a road level is almost entirely absent.

3 EventMapping: Architecture and Data Acquisition

Our proposed EventMapping service consists of multiple interconnected components organized in a three-layered architecture, presented in Fig. 1 along with the external services and relevant data flows. The architectural layers are decoupled to facilitate the future expansion and maintenance of the service, while the implementation and technical dependencies of each layer are kept locally. Specifically, the "Data" layer is represented

by the green highlighted rectangle and consists of the dedicated “Persistence Storage” component. This is an SQLite relational database responsible for storing and accessing all relevant data, such as the collected OSN content, the event types, and the geographic location, including the actual coordinates.

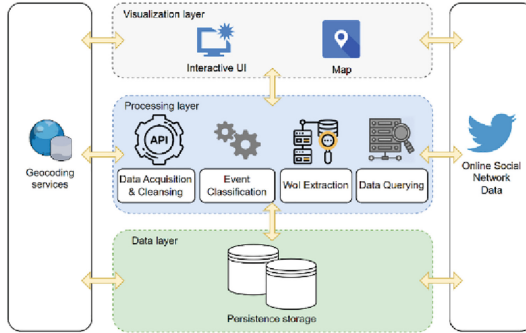


Fig. 1. The three-layered architecture of our service.

The “Processing” layer, represented by the blue highlighted rectangle, is responsible for all business logic, including OSN data collection, cleansing, transformation, querying, extraction, and enrichment tasks. Specifically, the layer was developed using the Python programming language and consists of four components: (a) the “OSN Data Acquisition & Cleansing” component, responsible for retrieving the OSN content, as well as for converting the terms describing geographic information into coordinates leveraging external geocoding services (Sect. 4.3); (b) the “Event Classification” component, where the type of the event described in a tweet is identified (Sect. 4.1); (c) the “Woi Extraction” component, involving the detection of the key terms related to geographic information, referred to as “Words of Interest” (Wois) (Sect. 4.2); and (d) the “Data Querying” component, handling the database-oriented tasks. All types of data, and especially the spatial information, are converted into the format required by the “Persistence Storage”, so that the content can then be visualized on the interactive interface and the map.

Finally, the “Visualization” layer is represented by the grey highlighted rectangle and consists of the “Map” component, responsible for visualizing the OSN events on a map, accessible via the web-based “Interactive User Interface (UI)” component. This layer has been implemented using the ReactJS framework, along with a Leaflet map for presenting the geospatial information and the derived coordinates of the tweets. Flask was used as the main web application framework.

As it can be seen in the right-hand side of Fig. 1, the collected Twitter content is accessible by the appropriate components of all layers, facilitating its parsing, analysis, or enrichment. The latter is achieved through the integration of geocoding services, as seen in the left-hand side of Fig. 1.

The tweets are collected from the official Twitter API. For our EventMapping service, we are only interested in the tweets from the Greek governmental accounts of the Fire Brigade (@Pyrosvestiki) and the Hellenic Police (@HellenicPolice), therefore we relied

on the dedicated endpoint returning up to the 3,200 most recent tweets of a specific account, along with their metadata. Despite that the available metadata of each tweet contains the geographical information, this field was not utilized in our work, since neither is populated by the two examined accounts, nor its existence guarantees that it would reflect the location of the announced event and not the location of the author of the post.

The Twitter API returns the data in UTF-8 (Unicode Transformation Format). However, analyzing text to extract the terms related to geospatial information requires it to be in a suitable format. Therefore, the collected original content is pre-processed according to those rules so that the text can be then analyzed by our geospatial information detection methods (Sect. 4). The pre-processing step accepts as input the original text in UTF-8, applies a series of RegEx, and returns its cleansed form in the same encoding.

In the context of our study, a total of 3,741 tweets were collected and analyzed, from the 8th of May 2020 to the 19th of October 2022 from the account of @HellenicPolice, and 3,284 tweets from the 7th of April 2021 to the 20th of October 2022 from the account of @Pyrosvestiki. The raw, pre-processed, translated, and geolocation data are available in the GitHub repository³ of EventMapping.

4 Event Classification and Geographic Terms Identification

In this section, we analytically present the classification process of the events detected in the tweets, the two methodologies implemented toward the identification of WoIs, as well as the three employed geocoding services for transforming the text into the actual coordinates. The experimental results are presented in Sect. 5.

4.1 Event Classification

As already mentioned, prior to processing the tweets for geographical information, these are first categorized according to the described event. Indicative categories are fire updates, search & rescue operations, and arrests. This classification is performed not only for statistical reasons, but also for filtering any events that we know beforehand do not contain geographic information, such as those referencing the suppression of an electronic crime. In case that a tweet is not classified into any of the available categories, it is placed into the default “Miscellaneous” category.

The categories were derived by considering the semi-supervised examination of approximately 3,000 tweets, and the size of each category, namely the number of tweets belonging to it. An indicative indicator for the improvement of categorization is the reduction of the number of tweets belonging to the default category while simultaneously increasing the percentage of correct categorizations. Eventually, a dictionary of terms was created for each category.

For classifying the social content, the original corpus of the tweets was used, along with the authoring government agency account. Each tweet is assumed to belong to a single category, with very few exceptions. The classifier is based on finding terms from a tweet related to each category’s dictionary using RegEx.

4.2 Geographic Terms Identification

Once the tweets have been categorized against the described event type, those of the valid categories are analyzed for detecting the WoIs related to geographical information, such as names of streets, cities, and so on. Specifically, two distinct methodologies were implemented, which are analyzed below.

Regular Expression Matching. Our first methodology for detecting WoIs related to geographical information is based on the creation of a complex pattern relying on RegEx, namely language-specific heuristics. According to the evaluation of [4] for a similar task, such approaches lead to superior results in identifying the desired information. Considering that the examined governmental accounts post exclusively in Greek, the derived pattern is based on the main syntactic rules of the Greek language. This pattern consists of two basic “capturing groups”.

The first group consists of many alternative words that indicate the existence of toponyms such as “street”, “municipality”, “in”, and so on. The structure of the first group is such that most words are ignored as they do not contribute to the final stage, which is the geocoding (Sect. 4.3). For example, the terms “prefecture of Ilia” and “Ilia” return the same results to most geocoders (“Ilia” is a prefecture of Greece). The second group consists of either contiguous words beginning with a capital letter, or words of one or two letters followed by a period and then contiguous words beginning with a capital letter. Thus, it is possible to identify toponyms such as “N. Makri” (“N” stands for “New”), “L. Alexandras” (“L” stands for “Leoforos”, namely “Avenue” in English), or “Ag. Andreas” (“Ag.” stands for “Agios”, namely “Saint” in English).

However, the aforementioned basic RegEx pattern often returns undesired results, as it detects terms that start with a capital letter but are not toponyms or relate to the location of the event (e.g., denoting where a police station is located). For this purpose, a second pattern was developed for rejecting such cases, based on our findings of how spatial information is described in the posts of the two examined accounts.

NLP on Greek content. The second methodology for detecting WoIs related to geographical information is based on one of the most widespread open-source libraries for NLP spaCy (See footnote 4). The library uses state-of-the-art neural networks for NLP tasks, such as labeling, analysis, NER, and text classification, while offering pre-trained models that users can leverage to train the models on different data sets. According to [12], from an overall viewpoint, NLTK and spaCy have similar results in terms of precision and recall. Moreover, spaCy supports more than 70 languages while offering pre-trained models for 23 languages, including Greek, having been trained on content from news and mass media. In this work, we relied on the most accurate of the available pre-trained model⁵ for the Greek language, in terms of f-score.

The OSN text is parsed using that model for the NER task, and the entity categories are assigned to the terms. In our case, we are only interested in the geopolitical entities, represented by the GPE label, such as cities, countries, regions, and states, and not in generic geographical features, represented by the LOC label, such as natural and/or

⁵ https://spacy.io/models/el#el_core_news_lg.

human-made landmarks and structures. Therefore, the annotated terms are filtered so that only the appropriate entities are extracted. Finally, these terms are concatenated with a blank character, and the resulting string is subjected to the same WoI rejection processing as in the “Regular Expression Matching” method.

4.3 Geocoding Services

Geocoding is the process of converting textual geographical information, such as a pair of coordinates or a toponym, into a location on the earth’s surface. The extracted location can then be used for representation on a map or for spatial analysis.

Within our EventMapping framework, the geocoding step is solved by relying on existing geocoding services. For data quality and evaluation reasons, two different services were used, one commercial and one free. These are Esri (ArcGIS Geocoder) and Nominatim (powering OpenStreetMap), respectively. The geocoding process was automatically performed by submitting the requests and retrieving the results using the official APIs of those services.

5 Evaluation Metrics

For evaluation purposes, two datasets were created and manually annotated, consisting of:

- 100 tweets from each Twitter account including all event types for evaluating the event classifier (Sect. 4.1).
- 100 tweets from each Twitter account containing geographical information, for evaluating a) the derived WoIs of the two geographic information extraction methodologies (Sect. 4.2), and b) the final outcome of the framework, namely the corresponding coordinates of the extracted terms.

5.1 Event Classifier Evaluation

As already mentioned in Sect. 4.1, the categories of the events were derived by considering the semi-supervised examination of approximately 3,000 tweets, whereas a dictionary of terms was created for each category. Each tweet is assumed to belong to a single category. The classifier is based on finding terms from a tweet related to each category’s dictionary, using RegEx.

For evaluating the event classifier, 100 tweets were selected from each Twitter account including all event types, and their event category was manually assigned. Then, this category was compared against the automatically assigned one. The combination of the categories’ dictionaries with the RegEx achieves an accuracy of 98% for both examined governmental Twitter accounts.

5.2 Geographic Information Identification Evaluation

Geographic Term Extraction Evaluation. For evaluating the two geographic information extraction methodologies (Sect. 4.2), 100 tweets were selected from each Twitter

account with categories that are likely to contain the location of the described event. For each of these tweets, all terms describing the location of the event were manually identified and stored, separated by a space. Then, these ground truth values were compared against the automatically derived ones (i.e., the WoIs) from the two methodologies using the precision, recall, and F1 metrics.

Precision is defined as the number of WoIs correctly detected by a methodology, namely terms belonging to the ground truth, out of the total WoIs extracted. Recall is defined as the number of WoIs correctly detected by the method, out of the total terms contained in the ground truth. The F1-score is the harmonic mean of precision and recall, representing both metrics in one value. Table 1 presents the results of the evaluation metrics of the two geographic information extraction methodologies, both on an individual account level and combined on all ground truth data.

As the evaluation metrics showcase, the “Regular Expression Matching” methodology outperforms the other approach in all three metrics, which is in line with the findings of [4]. The values of the three metrics are stable and very high for both examined accounts, namely approximately 0.91 for @HellenicPolice and approximately 0.95 for @Pyrosvestiki, while approximately 0.93 overall. This is expected, since the regular expressions are tailor-made rules derived from a) patterns of the main syntactic rules of the Greek language, b) patterns of Greek toponyms, c) the writing patterns of the examined governmental accounts, and d) the exclusion rules for fire brigade or police station locations not being related to the actual location of an event.

The “NLP on Greek content” methodology despite relying on the spaCy library which is trained on Greek content for NLP and NER tasks, in many cases it fails to categorize street names containing people’s names as geographic entities (GPE label). The F1 value on the combined dataset is approximately 0.78. The values of the three metrics show minimal variation in the case of @HellenicPolice, approximately 0.82, however greater variation for @Pyrosvestiki, as the Recall (approximately 0.72) is quite reduced compared to Precision (approximately 0.8). This is due to this account sharing street names containing people’s names, which spaCy fails to categorize as geographic entities. The results of this methodology have been improved by adding a layer of custom rules for rejecting geospatial abbreviations and undefined areas.

Coordinates Evaluation. The evaluation of the geocoding outcome, namely the corresponding coordinates of the extracted terms denoting geographical information, was performed using the same manually annotated ground truth dataset as in the previous section. For each tweet all terms referring to the location of the event were manually identified. Then, these ground truth values were submitted to Esri’s geocoding service, and the returned coordinates were stored. Finally, we compared these ground truth coordinates with the ones derived from the WoIs of the two methodologies.

To this end, a process was implemented that accepted as input two sets of coordinates, one for the ground truth location and one for each of the two methodologies. The aim is to measure the distance between the two points, expressed by their latitude and longitude in WGS 84 geometry. In case that it is less than 200 m, the resulting geolocation of the examined methodology is regarded as correct. This number derives from an empirical evaluation, and balances cases were: (a) only the broad names of cities or regions are mentioned, since all geolocation services return coordinates in a very small radius, (b)

Table 1. The evaluation metrics of the four geographic information extraction methodologies.

Metric / Account	@HellenicPolice	@Pyrsovestiki	Combined
<i>Regular Expression Matching</i>			
Precision	0.917	0.953	0.935
Recall	0.918	0.958	0.938
F1-score	0.912	0.953	0.933
<i>NLP on Greek content</i>			
Precision	0.827	0.804	0.815
Recall	0.830	0.721	0.776
F1-score	0.824	0.742	0.783

specific names of roads are mentioned with or without additional geographical details, which may span quite a few hundreds of meters.

Table 2 presents the results of the evaluation metrics of the two geographic information extraction methodologies both on an individual account level and combined on all ground truth data.

As the evaluation metrics showcase, the “Regular Expression Matching” methodology outperforms the other approach both on the individual and combined dataset level, with a weighted accuracy of 93.5%. This confirms that it not only identifies the most appropriate WoIs, but also the most important ones in terms of geocoding. Despite that the accuracy of the “NLP on Greek content” methodology on the @HellenicPolice dataset is 82%, it fails to achieve such a score for the @Pyrsovestiki dataset (62%) leading to a weighted average accuracy of 73%.

As it can be observed from Table 2, the NLP-based solution has a high degree of variation, with low accuracy in the content of @Pyrsovestiki. This is due to the level of detail provided by the two accounts: @Pyrsovestiki in most cases reports the location of the event in detail (e.g., at a road or crossroad level), whereas @HellenicPolice usually reports a broader area. The latter is easier to be defined by an NLP approach.

Table 2. The accuracy of the derived coordinates against the ground truth coordinates.

Methodology / Account	@HellenicPolice	@Pyrsovestiki	Combined
Regular Expression Matching	94%	93%	93.5%
NLP on Greek content	82%	64%	73%

6 Conclusions and Future Work

There are three aims of this study. Firstly, we propose an end-to-end framework for the identification and extraction of terms describing geographic information related to events from the Twitter messages of the Greek Fire Brigade and Hellenic Police, while in parallel visualizing these social posts on an interactive map. To this end, two methodologies are implemented and evaluated in detail. Secondly, the source-code of our EventMapping framework, including all described methodologies, collected OSN data, extracted geospatial and geocoding information, along with the evaluation results are open-source and available on our GitHub repository (See footnote 3). Considering the efforts of the Greek governmental agencies to inform the citizens constantly and accurately about the latest updates and actions, we believe that our open-source GIS can further improve information and public awareness.

The three-layered architecture of our GIS has been developed in an expandable approach so that its maintenance is facilitated, and additional functionality can be effortlessly incorporated. Moreover, each of the individual layers and components handling tasks such as receiving information, processing, storing, and analyzing tweets, as well as the interactive web interface are building blocks that can be used separately in new applications.

The main research contribution of our study is the identification and extraction of terms found in Greek tweets describing geographic information related to events. To this end, two distinct methodologies were implemented, optimized for posts specifically from the examined Greek governmental accounts. The methodologies were thoroughly evaluated (Sect. 5.2) against two manually annotated datasets. It was revealed that the NLP-based solution was outperformed by the rule-based approach. Specifically, the “Regular Expression Matching” methodology achieved the best results, with an F1-score of 0.93, whereas the NLP-based approach achieved an F1-score of 0.78. Moreover, we evaluated the coordinates derived from the extracted terms of these methodologies against the manually annotated ones (Sect. 5.2), and the “Regular Expression Matching” methodology outperformed the other approach with a weighted accuracy of 93.5%. This confirms that it not only identifies the most appropriate terms, but also the most important ones in terms of geocoding.

Compared to the related literature, our study differs in two important aspects. Firstly, our research is focused exclusively on OSN content written in Greek, a language generally not well-supported by the established NLP tools. Due to that, we opted to release our EventMapping and the described methodologies as an open-source project (See footnote 3). Secondly, where possible, our framework can identify fine-grained textual locations (e.g., at a road or crossroad level) in the OSN messages and their resulting coordinates, rather than providing broader-level information (e.g., on a county or region level). Moreover, we do not rely on any offline gazetteers, given that the Greek areas are under-represented and information on a road level is almost entirely absent.

Going forward, our plan is to extend this study in two areas. Firstly, we intend to implement and evaluate two additional methodologies towards the identification of terms describing geographic information. One involves the detection of words starting with a capital letter, based on the grammatical rule of words denoting locations and names, while the second involves the application of a NER on the translated content from Greek

to English, since all existing NER systems are trained on English content. Secondly, we aim to introduce the Google Maps geocoding service on top of the examined ones and compare the efficiency of the three geocoding services against a manually annotated dataset.

Acknowledgments. We acknowledge support of this work by the project “Par-ICT CENG: Enhancing ICT research infrastructure in Central Greece to enable processing of Big data from sensor stream, multimedia content, and complex mathematical modeling and simulations” (MIS 5047244) which is implemented under the Action “Reinforcement of the Research and Innovation Infrastructure”, funded by the Operational Programme “Competitiveness, Entrepreneurship and Innovation” (NSRF 2014–2020) and co-financed by Greece and the European Union (European Regional Development Fund).

References

1. DataReportal, <https://datareportal.com/reports/digital-2023-greece>, last accessed February 2023
2. Castillo, C.: Big crisis data: Social Media in Disasters and Time-Critical Situations. Cambridge University Press (2016). <https://doi.org/10.1017/CBO9781316476840>
3. Avvenuti, M., Cresci, S., Del Vigna, F., Fagni, T., Tesconi, M.: CrisMap: a Big Data Crisis Mapping System Based on Damage Detection and Geoparsing. *Inf. Syst. Front.* **20**(5), 993–1011 (2018). <https://doi.org/10.1007/s10796-018-9833-z>
4. Middleton, S., Kordopatis-Zilos, G., Papadopoulos, S., Kompatsiaris, Y.: Location Extraction from Social Media: Geoparsing, Location Disambiguation, and Geotagging. *ACM Trans. Inf. Syst.* **36**(4), 1–27 (2018). <https://doi.org/10.1145/3202662>
5. Avvenuti, M., Cresci, S., Nizzoli, L., Tesconi, M.: GSP (Geo-Semantic-Parsing): Geoparsing and Geotagging with Machine Learning on Top of Linked Data. In: Gangemi, A., et al. (eds.) *ESWC 2018. LNCS*, vol. 10843, pp. 17–32. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-93417-4_2
6. Cha, M., Gwon, Y., Kung, H.: Twitter geolocation and regional classification via sparse coding. In: *Ninth International AAAI Conference on Web and Social Media*, vol. 9, no. 1, pp. 582–585 (2015). <https://doi.org/10.1609/icwsm.v9i1.14664>
7. Scalia, G., Francalanci, C., Pernici, B.: CIME: Context-aware geolocation of emergency-related posts. *GeoInformatica* **26**(1), 125–157 (2021). <https://doi.org/10.1007/s10707-021-00446-x>
8. Halterman, A.: Mordecai: Full Text Geoparsing and Event Geocoding. *Journal of Open Source Software* **2**(9), (2017). <https://doi.org/10.21105/joss.00091>
9. Skoumas, G., Pfoser, D., Kyrellidis, A., Sellis, T.: Location Estimation Using Crowdsourced Spatial Relations. *ACM Trans. Spatial Algorithms Syst.* **2**(2), 23 pages (2016). <https://doi.org/10.1145/2894745>
10. Angel, A., Lontou, C., Pfoser, D., Efentakis, A.: Qualitative geocoding of persistent web pages. In: *16th ACM SIGSPATIAL international conference on Advances in geographic information systems*, pp. 1–10. Association for Computing Machinery, New York, NY, USA (2008). <https://doi.org/10.1145/1463434.1463460>
11. Arapostathis, S.G.: A Methodology for Automatic Acquisition of Flood-event Management Information From Social Media: the Flood in Messinia, South Greece, 2016. *Inf. Syst. Front.* **23**(5), 1127–1144 (2021). <https://doi.org/10.1007/s10796-021-10105-z>

12. Schmitt, X., Kubler, S., Robert, J., Papadakis M., LeTraon, Y.: A Replicable Comparison Study of NER Software: StanfordNLP, NLTK, OpenNLP, SpaCy, Gate. In: Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS), pp. 338–343, Granada, Spain (2019). <https://doi.org/10.1109/SNAMS.2019.8931850>