A Domain-Independent Method for Thematic Dataset Building from Social Media: The Case of Tourism on Twitter

Maxime Masson, Christian Sallaberry, Rodrigo Agerri, Marie-Noelle Bessagnet, Philippe Roose, Annig Le Parc Lacayrelle

A domain-independent method for thematic dataset building from social media: the case of tourism on Twitter

Maxime Masson[1], Christian Sallaberry[1], Rodrigo Agerri[2], Marie-Noelle Bessagnet[1], Philippe Roose[1], and Annig Le Parc Lacayrelle[1]

[1] LIUPPA, E2S, University of Pau and Pays Adour (UPPA)
{maxime.masson,christian.sallaberry}@univ-pau.fr
{marie-noelle.bessagnet,philippe.roose,annig.lacayrelle}@univ-pau.fr
[2] HiTZ Basque Center for Language Technologies - Ixa NLP Group, University of the Basque Country (UPV/EHU)
rodrigo.agerri@ehu.eus

Abstract. In this article, we propose a generic method to build thematic datasets from social media. Many research works gather their data from social media, but the extraction processes used are mostly *ad hoc* and do not follow a formal or standardized method. We aim at extending the processes currently used by designing an iterative, generic and domain-independent approach to build thematic datasets from social media with three modulable dimensions at its core: spatial, temporal and thematic. We experiment our method using data extracted from Twitter to build a thematic dataset about tourism in a highly touristic region. This dataset is then evaluated using both quantitative and qualitative metrics to highlight the value of this method. The application to this use case shows the effectiveness of our domain-independent method to generate thematic datasets from Twitter data.

Keywords: Social Media · Dataset Building · Social Web Analysis · Computational Social Science · Natural Language Processing

1 Introduction

Recently, we have seen a significant growth in available data sources covering many topics and the rise of user-generated content. Tourism in particular represents one of the biggest economic sectors in the world in which accurate thematic datasets are critical to better analyze and decipher trends [1].

In the domain of tourism, large datasets can have many practical use cases. On the one hand, they can be used for the purpose of supporting the decision-making process of tourism stakeholders for the improvement, development and planning of touristic cities and areas. This is done by analyzing the data to better understand the practice and needs of tourists. Such analysis is particularly useful for companies specialized in tourism marketing (such as *Destination Marketing Organization, DMO*) where understanding the desires and expectations

of tourists is key. On the other hand, tourism data can be analyzed for tourists themselves by building recommender systems. Those systems analyze the types of practices of a large number of tourists to be able to recommend better suited places, activities or tourist itineraries. Furthermore, different types of sources can be used to extract touristic data. Historically, these are mainly databases with 2 categories that stand out: (1) commercial databases, such as those from Online Travel Agencies (OTAs) and, (2) public databases, for example government issued or crowd sourced ones. The latter is a part of what is called User-Generated Content (UGC). In recent years, this category of data has grown significantly, ranging from social media (*Twitter, FourSquare*) to review sites (*TripAdvisor*).

As part of a local project, we needed to extract, analyze and present data focused on a specific theme defined by a domain expert (in our case, *tourism*). While many research works concentrate on proposing fully generic and adaptable processing pipelines to extract knowledge from flows of social media data [6] (such as NLP modules calibrated for short, informal messages), they do not propose a generic methodology to build such thematic datasets. In contrast to those previous approaches, the work presented here focuses on the upstream step of the data collection process. It aims at consolidating existing solutions when it comes to social media based Information Extraction (IE). Indeed, this contribution consists of proposing a domain-independent method to obtain high-quality data from social media to build thematic datasets. It is generic and based on an iterative and multi-dimensional (*spatial, temporal, thematic*) filtering process. It was hypothesized that representing the theme of the dataset to be built up in the form of a vocabulary can contribute to (1) an efficient thematic filtering of posts and to (2) the development of a domain independent process.

The article is organized as follows. Firstly, we introduce our project's motivations and the requirements our extraction method must meet. Secondly, we review the state of the art of the approaches used to gather data out of social media with a focus on Twitter data. Thirdly, our contribution is presented: a domain-independent, generic method for thematic dataset building from social media. We finally experiment and evaluate our method on a local touristic case.

2 Motivation

This work, carried out in the framework of a multilingual (*English*, *French*, *Spanish* and *Basque*) project, aims to collect, process, analyze and then value social media data related to the practice of tourism, visitor flows and the use of cultural heritage in the *Basque Country*, a cross-border highly touristic area.

We decided to adopt a trajectory-based analysis and therefore build multidimensional trajectories from local visitors. Our trajectory model has 3 dimensions at its core. The spatial dimension (*where*) (1) refers to the set of places (*municipalities, natural areas, etc.*) visited by tourists. The temporal dimension (*when*) (2) can be seen at different levels: the period of the trip (*season, year, month, time interval etc.*) as well as the date and duration of each activity performed. Finally, the thematic dimension (*what*) (3) is purely semantic, it describes the

"*what*" of the tourist practice, such as the activities performed, the conditions of the trip, the tourist's feelings, whether they were accompanied or not, etc. We aim at moving beyond the well-known concept of spatial tracks and instead focus on trajectories where the thematic dimension has a greater weight than the spatial one, where places could be represented by themes and region by cluster of themes. **Fig. 1** shows an example of a trajectory in a tourist thematic space.

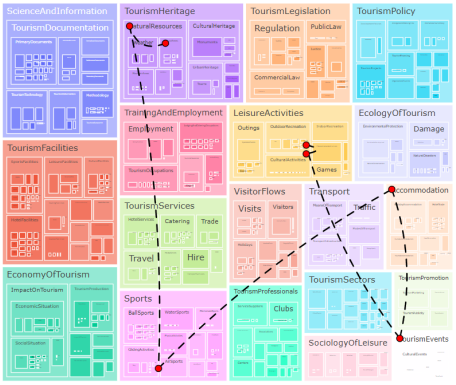

Fig. 1. Trajectory of a visitor in what could be the tourism thematic space.

With thousands of trajectories, we could run pattern analysis on it to detect affinities between types of tourist activities, categories of cultural heritage or even recurring sequences of activities. To build these thematic trajectories, social media is chosen as our primary source of data for several reasons: (1) the ease of access, no need for time-consuming collection campaigns, (2) the diversity, most tourism aspects are covered and (3) the massive amount of data available.

Although our project is mostly focused on the tourism domain, we have already identified future needs for the generation of datasets related with other specific domains such as wine culture or education. It therefore appeared necessary to move towards a method that is as independent of the domain as possible.

3 Related Work

Due to the extensive amount of research work on the subject, we focus on Twitter data that we have chosen as a main data source for our project. For the **spatial dimension (1)**, two different approaches are generally used and sometimes combined to obtain better results. First, filtering can be performed on the **spatial metadata** attached to the post, either by providing the social media API with a bounding box encompassing the area of study, or by basing it on a precise location (*referenced with a latitude/longitude*) with a radius around it (*location nearness*) [7]. The advantage of this metadata-based approach is that

it is extremely accurate, as the location is generally determined using the GPS of the device. The biggest problem that arises is the low amount of posts that have metadata attached. For Twitter, it was estimated that around 1% of tweets have spatial metadata attached to them [9]. Thus, relying solely on metadata misses many potentially relevant posts. The second method is applied to the content of the post and is done via toponym filtering [8, 7]. A list of place names and their associated abbreviations is compiled and all posts containing them are extracted. The big drawback of this method is the potentially large amount of noise. Disambiguation approaches are used to mitigate this problem, for example by combining this list of toponyms with exclusion lists [10].

For the **temporal dimension (2)**, the extraction is usually done according to the timestamp. It is an easy method to implement and usually quite accurate. Exchanges on social media are often done in real time, so it is not necessarily mandatory to set up a complex temporal entity extraction system.

When it comes to the **thematic dimension (3)**, multiple approaches are used. Content-based approaches use thematic keywords directly related to the theme of study (such as event names [8]), the use of too specific keywords can restrict too much the number of returned posts, so some research works associate several words together to establish filtering rules. Some social media have the concept of *hashtag* allowing to identify topics of discussion, which have often been used as a filtering tool [3]. Other thematic filtering methods are applied to the metadata and include: the language of the post, its source, etc. To reduce noise, some research work use only the posts from a pre-selected list of accounts known to validate certain desired criteria (e.g., known for speaking regularly about a specific topic [2]). Associated replies and comments can also be extracted [3].

Although, in recent years, various efforts have been made to design more generic processing techniques for social media data [6], when it comes to dataset building, each research project usually comes with its own extraction and filtering flow. While some common techniques are shared, there is not yet a fully generic and domain-independent procedure for dataset building from social media.

4 A new method for thematic dataset building from social media

We propose a method for creating thematic datasets from social media. It aims at formalizing a generic approach to extract social media-sourced data related to a given domain and possibly a given time period or spatial area. This method is designed around the following properties:

- **Multi-dimensional**: the method is based on 3 dimensions: spatial, temporal, thematic. One can combine two, three of them or only use one.
- **Generic**: it can be implemented with any social media that has a post system. This genericity also extends to languages.
- **Domain-independent**: it supports any target theme for the dataset.

– **Iterative, incremental**: the method is designed around an iterative and incremental process, namely, each iteration aims at refining the following iterations to have a dataset as qualitative as possible with a minimal noise.

Before proceeding with the collection process (the steps presented in **Fig. 2**), it is necessary to define the future dataset along the **spatial** (*territorial footprint of the data*), **temporal** (*temporal scope*) and **thematic** (*the theme, the semantics as of a vocabulary, thesaurus or ontology*) dimensions. According to our review of the state of the art, these dimensions cover the majority of use cases in dataset building. The presence of all these dimensions is not mandatory. We could therefore have only *tempo-thematic* or *spatio-thematic* datasets.

Our method being **iterative and incremental**, users can refine the dimensions later in the process, until they are satisfied with the obtained dataset. This process is semi-automatic. After this preliminary definition step, it is recommended to define a *calibration dataset*, It is a subset of the main one with a more restrictive definition used to calibrate the main, wider collection process.

4.1 Filtering the Flow of Posts

Once the dataset have been defined, we can now move on to the collection process. It is applied sequentially to 2 sets of posts with different features.

1. **Geotagged posts**: posts whose authors have activated GPS location (about 2% [9] for Twitter). This reduced set of posts is handled first (**Fig. 2**, ①).
2. **All posts**: posts containing a text (no image-only posts) (**Fig. 2**, ②).

The flow of associated media (**Fig. 2**, ③) is a future research axis and will therefore not be discussed in this article.

Each set is extracted following the procedure described in **Fig. 2**. The order of the steps is only indicative. In this sense, it is necessary to think about which steps can be delegated to the internal filtering system of the social media and which ones must necessarily be carried out locally (the latter have to be processed last). There is no universal answer to this question, it depends on the extent of the search functionality of the social media and the dimensions' complexity.

Pre-processing aims to exclude accounts, keywords or hashtags that we know should not appear in our final dataset while being prone to fit within our dimensions, for example: excluding professional, institutional or promotional accounts, excluding problematic keywords, or excluding certain languages. This step is especially useful to exclude places with the same name but actually unrelated to each other (*toponymic homonyms*), which could distort the spatial filtering process. Usually, these criteria are not known at the beginning of the process so this step is empty, although it will be filled in future iterations.

The following steps are optional and depend on the dataset definition. For a more efficient collection process, it is advisable to order the dimensions from the one believed to be the most restrictive, to the less restrictive (*in order to process as few unnecessary posts as possible*) and to delegate as much filtering as possible to the social media native API system.

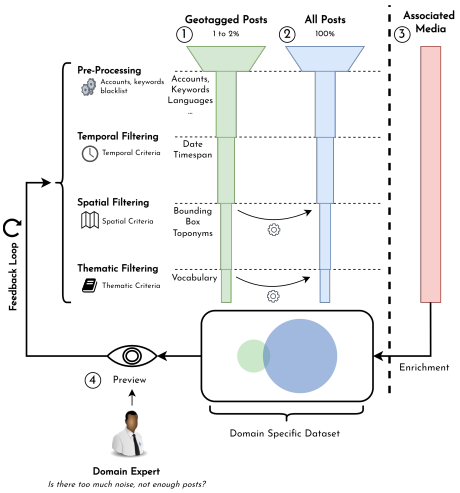

Fig. 2. Overview of the extraction process.

Temporal filtering is done using the timestamp of the posts. This method has been extensively used and covers most of the use cases. Indeed, social media are instantaneous informal exchange zones, i.e. users generally talk about the present moment. In specific cases, a temporal entity detection system could be applied on the content of the posts. This would allow to be more precise in cases where users talk about past or future events (*yesterday, next week*).

For **spatial filtering** the approach to be used differs depending on the two sets of posts described above. In the case of geotagged posts, we propose to check whether the position of the post is contained within a bounding box. This method is highly accurate because it is based on the devices' GPS. Moreover, it can usually be offsetted to the social media API. For the posts without spatial metadata attached, the approach is a bit different. We rely on the post content to do the selection. A list of toponyms contained within the area of study is provided as input and is matched within post content to determine which ones are about this area. The risk of noise is higher with this approach, but our iterative method allows to refine this step further in the process.

The objective of **thematic filtering** is to keep only the posts which are related to the theme, with the latter being defined using a given vocabulary. It is done by aligning this vocabulary with the post content (*entity linking* [4]).

4.2 Dataset Preview and Iteration

Each time an iteration of this process is performed, a dataset is obtained. We can then preview the resulting dataset and evaluate it, this step is usually performed by a domain specialist (**Fig. 2**, ④). Their role is to determine by reviewing a set of randomly selected posts, whether there is too much noise or not enough

posts and try to identify certain types of recurring posts to exclude or that are missing. The criteria to add, remove, extend or narrow are then decided and a new iteration of the process can start using those refined criteria, and so on until the final dataset is deemed satisfactory (*feedback loop*).

5 Experimentation and Evaluation

To experiment our methodology, we chose the social media Twitter[3]. Concerning the preliminary definition step, we rely on the *Thesaurus on Tourism & Leisure* of the *WTO*[4]. This resource covers roughly 1,300 touristic concepts. For the purpose of this experiment, we reduce the spatial extent of the data. We wish to gather content from only one specific sub-area which will serve as our **spatial dimension**: the *French Basque Coast*, which is broadly considered to be among the most touristic places in the region. Also, in an effort to show the full range of our dimensions, we will focus solely on the summer of 2019.

We compute **quantitative statistics** on the collected data such as the number of users, tweets, detected concepts or locations. These quantitative metrics will allow us to determine if our method can be implemented on Twitter data and if a consistent number of tweets is collected without having too few posts or excessive noise. Then, we want to evaluate the effectiveness of our thematic filtering process. We take as reference the **context annotations** generated by Twitter. These are labels that Twitter automatically attaches to tweets based on their content. The approach used to create those is not known but among these annotations, one is about *Travel*. After reviewing them, we realized that these annotations were usually quite accurate but that many relevant tweets did not have them (e.g., *Twitter annotates well but doesn't annotate enough tweets*) making it difficult to build large datasets relying solely on them. It therefore seemed relevant to calculate what proportion of tweets suggested as related to travel by Twitter our system select. Lastly, we needed to evaluate whether all other tweets not tagged as "*Travel*" we import are relevant or just noise. We therefore perform a **qualitative** analysis on the resulting dataset. We randomly sampled 20, 50, and 100 posts of the resulting datasets and calculated the thematic accuracy @20, @50, and @100 to check the reliability of our process at each iteration. Two tourism experts will manually analyze their content and annotate whether or not they relate to tourism or not. This qualitative metric is meant to demonstrate the role of our method's multiple iterations in improving the accuracy and evaluate the overall quality of the results.

5.1 Method Implementation

The whole process (Table 1) was developed with the *Tweepy*[5] library used for the interaction with the Twitter API. We performed 3 refinement iterations and we only collected tweets in French, English and Spanish posted in summer 2019.

[3] https://developer.twitter.com/en/products/twitter-api/academic-research
[4] https://www.e-unwto.org/doi/book/10.18111/9789284404551
[5] https://www.tweepy.org

Table 1. Application of the method using our dataset requirements.

		Iteration 1		Iteration 2		Iteration 3	
		Geotagged Posts	**All Posts**	**Geotagged Posts**	**All Posts**	**Geotagged Posts**	**All Posts**
Pre Processing	Criteria	**Lang**: FR, ES, EN. **Blacklist** *retweets, quotes*		− professional accounts		− blacklist G7-related keywords & #	
Temporal Filtering	Criteria	Summer 2019					
	Tweets	> 1 billion tweets					
Spatial Filtering	Criteria	Basque Bounding Box	625 Basque OSM places	Basque Bounding Box	579 Basque OSM places	Basque Bounding Box	550 Basque OSM places
	Tweets	**7,003**	**> 2,700,000**	**6,689**	**148,860**	**6,127**	**59,878**
Thematic Filtering	Criteria	Full WTO Thesaurus	*Too many tweets*	Refined WTO Thesaurus		Refined++ WTO Thesaurus	
	Tweets	**3,447**		**2,390**	**56,968**	**2,098**	**25,281**
Stats	Hashtags (#)	3,730 hashtags		3,620 hashtags	44,411 hashtags	3,341 hashtags	24,263 hashtags
	Users (@)	1,112 users		865 users	30,126 users	796 users	14,114 users
	Places	32 locations		31 locations	194 locations	31 locations	184 locations
	Mapped Concepts	462 concepts		245 concepts	540 concepts	235 concepts	458 concepts

For geotagged tweets, we simply filter on the bounding box of the area of study and get about 7,000 tweets. Those are then thematically filtered using the whole thesaurus vocabulary in conjunction with the IAM Entity Linker [4], a dictionary-based approach for semantic annotation. We get 3,447 tweets as output. For non-geotagged tweets, we use 625 multilingual toponyms extracted from OSM[6] contained within the area of study. In the first iteration, we use all of them indiscriminately to filter and therefore retrieve too many tweets (more than 2.5 million). We decide to stop the process and refine it at the next iteration.

The feedback from the 1st iteration leads us to blacklist professional or institutional accounts (not interesting for our analysis). We also refine the toponym list and remove 46 common place names (such as *Roman Theater*). Lastly, the tourism thesaurus is reduced to exclude some branches deemed irrelevant by our project domain experts. We get about 60,000 tweets at the end of the 2nd iteration and notice a large number of tweets about the G7 an important event taking place in 2019 in the region. A 3rd iteration is carried out with additional pre-processing filters to blacklist G7-related keywords and hashtags. The final dataset is made of about 2,098 geotagged tweets, 25,281 tweets in total belonging to 15,000 users and 458 unique concepts have been found in these tweets.

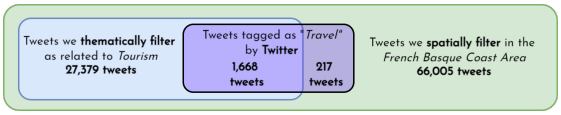

Fig. 3. Number of tweets annotated "*Travel*" by Twitter among those we select.

Out of those: 1,668 have the Twitter "*Travel*" annotation. In other words, only about 6% of the tweets collected were identified as relating to tourism by Twitter, that is not much and it is why we aim at collecting more than those.

[6] https://www.openstreetmap.org/

5.2 Results Analysis

Fig. 3 shows the different sets of tweets of the 3rd (last) iteration (*union of both geotagged non-geotagged ones*) and among these sets of tweets, the proportion of those annotated with the "*Travel*" context annotation by Twitter. We observe that among the tweets from the area of study (66,005), most of those tagged as "*Travel*" by Twitter are selected by our system (1,668 selected, 217 excluded). We get an accuracy of 0.884. We also notice a high number of selected tweets which are not tagged by Twitter (25,711) but that we select. It means that our process detects many more potentially relevant items. This can be rather a positive aspect as Twitter context annotations seem to be missing from a lot of tourism-related tweets. However, we now need to determine whether all those are just noise or other tourism-related tweets that Twitter has not annotated.

For this purpose, a qualitative analysis is set up. 100, 50 and 20 tweets are extracted randomly from the datasets at different stages of the method and manually evaluated by experts to determine whether they have been correctly or wrongfully selected. Table 2 shows the result of the evaluation done by two experts on 20, 50 and 100 tweets, randomly picked just after the thematic filtering for both sets of tweets. We compute the mean thematic accuracy between those two experts and, for the 3rd iteration @100, the associated Cohen's Kappa (κ) which measures the degree of agreement between them to mitigate subjectivity.

Table 2. Qualitative analysis of the results.

		Iteration 1		**Iteration 2**		**Iteration 3**	
		Geotagged	**All**	**Geotagged**	**All**	**Geotagged**	**All**
Thematic Accuracy	(@ 20)	0.75		0.60	0.30	**0.83**	**0.72**
	(@ 50)	0.64		0.60	0.30	0.77	0.74
	(@ 100)	**0.52**		**0.59**	0.35	**0.74** (κ 0.74)	**0.65** (κ 0.48)

The question asked was "*Is this tweet related to Tourism?*". We obtain a mean thematic accuracy ranging from 0.83 (@20) to 0.74 (@100) for geotagged tweets and 0.72 (@20) to 0.65 (@100) for non-geotagged ones. That means, by extrapolation, potentially 65% to 83% of the tweets not selected by Twitter might actually be relevant to the topic of tourism and our assumption that Twitter is not annotating enough content is correct.

We also observe a thematic accuracy @100 starting at 0.52 at the 1st iteration for geotagged tweets which increases to 0.59 and then 0.74 in iterations 2 and 3 which clearly highlight the effect of filter refinement and feedback loop between each iterations. Overall, experts seem to agree on the outcome as shown by the relatively high κ score. Thematic accuracy on all non-geotagged tweets is slightly lower but follow a similar increasing trend. The final dataset thematic accuracy is acceptable but could have been increased even more by doing more iterations, we limited ourself to 3 for this experiment. To go further, we observed the **same** sample of incorrect tweet from the first iteration throughout the method to see what proportion of it would get removed in future ones. We use the set of tweets

used for the @100 thematic accuracy measure of the 1st iteration. The accuracy is average (0.52) which means: out of 100 tweets evaluated, 48 were incorrectly selected. The 2nd iteration removes 27 of them, so 21 are remaining. Then finally, the last iteration leaves 15 remaining. This is consistent with the accuracy we calculated on random samples previously ($\approx$ 70% of correct tweets).

6 Conclusion

We proposed a domain-independent and generic method for building thematic datasets from social media based on 3 dimensions. The objective is to move away from *ad hoc* collection processes and to propose a robust method to build focused datasets. We are thinking of going further by extending our dimensions by taking inspiration from the 5W1H [5] dimensions. The 5W1H is a framework widely used in problem solving and question answering based on the 6 interrogative words: *Who, What, When, Where, Why, How*. We already have the *When* (temporal), *Where* (spatial) and *What* (thematic) but we could imagine other dimensions for the *Who* (the users, the persons they are referring to), the *why* (reasoning behind an action) or the *how* (in what way is it carried out).

References

1. Aguiar, A., Szekut, A.: Big data and tourism: opportunities and aplications in tourism destination management. Applied Tourism **4**, 36 (09 2019)
2. Chiruzzo, L., Castro, S., Rosá, A.: Haha 2019 dataset: A corpus for humor analysis in spanish. In: Proceedings of The 12th Language Resources and Evaluation Conference. pp. 5106–5112 (2020)
3. Cignarella, A.T., Lai, M., Bosco, C., Patti, V., Paolo, R., et al.: Overview of the task on stance detection in italian tweets. In: EVALITA 2020 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. pp. 1–10. Ceur (2020)
4. Cossin, S., Jouhet, V., Mougin, F., Diallo, G., Thiessard, F.: Iam at clef ehealth 2018: Concept annotation and coding in french death certificates. arXiv preprint arXiv:1807.03674 (2018)
5. Han, S., Lee, K., Lee, D., Lee, G.G.: Counseling dialog system with 5w1h extraction. In: Proceedings of the SIGDIAL 2013 Conference. pp. 349–353 (2013)
6. Sathick, J., Venkat, J.: A generic framework for extraction of knowledge from social web sources (social networking websites) for an online recommendation system. International Review of Research in Open and Distributed Learning **16**(2), 247–271 (2015)
7. Scholz, J., Jeznik, J.: Evaluating geo-tagged twitter data to analyze tourist flows in styria, austria. ISPRS International Journal of Geo-Information **9**(11), 681 (2020)
8. Shimada, K., Inoue, S., Maeda, H., Endo, T.: Analyzing tourism information on twitter for a local city. In: 2011 First ACIS International Symposium on Software and Network Engineering. pp. 61–66. IEEE (2011)
9. Sloan, L., Morgan, J.: Who tweets with their location? understanding the relationship between demographic characteristics and the use of geoservices and geotagging on twitter. PloS one **10**(11), e0142209 (2015)
10. Zenasni, S., Kergosien, E., Roche, M., Teisseire, M.: Spatial Information Extraction from Short Messages. Expert Systems with Applications **95**, 351 – 367 (Apr 2018)