Information Processing and Management 56 (2019) 1119–1132

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

# On fine-grained geolocalisation of tweets and real-time traffic incident detection

Jorge David Gonzalez Paule*,[a], Yeran Sun[b], Yashar Moshfeghi*,[c]

[a] School of Computing Science, University of Glasgow, UK
[b] Urban Big Data Centre, University of Glasgow, UK
[c] Department of Computer & Information Sciences, University of Strathclyde, UK

ARTICLE INFO

*Keywords:*
Twitter
Fine-grained geolocation
Majority voting
Information Retrieval
Traffic incident detection

ABSTRACT

Recently, geolocalisation of tweets has become important for a wide range of real-time applications, including real-time event detection, topic detection or disaster and emergency analysis. However, the number of relevant geotagged tweets available to enable such tasks remains insufficient. To overcome this limitation, predicting the location of non-geotagged tweets, while challenging, can increase the sample of geotagged data and has consequences for a wide range of applications. In this paper, we propose a location inference method that utilises a ranking approach combined with a majority voting of tweets, where each vote is weighted based on evidence gathered from the ranking. Using geotagged tweets from two cities, Chicago and New York (USA), our experimental results demonstrate that our method (statistically) significantly outperforms state-of-the-art baselines in terms of accuracy and error distance, in both cities, with the cost of decreased coverage. Finally, we investigated the applicability of our method in a real-time scenario by means of a traffic incident detection task. Our analysis shows that our fine-grained geolocalisation method can overcome the limitations of geotagged tweets and precisely map incident-related tweets at the real location of the incident.

## 1. Introduction

In recent years, social media services have gained increasing popularity within the research community. Specifically, Location-Based Social Networks (LBSN) (Zheng & Xie, 2011), such as Twitter, have become very popular as its data is generated in real-time and can contain spatially fine-grained geolocations (i.e. at a street, building or neighbourhood level). Such characteristics has provided new opportunities for a broad range of real-time applications such as real-time event detection (Atefeh & Khreich, 2015; Crooks, Croitoru, Stefanidis, & Radzikowski, 2013; Sakaki, Okazaki, & Matsuo, 2010; Walther & Kaisser, 2013; Watanabe, Ochi, Okabe, & Onai, 2011; Xia et al., 2014; Zhang, Zhou et al., 2016), sentiment analysis (Baucom, Sanjari, Liu, & Chen, 2013), urban planning (Frias-Martinez, Soto, Hohwald, & Frias-Martinez, 2012), topic detection (Hong, Ahmed, Gurumurthy, Smola, & Tsioutsiouliklis, 2012a), and disaster and emergency analysis (Ao, Zhang, & Cao, 2014; Imran, Castillo, Diaz, & Vieweg, 2015; McCreadie, Macdonald, & Ounis, 2016).

Several of aforementioned applications depend on the availability of sufficient fine-grained geotagged tweets. However, since only a very small sample of tweets in the Twitter stream (1% to 2%) contain geographical information (Graham, Hale, & Gaffney, 2014), the effectiveness of such applications is limited. Thus, to increase this sample, geolocating (or geolocalising) non-geotagged

* Corresponding author.
*E-mail addresses:* j.gonzalez-paule.1@research.gla.ac.uk (J.D.G. Paule), yeran.sun@glasgow.ac.uk (Y. Sun), yashar.moshfeghi@strath.ac.uk (Y. Moshfeghi).

https://doi.org/10.1016/j.ipm.2018.03.011
Received 31 July 2017; Received in revised form 24 March 2018; Accepted 27 March 2018
Available online 07 April 2018

tweets has become an important yet challenging task. In this paper, we tackle this problem by presenting a method for geolocalising non-geotagged tweets at a fine-grained level.[1] To better geolocalise tweets, we propose a novel approach that combines evidence gathered from geotagged tweets that are similar based on their contents to a given non-geotagged tweet.

Other approaches have been proposed in the past to provide fine-grained geolocalisation of tweets; e.g. Kinsella, Murdock, and O'Hare (2011) and Paraskevopoulos and Palpanas (2015). In these works, for each predefined geographical area an aggregation of the geotagged tweets belonging to that area is performed by concatenating their texts into a document. Then, a vector representation of that area is created from the generated document using a bag-of-words approach. To geolocate a given tweet, the most similar area to that tweet is returned based on its content similarity using the generated vectors.

Although approaches mentioned above have provided important insights on how to tackle fine-grained geolocalisation of tweets, due to the noisy nature of Twitter data (Teevan, Ramage, & Morris, 2011), such an aggregation method can diminish the importance of infrequent yet relevant geotagged tweets and affect the accuracy of matching algorithms, decreasing the accuracy of the geolocalisation. In addition, by aggregating the tweets into a single document, the important evidence contained in the meta-data of the tweet is lost. In contrast to previous works, we avoid aggregating the geotagged tweets and treat each one individually as a single document, thus representing each area as multiple bag-of-words vectors. This way, we give each tweet the same chance during the matching process and enable the use of evidence in its meta-data for better geolocalisation.

To tackle the problem of fine-grained geolocalisation of tweets, we adopted a weighted majority voting algorithm. We estimate the geographical location of a given non-geotagged tweet by collecting the geolocation votes of the content-based most similar geotagged tweets to that tweet. In this paper, we refined our previous work (Gonzalez Paule, Moshfeghi, Jose, & Thakuriah, 2017) by incorporating a new weighting function that combines the evidence gathered from the geotagged tweets. These weights are calculated based on the credibility of the users of the geotagged tweet and the degree of content similarity with respect to the non-geotagged tweet. The credibility of the user is calculated as a score that represents the user's posting activity and its relevance to the physical location they are posting from. To validate our approach, we then performed an exhaustive study across three test collections, which contain tweets gathered from two different cities. Our experimental results showed (statistically) significant improvements regarding accuracy and reduction of geographical distance error compared to baselines.

Finally, we extend our previous work (Gonzalez Paule et al., 2017) by studying the applicability of our fine-grained geolocalisation approach in a real-time scenario. We integrated our approach into a real-time traffic incident detection task for geolocalising incident-related tweets. Our analysis demonstrates that our geolocalisation method can overcome the limitations of geotagged tweets, mapping more precisely the incident-related tweets to the real incident locations.

The contributions of this paper can be summarised as follows:

- First, we proposed a novel approach for fine-grained geolocalisation of non-geotagged tweets, which adopts a weighted majority voting algorithm to combine evidence gathered from the content-based most similar geotagged tweets.
- Second, we perform an exhaustive evaluation of our approach on three test sets of geotagged tweets posted on two different cities, including a big dataset gathered from one of the major cities in the USA (i.e. New York).
- Finally, we performed a traffic incident detection task to demonstrate the applicability of our approach in a real-time practical scenario.

The rest of the paper is organised as follows. In Section 2 we discuss previous research and motivate our work. We introduce our approach for fine-grained geolocalisation of non-geotagged tweets in Section 3. Section 4 presents our experimental setup and discusses our results. Finally, we show the applicability of our model in a real-time scenario in Section 6.

## 2. Background

First approaches in the literature on geolocalising social media data (Chang, Lee, Eltaher, & Lee, 2012; Cheng, Caverlee, & Lee, 2010; Eisenstein, O'Connor, Smith, & Xing, 2010; Han & Cook, 2013) addressed the localisation of Twitter users rather than individual social media posts. To achieve this, these approaches extract all the tweets posted by a single user to infer their city or home location. However, not all the tweets of a user are geotagged, which leads to a sparse dataset. As we aim to locate individual non-geotagged tweets, our work can be placed one step before the Twitter user geolocalisation task, reducing the sparsity of the data.

There is a lot of research aiming to identify the problem of geolocalising individual non-geotagged tweets. For example, Schulz, Hadjakos, Paulheim, Nachtwey, and Mühlhäuser (2013) tackled this problem by exploiting different spatial indicators of a tweet – i.e. tweet text or user profile – and mapping them to different geospatial datasets such as *DBpedia Spotlight* or *Geonames*. More recently, other works tackled this problem by dividing the geographical space into areas of a given size and then modelled the language for each area (Hulden, Silfverberg, & Francom, 2015; Kinsella et al., 2011; Paraskevopoulos & Palpanas, 2015; Roller, Speriosu, Rallapalli, Wing, & Baldridge, 2012; Wing & Baldridge, 2011). Then, for a given non-geotagged tweet the most likely area is returned based on the probability that the tweet was issued in that area. However, these studies work at a coarse-grained level of granularity – i.e. zip codes to city or country level. In contrast, the problem we aim to tackle in this work is the geolocalisation of Twitter posts at a fine-grained level – i.e. street or neighbourhood level.

An example of previous research on fine-grained geolocalisation is the work by Kinsella et al. (2011). They attempted to predict

[1] Specifically, fine-grained locations are defined as squared areas of size 1 km in this work.

location from country level to postal code level. As a result, the accuracy of their model decreases significantly when trying to predict at such fine-grained level. Another example of fine-grained geolocalisation is the work by Paraskevopoulos and Palpanas (2015). The authors refined the approach proposed by Kinsella et al. (2011) by dividing the geographical space into fine-grained squares of size 1 km. Finally, Flatow, Naaman, Xie, Volkovich, and Kanza (2015) followed a completely different approach by estimating the geographical centre of word n-grams using a Gaussian Mixture Model.

In this paper, inspired by Paraskevopoulos and Palpanas (2015), we follow the strategy of dividing the city into squares of size 1 km. The work mentioned above perform a concatenation of texts of tweets belonging to each square to represent that area as a single bag-of-word vector. However, in contrast to this work we consider each tweet individually, representing each area as multiple bag-of-word vectors during the prediction process.

Also, our approach takes into account the credibility of tweets. Considering the quality of sources to verify the information generated from them is related to the truth discovery problem (Li et al., 2016). Different algorithms have been proposed to address the problem (Yin, Han, & Yu, 2007). In this work, we have decided to apply a voting approach due to its simplicity and effectiveness.

Some works have attempted to measure the quality of the information from Twitter users (Marshall, Syed, & Wang, 2016; Wang, Marshall, & Huang, 2016; Zhang, Han, Wang, & Huang, 2016) and specifically for event detection and disaster and emergency management (Castillo, Mendoza, & Poblete, 2011; McCreadie et al., 2016). For example, McCreadie et al. (2016) considered the idea of assigning a credibility score to tweets, but for the disaster and emergency detection task. They computed the credibility score using regression models with text features and user information. This credibility score is utilised to inform the user about the veracity/credibility of events derived from social media. We also incorporate the credibility of tweets in our fine-grained geolocalisation approach. But, in contrast to McCreadie et al. , we incorporate this score as a weight for each vote in our adopted majority voting approach.

The majority voting algorithm is a well known, fast and effective strategy widely adopted for prediction and re-ranking tasks (Chiang, Lo, & Lin, 2012; Mosbah & Boucheham, 2015; Rokach, 2010). However, to the best of our knowledge, this is the first time the majority voting is considered to tackle the geolocation of tweets. Next section describes our approach in detail.

## 3. Fine-grained geolocalisation approach

Our proposed approach consists of three stages. First, we divide the geographical area of interest into a grid of 1 km squared areas, and associate each geotagged tweet to an area based on its location. As discussed in Section 2, the grid approach is a popular technique to represent geographical areas at different levels of granularity in the literature (Kinsella et al., 2011; Paraskevopoulos & Palpanas, 2015). Second, we obtained the Top-N content-based most similar geotagged tweets to each non-geotagged tweet using a retrieval model (see Section 4.2). Finally, we combine the evidence gathered from the above mentioned Top-N tweets by adopting a weighted majority voting algorithm, which we introduce in the following section.

### 3.1. Weighted majority voting

In order to combine evidence gathered from the Top-N content-based most similar geotagged tweets to a non-geotagged tweet $t_{ng}$, we adopt a weighted majority voting algorithm (Blum, 1996; Boyer & Moore, 1991; Chiang et al., 2012; Littlestone & Warmuth, 1992; Mosbah & Boucheham, 2015; Rokach, 2010) as follows. Each element of the Top-N tweets is represented as a tuple ($t_i$, $l_i$, $u_i$), where $l_i$ is the location associated with the geotagged tweet $t_i$ posted by the user $u_i$. Finally, we select the most frequent location within the Top-N set as the inferred location for the non-geotagged tweet. We can formalise it as:

$$Location(t_{ng}) = \underset{l_j \in L}{\operatorname{argmax}} \left( \sum_{i=1}^{N} W_{l_i}(\alpha, t_{ng}) * Vote(t_i^{l_i}, l_j) \right) \tag{1}$$

where $L$ is the set of unique locations ($l_j$) associated with the Top-N geotagged tweets, and $t_i^{l_i}$ is the location of the $i$th tweet in the rank. Then, a vote is given to the location $l_j$ by the tweet $t_i$ as follows:

$$Vote(t_i^{l_i}, l_j) = \begin{cases} 1 & t_i^{l_i} = l_j \\ 0 & t_i^{l_i} \neq l_j \end{cases} \tag{2}$$

The vote from tweet $t_i$ is weighted by:

$$W_{l_i}(\alpha, t_{ng}) = \alpha \cdot Credibility(u_i) + (1 - \alpha) \cdot Sim(t_i, t_{ng}) \tag{3}$$

where $\alpha \in [0, 1]$, and $Credibility(u_i)$ is the credibility of user $u_i$ that posted the tweet $t_i$ (see Section 3.2). $Sim(t_i, t_{ng})$ is the content-based similarity of the geotagged tweet ($t_i$) with the non-geotagged tweet ($t_{ng}$) given by a retrieval model (see Section 3.3). Finally, the location $l_j$ that obtains the highest number of votes is returned as the final predicted geolocation for a given non-geotagged tweet.

We decided to use a linear combination as our weighting function in order to study the effectiveness of each of the components together and separately. Therefore, when using $\alpha = 1$ only the credibility score is considered, whereas the content-based similarity is considered when $\alpha = 0$. Likewise, when $\alpha = 0.5$ both components are considered equally.

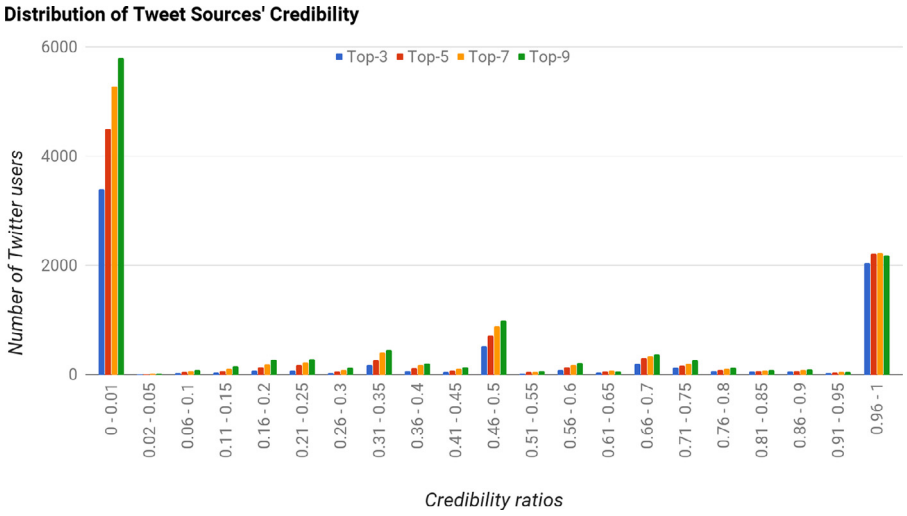


**Fig. 1.** Distribution of tweet users' credibility. The figure presents the number of Twitter users (y-axis) distributed over different values of credibility ratios (x-axis).

### 3.2. Extracting credibility from a tweet's user

We believe that some users generate content that is highly relevant to the physical location they are posting from. However, to ensure that the content is credible we compute a score based on the posting activity of the user. Finally, we utilise this score to weight the vote of a tweet in our adapted majority voting algorithm, as discussed above in Section 3.1.

To obtain the credibility score we use the training and validation sets introduced in Section 4.1, and compute it as follows. First, we obtain the Top-N most content-based similar tweets from the training set for every tweet in the validation set. Second, for each of the tweets in the Top-N ($t_{si}$) we calculate their geographical distance with respect to their corresponding validation tweet ($t_{vi}$). Finally, for each user $u_i$ we define a set *TN* that contains all the tweets $t_{si}$ posted by the user. Finally, the credibility of user $u_i$ is given by the ratio of all tweets in *TN* placed within less than 1 km distance from their corresponding $t_{vi}$ tweet in the validation set, formally defined as:

$$Credibility(u_i) = \frac{|\{t_{si} \in TN \mid distance(t_{si}, t_{vi}) \leq 1km\}|}{|TN|} \quad (4)$$

Fig. 1 shows the distribution of credibility ratios when considering different cut-off points for N across all users evaluated in the validation set for the city of Chicago (see Section 4.1). As can be observed, an important chunk of the user population exhibit a low ratio ( $\leq 0.01$). On the other hand, the rest of the population is uniformly distributed except $0.46 - 0.5$ and $0.96 - 1$, where there is a noticeably higher concentration of users. We observed similar patterns in all the cities considered in Section 4.1.

### 3.3. Similarity score and tweet geolocation

Previous research (Grabovitch-Zuyev, Kanza, Kravi, & Pat, 2007; Hong, Ahmed, Gurumurthy, Smola, & Tsioutsiouliklis, 2012b) has shown the correlation between the content of the tweets and their geographical location. This is because high similar tweets are related to the same topic/event, and therefore they are likely to be posted in the same location. Based on this assumption, we believe that the level of content-similarity with the content of the Top-N geotagged tweets is a strong indicator of the actual geolocation for a given non-geotagged tweet.

For example, given the non-geotagged tweet "*Welcome to my birthday party at 7th avenue*", and the geotagged tweet "*Amazing birthday party in a nightclub at 7th avenue*", their contents are highly related as they refer to the same event (*birthday party at 7th avenue*). Therefore, they will be associated with a high similarity score, and it is very likely that they were posted in the same geographical locations.

However, we can find some cases in which the level of similarity is not sufficient to ascertain whether any two tweets share a geographical location. For example, given the non-geotagged tweet "*Happy Birthday to my friend David*", and the geotagged tweet "*Amazing birthday party at 7th avenue*", they will be associated with a low similarity score as both tweets contain the term "*birthday*", but they are not referring to the same event. This indicates that although the topics are related to a birthday event, they may or may not be referring to the same event in the same location.

To this end, we introduced the similarity score $Sim(t_i, t_{ng})$ in Eq. (3) in Section 3.1. The contribution of the similarity component is controlled by the values of $\alpha$. In addition, the lower the value of $\alpha$ the higher the contribution of the content-based similarity score to the total weighting of each tweet vote.

### 3.4. Time complexity

In this Section, we discuss the time complexity of our proposed approach. To do that, we consider the two main components of our approach explained previously in this section: the ranking component, and the majority voting algorithm. Other elements such as the credibility ratio, the creation of the grid of squared areas, and the association of each geotagged to its corresponding area can be computed off-line, and therefore will not be considered.

The ranking task in information retrieval has been widely investigated and optimised over the time. In this work, we adopted the most common approach that implements an inverted-index that maps words to documents. This way, the number of documents to explore is reduced to only the $D_p$ documents that contain any of the words of the query (Cutting & Pedersen, 1997). This corresponds to an overall time complexity of $O(D_p)$. Nevertheless, more advanced implementations have reduced this complexity further (Perry & Willett, 1983).

For our second component, we implemented a majority voting algorithm that performs the task in linear time $O(N)$, where $N$ is the Top-N elements in the rank considered by the majority voting algorithm. This is achieved by storing each element's frequency in a hash table, so only one iteration over the input candidates is needed to count frequencies.

Overall, our proposed approach performs the geolocalisation of a non-geotagged tweet in time $O(D_p + N)$. This complexity shows the convenience of our approach for real-time applications (i.e. real-time traffic incident detection).

## 4. Experimental setup

In this section, we describe the experimental setup that supports the evaluation of our proposed approach for fine-grained geolocalisation of non-geotagged tweets.

### 4.1. Data

Previous studies have shown that geotagged and non-geotagged data have the same characteristics (Han, Cook, & Baldwin, 2014). Thus, models built from geotagged data can potentially be generalised to non-geotagged data. Moreover, as we only use geotagged data from specific cities, we assume that the city-level (or similar) of a tweet is known and focus on detecting their fine-grained location.[2] Therefore, we experimented over a ground truth sample of English geotagged tweets located in two different cities. Tweets were collected from the Twitter Public stream.[3]~Our first dataset contains 131,273 geotagged tweets posted on March 2016 in Chicago (USA). Also, we collected two different datasets with tweets posted in New York City (USA) containing 155,114 from March 2016, and 1,318,065 geotagged tweets from September 2014. As a preprocessing step, usernames and hashtags were preserved as tokens, all hyperlinks were removed from tweets, and re-tweets were preserved in the dataset. Then, for each tweet we removed punctuations, removed stop-words (Manning et al., 2008), applied Porter Stemmer (Porter, 1980) and tokenised to extract words (1-gram). Finally, using Apache Lucene,[4] we created an inverted index that maps words to tweets.

To evaluate our approach, we divided each dataset into three subsets. We used the first three weeks of tweets in our collection (i.e. the first three weeks of March and September) as a training set. We then randomly divided the last week data into validation and test sets to ensure that they have similar characteristics. Table 1 describes the distribution of tweets for the three datasets.

### 4.2. Models

In this section, we describe the baseline models, as well as the different configurations of our approach utilised in our experiments. In total, we have implemented three state-of-the-art approaches (explained in detail below) as strong baselines. Based on their general approach, we have divided the baselines models into two groups: grid-based approaches and density-based approaches.

#### 4.2.1. Grid-based baseline models

Our grid-based approaches model the space by dividing the geographical area using a grid structure. To adapt the baselines to the task of fine-grained geolocalisation, for each city mentioned in Section 4.1, we created a grid structure of squared areas with a side length of 1 km (denoted by "fine-grained grid").

**Hulden.** Firstly, we implemented the work by Hulden et al. (2015) (denoted by "Hulden"). Following authors approach, we have modelled each cell of the 1 km grid using Multinomial Naive Bayes (denoted by "Naive_Bayes") and Kullback–Leibler divergence (denoted by "Kullback_Leiber") using words as features. We also report standard Naive Bayes and Kullback–Leiber versions using kernel density estimation, denoted as "Naive_Bayes + KDE" and "Kullback_Leiber + KDE" respectively.

**Paraskevopoulos.** In this baseline, for each of the cells in the 1 km "fine-grained grid" described above, we created a document by concatenating the text of the tweets associated with that cell. We then indexed these documents (see Section 4.1). After indexing the documents, we retrieve the most content-based similar document (Top-1) for each non-geotagged tweet. As our retrieval model,

[2] The city-level location of tweets can be derived from previous works such as (Cheng et al., 2010; Kinsella et al., 2011; Schulz, Hadjakos et al., 2013).
[3] https://dev.twitter.com/streaming/public.
[4] http://lucene.apache.org/core/.

**Table 1**
Number of tweets distributed between training, validation and testing for our three datasets. We collected geotagged tweets from the Twitter Public Stream posted in two different cities: Chicago (USA) ("Chicago") and New York City (USA) ("NYC").

| | | Number of Tweets | | |
|---|---|---|---|---|
| Dataset | Collection Time | *Training* | *Validation* | *Testing* |
| Chicago | March 2016 | 111,627 | 9823 | 9823 |
| NYC_1 | March 2016 | 128,746 | 13,184 | 13,184 |
| NYC_2 | September 2014 | 1,123,125 | 97,470 | 97,470 |

we implemented TF-IDF, to follow (Paraskevopoulos & Palpanas, 2015) (denoted by "TF-IDF"). We also tried four alternative retrieval models including Language Model with Dirichlet Smoothing (denoted by "LMD"), Divergence From Randomness (denoted by "DFR"), IDF (denoted by "IDF") and BM25 (denoted by "BM25").

It is important to note that the results presented in this paper are based on an adaptation of (Paraskevopoulos & Palpanas, 2015) since we have removed stop-words (Manning et al., 2008) and applied Porter stemming (Porter, 1980), while the original work did not take these actions.[5] We also did not consider time dimension as it is out of the scope of this work, in contrast to Paraskevopoulos and Palpanas (2015) model.

Additionally, note that Paraskevopoulos and Palpanas (2015) extended Kinsella et al. (2011) approach to work at a fine-grained level by using a 1 km grid, and TF-IDF as retrieval model instead of Language Model with Dirichlet Smoothing. Therefore, our adaptation of Paraskevopoulos that use Language Models ("LMD") follows the work by Kinsella et al. (2011).

#### 4.2.2. Density-based baseline model

**Flatow.** We also implemented the work by Flatow et al. (2015) (denoted by "Flatow") that utilises Gaussian Mixture Models to assign a location to geospecific word n-grams, based on the number of tweets that contains the n-grams and its spatial density. An n-gram is geospecific if we can create an ellipse that covers a predefined maximum area ($s$) and contains at least a certain ratio of the total tweets ($t$). As we are working at a fine-grained level, we fixed $s$ to 1 $km^2$ and experimented with different values of $t$ (0.5, 0.6, 0.7, 0.8), being $t = 0.8$ the best performing one.

#### 4.2.3. WMV model

Our proposed approach explained in Section 3 (denoted by "WMV") was implemented. We used the same squared areas of the fine-grained grid defined for the baseline models. However, in WMV model, each of these defined squared areas was represented as multiple bag-of-word vectors where each vector represents a single geotagged tweet associated with that area. By doing this, we indexed each tweet as a single document for the retrieval task. All tweets were preprocessed following the same step explained in Section 4.1.

After indexing the tweets, a retrieval task was performed to obtain the Top-N content-based most similar geotagged tweets for each non-geotagged tweet. Similarly to the baselines, we investigated the same five retrieval models to maximise the performance of our approach, returning the longitude and latitude coordinates of the Top-1 tweet as the predicted location. The results indicated that using IDF gave us the best performance. This is consistent with previous research findings in microblog information retrieval (Rodriguez Perez & Jose, 2015).

Finally, we apply our weighted majority voting algorithm on top of the retrieval task. In our experimental evaluation we considered the Top-3, -5, -7 and -9 content-based most similar tweets obtained from the retrieval task, and different values of $\alpha$ for the weighting function (0.0, 0.25, 0.55, 0.75, 1.0). The final predicted location is the predefined area that obtains the majority of the votes.

### 4.3. Metrics

We reported the following metrics for evaluating the effectiveness of our approach.

- **Average error distance (km).** We compute the distance on Earth (Haversine formula (Robusto, 1957)) between the predicted location and the real coordinates of the tweet in our ground truth. As described in Section 4.2, the output of our models can be either a tweet or a squared area. When our prediction is a single tweet, we compute the distance between two coordinates; when our prediction is an area, the distance between the ground truth coordinate and the centroid of the area is calculated.
- **Accuracy.** We measure the accuracy of our model in two ways. First, we calculate whether the centroid of the predicted area lies within a radius of 1 km from the real location of a tweet (denoted by **"Accuracy@1 km"**).
  Second, we calculate whether the real location of a tweet falls within the predicted area or not (denoted by **"Accuracy@GRID"**). Note that in "WMV", "Huldens" and "Paraskevopoulos" models, this area corresponds to a squared area of

[5] We also tried an implementation without removing stop-words nor applying Porter stemming, which resulted in the lower performance and hence we did not report them.

**Table 2**
Results for "Chicago" dataset. The table presents the Average Error Distance in kilometres (AED), Accuracy at Grid (A@Grid), Accuracy at 1 km (A@1km) and Coverage for our proposed approach ("WMV") against the baselines using the Top-N (@TopN) elements in the rank. Significant differences ($p < 0.01$) with respect to the best density-based baseline ("Flatow") and the best grid-based baseline ("Paraskevopoulos_LMD") are denoted by * and † respectively.

| Chicago | | | | | |
|---|---|---|---|---|---|
| *Model* | *Config* | *AED* | *A@Grid* | *A@1km* | *Coverage* |
| Hulden | *NaiveBayes + KDE* | 7.461 | 15.87% | 30.00% | 100% |
| Hulden | *KullbackLeibler + KDE* | 7.570 | 9.621% | 25.66% | 100% |
| Hulden | *NaiveBayes* | 6.215 | 48.61% | 51.72% | 100% |
| Hulden | *KullbackLeibler* | 6.994 | 46.39% | 49.32% | 100% |
| Paraskevopoulos | *TFIDF* | 8.686 | 37.37% | 40.95% | 99.98% |
| Paraskevopoulos | *LMD* | 7.025 | 40.32% | 44.20% | 99.98% |
| Paraskevopoulos | *IDF* | 13.987 | 11.12% | 12.96% | 99.98% |
| Paraskevopoulos | *DFR* | 8.839 | 34.11% | 37.54% | 99.98% |
| Paraskevopoulos | *BM25* | 8.234 | 36.16% | 39.52% | 99.98% |
| Flatow | $t = 0.8$, $s = 1$ km$^2$ | 2.903 | 15.21% | 59.62% | 69.85% |
| WMV@Top-3 | $\alpha = 0.0$ | 2.389* † | 72.36%* † | 75.27%* † | 67.06%* † |
| WMV@Top-3 | $\alpha = 0.25$ | 2.333* † | 72.85%* † | 75.85%* † | 66.68%* † |
| WMV@Top-3 | $\alpha = 0.55$ | 2.365* † | 72.38%* † | 75.60%* † | 66.13%* † |
| WMV@Top-3 | $\alpha = 0.75$ | 2.898* † | 67.69%* † | 71.14%* † | 75.22%* † |
| WMV@Top-3 | $\alpha = 1.0$ | 3.768* † | 60.27%* † | 63.85%* † | 86.30%* † |
| WMV@Top-5 | $\alpha = 0.0$ | 1.662* † | 79.67%* † | 82.32%* † | 58.33%* † |
| WMV@Top-5 | $\alpha = 0.25$ | 1.643* † | 79.76%* † | 82.44%* † | 58.17%* † |
| WMV@Top-5 | $\alpha = 0.55$ | 1.665* † | 78.94%* † | 81.97%* † | 59.45%* † |
| WMV@Top-5 | $\alpha = 0.75$ | 1.857* † | 77.07%* † | 80.17%* † | 61.77%* † |
| WMV@Top-5 | $\alpha = 1.0$ | 3.182* † | 65.63%* † | 69.17%* † | 75.15%* † |
| WMV@Top-7 | $\alpha = 0.0$ | 1.343* † | 82.98%* † | 85.23%* † | 53.71%* † |
| WMV@Top-7 | $\alpha = 0.25$ | 1.349* † | 82.91%* † | 85.26%* † | 53.56%* † |
| WMV@Top-7 | $\alpha = 0.55$ | 1.428* † | 81.99%* † | 84.44%* † | 54.49%* † |
| WMV@Top-7 | $\alpha = 0.75$ | 1.584* † | 80.41%* † | 83.00%* † | 55.94%* † |
| WMV@Top-7 | $\alpha = 1.0$ | 2.592* † | 71.39%* † | 74.27%* † | 65.10%* † |
| WMV@Top-9 | $\alpha = 0.0$ | 1.208* † | 84.24%* † | 86.75%* † | 51.19%* † |
| WMV@Top-9 | $\alpha = 0.25$ | 1.254* † | 83.86%* † | 86.44%* † | 51.18%* † |
| WMV@Top-9 | $\alpha = 0.55$ | 1.317* † | 82.99%* † | 85.65%* † | 51.60%* † |
| WMV@Top-9 | $\alpha = 0.75$ | 1.489* † | 81.29%* † | 84.06%* † | 52.83%* † |
| WMV@Top-9 | $\alpha = 1.0$ | 2.166* † | 74.69%* † | 77.65%* † | 58.84%* † |

side length 1 km, whereas for "Flatow" this area corresponds to an ellipse.

- **Coverage.** We consider Coverage as the fraction of tweets in the test set from which our approach finds a geolocation regardless of the distance error.

## 5. Results

In this section, we present our experimental results on fine-grained geolocalisation of tweets using our adapted weighted majority voting approach compared to the baseline models. We report the average error distance, accuracy, and coverage for our approach evaluated on the "Chicago", "NYC_1" and "NYC_2" datasets. These results are presented in Table 2, Table 3 and Table 4 respectively. A paired *t*-test was conducted to assess if the difference in effectiveness between the models is statistically significant.

### 5.1. Performance

As shown in Tables 2–4, our approach ("WMV") (statistically) significantly outperforms grid-based baseline models (i.e. "Parakevopoulos" and "Hulden") in terms of accuracy and error distance, regardless of the value of N in all datasets. However, this increase of accuracy and error distance is accompanied with the cost of a decrease in coverage. Additionally, our findings show that, as the number of voting candidates (i.e. Top-N) increases, our approach achieves lower error distance, higher accuracy, but lower coverage. Therefore, considering the Top-3 tweets resulted in the best trade-off regarding error distance, accuracy and coverage.

In addition, our results suggest that the aggregation of tweets to represent an area as a single vector leads to a decrease in accuracy when working at a fine-grained level of granularity. Consistently with this observation, our density-based baseline ("Flatow") behaves better than other grid-based approaches ("Hulden" and "Parakevopoulos"). Nevertheless, compared to "Flatow" our approach ("WMV") still performs better when using Top-3 and $\alpha = 0.75$. Moreover, as the values of $\alpha$ are close to 0.0, our model outperforms "Flatow" in terms of average error distance.

Finally, despite their geographical and cultural differences, our approach performs similarly across the two cities investigated in our experiments. Such similarity in performance suggests that our approach can be generalised and adapted to different cities. Also, we evaluated our approach in two datasets with different sample sizes from the same city, "NYC_1" in Table 2, and "NYC_2" in

**Table 3**
Results for "NYC_1" dataset. The table presents the Average Error Distance in kilometres (AED), Accuracy at Grid (A@Grid), Accuracy at 1 km (A@1km) and Coverage for our proposed approach ("WMV") against the baselines using the Top-N (@TopN) elements in the rank. Significant differences ($p < 0.01$) with respect to the best density-based baseline ("Flatow") and the best grid-based baseline ("Paraskevopoulos_LMD") are denoted by * and † respectively.

| NYC_1 | | | | | |
|---|---|---|---|---|---|
| *Model* | *Config* | *AED* | *A@Grid* | *A@1km* | *Coverage* |
| Hulden | *NaiveBayes + KDE* | 6.648 | 9.72% | 23.57% | 100% |
| Hulden | *KullbackLeibler + KDE* | 6.568 | 8.64% | 19.92% | 100% |
| Hulden | *NaiveBayes* | 6.309 | 39.35% | 43.78% | 100% |
| Hulden | *KullbackLeibler* | 7.129 | 37.40% | 41.76% | 100% |
| Paraskevopoulos | *TFIDF* | 7.505 | 34.79% | 38.39% | 99.98% |
| Paraskevopoulos | *LMD* | 7.169 | 34.10% | 37.29% | 99.98% |
| Paraskevopoulos | *IDF* | 12.755 | 10.21% | 12.78% | 99.98% |
| Paraskevopoulos | *DFR* | 7.609 | 32.93% | 36.28% | 99.98% |
| Paraskevopoulos | *BM25* | 7.460 | 34.88% | 38.25% | 99.98% |
| Flatow | $t = 0.8$, $s = 1$ km$^2$ | 2.903 | 16.29% | 65.52% | 72.46% |
| WMV@Top-3 | $\alpha = 0.0$ | 2.491* † | 67.35%* † | 71.83%* † | 58.65%* † |
| WMV@Top-3 | $\alpha = 0.25$ | 2.427* † | 67.95%* † | 72.52%* † | 58.44%* † |
| WMV@Top-3 | $\alpha = 0.55$ | 2.460* † | 67.18%* † | 72.50%* † | 60.84%* † |
| WMV@Top-3 | $\alpha = 0.75$ | 3.179* † | 60.35%* † | 66.19%* † | 71.05%* † |
| WMV@Top-3 | $\alpha = 1.0$ | 3.939* † | 53.63%* † | 59.68%* † | 82.36%* † |
| WMV@Top-5 | $\alpha = 0.0$ | 1.784* † | 75.37%* † | 79.48%* † | 49.02%* † |
| WMV@Top-5 | $\alpha = 0.25$ | 1.727* † | 75.37%* † | 80.03%* † | 49.04%* † |
| WMV@Top-5 | $\alpha = 0.55$ | 1.768* † | 75.01%* † | 79.65%* † | 51.08%* † |
| WMV@Top-5 | $\alpha = 0.75$ | 2.021* † | 72.24%* † | 77.41%* † | 54.21%* † |
| WMV@Top-5 | $\alpha = 1.0$ | 3.824* † | 56.04%* † | 61.58%* † | 74.90%* † |
| WMV@Top-7 | $\alpha = 0.0$ | 1.414* † | 79.15%* † | 83.12%* † | 44.86%* † |
| WMV@Top-7 | $\alpha = 0.25$ | 1.428* † | 79.06%* † | 83.02%* † | 44.92%* † |
| WMV@Top-7 | $\alpha = 0.55$ | 1.482* † | 78.26%* † | 82.50%* † | 46.31%* † |
| WMV@Top-7 | $\alpha = 0.75$ | 1.628* † | 76.42%* † | 81.28%* † | 48.32%* † |
| WMV@Top-7 | $\alpha = 1.0$ | 3.286* † | 60.96%* † | 66.64%* † | 64.74%* † |
| WMV@Top-9 | $\alpha = 0.0$ | 1.237* † | 81.06%* † | 84.98%* † | 42.01%* † |
| WMV@Top-9 | $\alpha = 0.25$ | 1.250* † | 80.94%* † | 85.00%* † | 41.99%* † |
| WMV@Top-9 | $\alpha = 0.55$ | 1.329* † | 79.78%* † | 84.01%* † | 43.02%* † |
| WMV@Top-9 | $\alpha = 0.75$ | 1.509* † | 78.17%* † | 82.62%* † | 44.71%* † |
| WMV@Top-9 | $\alpha = 1.0$ | 2.730* † | 66.17%* † | 71.61%* † | 55.86%* † |

Table 4. As a result, we observed similar behaviour in performance when our approach is applied in a bigger dataset ("NYC_2"), but with a cost of a decrease in performance.

### 5.2. Effects of the similarity score

As introduced in Section 3.3, we believe that the similarity between the contents of a given non-geotagged tweet, and the Top-N content-related geotagged tweets, can be indicative of the geolocation for non-geotagged tweets The effects of Eq. (3) can be observed in Tables 2–4. As the values of alpha decrease, our approach achieves higher accuracy, and reduce the average error distance. This pattern can be observed for any of the investigated values of N for the Top-N tweets in the rank. This demonstrates the validity of our assumption that similar tweets are likely to be posted in the same geographical area.

## 6. Applicability on real-time traffic incident detection

To show the usefulness of our approach in a practical scenario, we integrated our fine-grained geolocalisation approach into a real-time traffic incident detection task. The objective of the traffic incident detection task is to identify traffic-related content in the Twitter stream, and provide this information in real-time to end-users and transportation managers.

The advantages of using Twitter data for detecting traffic incidents are many-fold. First, Twitter data provides first-hand detailed descriptions of the incidents, which are reported by human beings. Second, people move everywhere and provide a wider coverage of the transportation network than traditional traffic detection systems, which use sensors that are placed in the main roads of the transportation network. Therefore, such characteristics of Twitter data can be complementary to expensive traditional systems, which are unable to provide first-hand detailed information of the incident. For these reasons, many research efforts have tackled the problem of filtering the Twitter stream to obtain quality traffic incident-related information (D'Andrea, Ducange, Lazzerini, & Marcelloni, 2015; Gu, Qian, & Chen, 2016; Mai & Hranac, 2013; Qian, 2016; Schulz, Ristoski, & Paulheim, 2013). In addition, knowing the precise location of the incident is crucial for reliably perform such task; however, previous works are limited by the small number of geotagged data available in the Twitter stream.

In this study, we aim to investigate to what extent fine-grained geolocalisation can provide reliable geotagged data to address this

**Table 4**
Results for "NYC_2" dataset. The table presents the Average Error Distance in kilometres (AED), Accuracy at Grid (A@Grid), Accuracy at 1 km (A@1km) and Coverage for our proposed approach ("WMV") against the baselines using the Top-N (@TopN) elements in the rank. Significant differences ($p < 0.01$) with respect to the best density-based baseline ("Flatow") and the best grid-based baseline ("Paraskevopoulos_LMD") are denoted by * and † respectively.

| NYC_2 | | | | | |
|---|---|---|---|---|---|
| *Model* | *Config* | *AED* | *A@Grid* | *A@1km* | *Coverage* |
| Hulden | *NaiveBayes + KDE* | 12.671 | 4.74% | 11.25% | 100% |
| Hulden | *KullbackLeibler + KDE* | 13.164 | 5.15% | 12.16% | 100% |
| Hulden | *NaiveBayes* | 14.335 | 17.05% | 19.91% | 100% |
| Hulden | *KullbackLeibler* | 15.521 | 17.01% | 19.68% | 100% |
| Paraskevopoulos | *TFIDF* | 13.981 | 17.11% | 19.54% | 84.59% |
| Paraskevopoulos | *LMD* | 13.831 | 17.64% | 19.96% | 99.98% |
| Paraskevopoulos | *IDF* | 17.248 | 6.20% | 7.58% | 99.98% |
| Paraskevopoulos | *DFR* | 14.266 | 17.92% | 19.40% | 99.98% |
| Paraskevopoulos | *BM25* | 13.576 | 17.48% | 19.82% | 99.98% |
| Flatow | $t = 0.8$, $s = 1$ km$^2$ | 7.356 | 11.42% | 36.06% | 44.48% |
| WMV@Top-3 | $\alpha = 0.0$ | 6.134* † | 48.20%* † | 52.10%* † | 31.69%* † |
| WMV@Top-3 | $\alpha = 0.25$ | 6.175* † | 47.96%* † | 52.02%* † | 32.27%* † |
| WMV@Top-3 | $\alpha = 0.55$ | 6.148* † | 47.66%* † | 52.30%* † | 33.74%* † |
| WMV@Top-3 | $\alpha = 0.75$ | 7.243* † | 40.82%* † | 45.55%* † | 42.08%* † |
| WMV@Top-3 | $\alpha = 1.0$ | 11.53* † | 22.94%* † | 26.31%* † | 81.73%* † |
| WMV@Top-5 | $\alpha = 0.0$ | 4.867* † | 54.89%* † | 58.82%* † | 24.84%* † |
| WMV@Top-5 | $\alpha = 0.25$ | 4.921* † | 54.34%* † | 58.42%* † | 25.76%* † |
| WMV@Top-5 | $\alpha = 0.55$ | 4.94* † | 53.99%* † | 58.30%* † | 26.63%* † |
| WMV@Top-5 | $\alpha = 0.75$ | 5.444* † | 50.31%* † | 54.82%* † | 29.74%* † |
| WMV@Top-5 | $\alpha = 1.0$ | 11.921* † | 20.78%* † | 23.57%* † | 81.45%* † |
| WMV@Top-7 | $\alpha = 0.0$ | 4.188* † | 59.00%* † | 63.24%* † | 21.04%* † |
| WMV@Top-7 | $\alpha = 0.25$ | 4.240* † | 59.54%* † | 62.66%* † | 21.72%* † |
| WMV@Top-7 | $\alpha = 0.55$ | 4.257* † | 57.88%* † | 62.28%* † | 22.43%* † |
| WMV@Top-7 | $\alpha = 0.75$ | 4.682* † | 54.66%* † | 59.23%* † | 24.50%* † |
| WMV@Top-7 | $\alpha = 1.0$ | 11.747* † | 20.89%* † | 23.66%* † | 73.34%* † |
| WMV@Top-9 | $\alpha = 0.0$ | 3.988* † | 60.14%* † | 64.37%* † | 19.11%* † |
| WMV@Top-9 | $\alpha = 0.25$ | 4.018* † | 59.54%* † | 63.97%* † | 19.75%* † |
| WMV@Top-9 | $\alpha = 0.55$ | 4.022* † | 58.73%* † | 63.33%* † | 20.38%* † |
| WMV@Top-9 | $\alpha = 0.75$ | 4.356* † | 55.99%* † | 59.23%* † | 24.50%* † |
| WMV@Top-9 | $\alpha = 1.0$ | 11.475* † | 21.62%* † | 24.36%* † | 65.25%* † |

**Table 5**
Average distance to traffic incidents in kilometres of geotagged incident-related tweets and geolocalised incident-related tweets. A *t*-test was conducted to assess that differences are statistically significant, and are denoted by † ($p < 0.01$).

| | *AVG Distance* |
|---|---|
| Geotagged incident-related | 3.300 km† |
| Geolocalised incident-related | 2.911 km† |

task. To this end, we performed an exhaustive analysis to determine differences in the spatial proximity between incident-related tweets – geotagged and geolocalised – with respect to traffic incidents reported in the city of New York, USA.

### 6.1. Extracting incident-related tweets

In order to perform the real-time incident detection task, we obtained a ground truth dataset of human labeled incident-related tweets[6] generated by Schulz, Guckelsberger, and Janssen (2017); Schulz, Ristoski et al. (2013). The dataset contains 1858 tweets posted from January 2014 to March 2014 in a 15 km radius around the city centre of New York. The tweets are categorised as "crash" and "non-incident".

To obtain incident-related tweets, we first trained a classifier on this ground truth dataset to determine whether a tweet is incident-related or not (following previous works (Schulz, Ristoski et al., 2013)). The final evaluation showed that our classifier was able to correctly identify incident-related tweets with an accuracy of 90.45%, which is consistent with similar works (D'Andrea et al., 2015; Gu et al., 2016; Schulz, Ristoski et al., 2013). We utilised our classifier to filter 1.3 million geotagged tweets posted in New

[6] http://www.doc.gold.ac.uk/~cguck001/IncidentTweets/.

York between October 2014 and December 2014. As a result, we obtained a total of 597 tweets categorised as incident-related (labelled as "crash") for our analysis.

### 6.2. Fine-grained geolocalisation of incident-related tweets

We estimated the geolocation of the incident-related tweets using our fine-grained geolocalisation approach introduced in Section 3. This approach was trained and evaluated on the "NYC_2" dataset described in Section 4.1.

To configure our geolocalisation method, we selected the parameters that give us a reasonable trade-off between accuracy and coverage. Thus, we configured our method to utilise the Top-3 content-based most similar geotagged tweets, and $\alpha = 0.55$. As can be observed in Table 4, this configuration achieved an average error distance of 6.248 km, 52.30% accuracy, and 33.74% coverage.

Finally, we ran our fine-grained geolocalisation method on the 597 incident-related tweets obtained in Section 6.1. As a result, our fine-grained geolocalisation approach was able to estimate the location of 42.84% of the incident-related tweets with an average error of 8.518 km, where 24.56% of them were placed within 1 km distance from their real location.

### 6.3. Effectiveness analysis

To assess the effectiveness of the geolocalised incident-related tweets, we investigated whether the spatial distribution pattern of the tweets is clustered around the actual locations of the incidents. We perform this analysis by using the original locations (geotagged), and the locations estimated by our method (geolocalised) in Section 6.2.

***Traffic incidents data***. Both geotagged and geolocalised tweets were compared against a ground truth dataset containing traffic incidents reported by the New York Police Department[7].~Specifically, we extracted motor vehicle collisions which occurred on any road in New York during the period of study (October-December 2014). According to previous research (Hakkert & Mahalel, 1978; Thomas, 1996), the majority of traffic incidents occur in road intersections. Therefore, to obtain a more realistic representation of the locations of traffic incidents, we collected the location of road intersections provided by the New York Open Data Portal.[8]

***Comparing against traffic incidents***. The bivariate K function (Bivand & Gebhardt, 2000; Gavin, 2010; Lotwick & Silverman, 1982; Rowlingson & Diggle, 1993) was used to explore how incident-related tweets are distributed around road intersections in comparison with traffic incidents. Generally, the variance stabilized $K$-function ($L$-function) is used in data analysis. This function is used to examine repulsive, attractive, and random relationships between two point sets within a distance window ($d$). The bivariate $L$-function for two point sets ($A$ and $B$) can be written as:

$$L_{AB}(d) = \sqrt{\frac{K_{AB}(d)}{\pi}} - d \tag{5}$$

and

$$K_{AB}(d) = \frac{S}{N_A N_B} \sum_{i=1}^{N_A} \sum_{i=1}^{N_B} w(A_i, B_j) I(|A_i - B_j| < d) \tag{6}$$

where S is area of study area, $N_A$ and $N_B$ are the number of points in set $A$ and $B$. $A_i$ and $B_j$ are locations of points, and $|A_i - B_j|$ is the distance between $A_i$ and $B_j$. $I$ is the identity function, and $w(A_i, B_j)$ is an edge correction, set to 2 if $|A_i - B_j|$ is greater than the distance of $A_i$ to the nearest "edge" of the record, otherwise it is set to 1.

### 6.4. Results

The bivariate $L$-function, introduced in Section 6.3, was calculated with a Monte Carlo envelope constructed at the 95% confidence level with 999 Monte Carlo simulations. Specifically, bivariate $L$-statistics were computed for three pairwise point sets: 1) traffic incidents versus road intersections, 2) geotagged incident-related tweets versus road intersections, and 3) geolocalised incident-related tweets versus road intersections. Figs. 2 and 3 shows bivariate $L$ statistics for the three pairwise point sets respectively at different distances ($d$).

As can be observed in Fig. 2, traffic incidents are significantly clustered around road intersections ($L(d)$ falls above the simulation envelope) when distance is below 3000 m (see Fig. 2). This indicates that traffic incidents tend to take place around road intersections, and supports the assumption that road intersections represent realistic locations of traffic incidents (Hakkert & Mahalel, 1978).

On the other hand, as can be observed in Fig. 3 geotagged incident-related tweets are randomly distributed around road intersections ($L(d)$ falls below or overlaps the simulation envelope) at almost all distances (see left Fig. 3). In contrast, geolocalised incident-related tweets are clustered around road intersections ($L(d)$ falls above the simulation envelope) when distance is below

[7] https://data.cityofnewyork.us/Public-Safety/NYPD-Motor-Vehicle-Collisions/h9gi-nx95.
[8] https://opendata.cityofnewyork.us/.

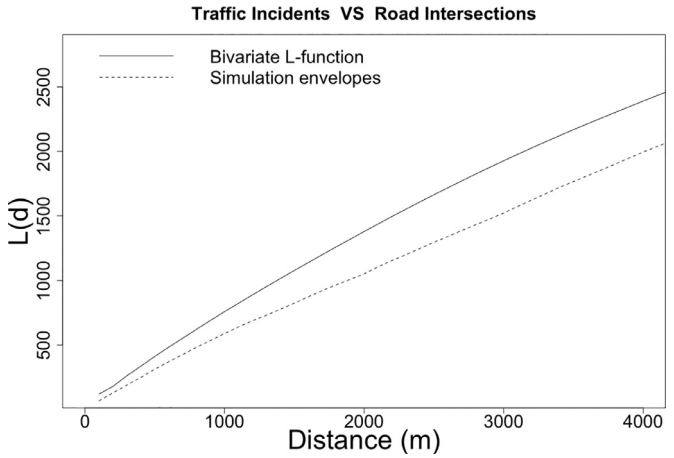


**Fig. 2.** Spatial relationship between traffic incidents and road intersections. The y-axis contains the values of the Bivariate L-function ($L(d)$) and different distances (x-axis). L-function values falling above the simulation envelope (doted line) mean attractive relationship.

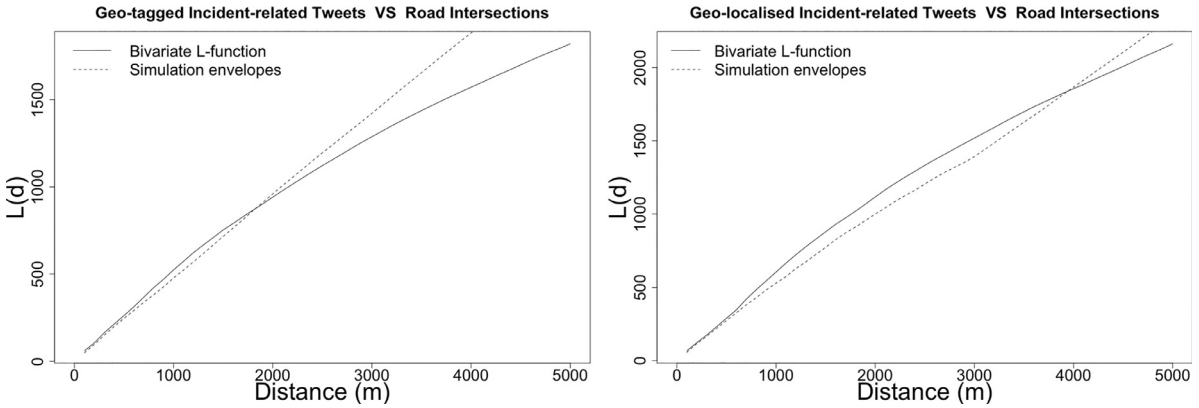


**Fig. 3.** Spatial relationship between: geotagged incident-related tweets and road intersections (left Figure), and geolocalised incident-related tweets and road intersections (right Figure). The y-axis contains the values of the Bivariate L-function ($L(d)$) and different distances (x-axis). L-function values falling above the simulation envelope (doted line) mean attractive relationship.

3000 m (see right Fig. 3).

These results show that geotagged incident-related tweets are randomly distributed around realistic locations of traffic incidents. In contrast, geolocalised incident-related tweets are clustered around the realistic locations of traffic incidents. A possible explanation to this is that some Twitter users post their incident-related tweets after they have left the actual place of the incident. In this cases, the location of the incident-related tweets could not represent the actual location of incident. This is consistent with previous findings (Mai & Hranac, 2013).

In addition, we assigned each incident-related tweet – geotagged and geolocalised – to the most likely traffic incident according to spatio-temporal proximity criteria. To do this, each traffic incident is linked to all the incident-related tweets that are posted within 30 minutes before and after the incident time. Then, the distances in kilometres between each traffic incident and the linked tweets are computed. As a result of this, geotagged incident-related tweets are located at an average distance of 3.300 km from the traffic incidents; while geolocated incident-related tweets are located at an average distance of 2.911 km. After conducting a *t*-test, the *p*-value ($p < 0.01$) showed that geolocated incident-related tweets are statistically significant closer to traffic incidents compared to geotagged incident-related tweets.

These results show the applicability of our approach, and how geolocalisation can be used for performing real-time traffic incident detection. Furthermore, we observed a mismatch between the coordinates reported by the geotagged tweets and the actual locations of the incidents. This is because some Twitter users may not post in the place of the event, but they mention the actual location in the text. We showed that our fine-grained geolocalisation approach could reduce this issue, and provide precisely located tweets that are highly correlated with the real locations of traffic incidents.

## 7. Conclusions

In this work, we proposed an approach for fine-grained geolocalisation of tweets by adopting a weighted majority voting

algorithm. The weight of each tweet vote is obtained by calculating the credibility of its user.

Inspired by grid-based approaches (Hulden et al., 2015; Kinsella et al., 2011; Paraskevopoulos & Palpanas, 2015), we worked at a fine-grained level by dividing a city into a set of predefined geographical areas of size 1 km. However, in contrast to these works, we did not concatenate the text of tweets into a document to create a single bag-of-word vector to represent a predefined area. Instead, our approach treats each tweet individually as a single document and represents each area as multiple bag-of-word vectors.

To demonstrate the effectiveness of our approach, we conducted an experiment on three datasets of English geotagged tweets collected during March 2016 from two different cities, Chicago and New York, with 131,273 and 155,114 tweets respectively. Moreover, we collected a bigger dataset from New York with 1.3 million tweets posted in October 2014. We implemented grid-based as well as density-based state-of-the-art approaches as baselines to compare against our model.

Our experimental results show that our weighted majority voting approach (statistically) significantly outperforms the baselines in terms of accuracy and error distance, in both cities, across all the investigated values of N for the Top-N tweets, with the cost of decrease in coverage in the two cities of study. We also observed that, as the number of voting candidates (i.e. Top-N) increases, our approach achieves lower error distance and higher accuracy, but lower coverage. Our findings suggest that the aggregation of tweets to represent an area as a single vector leads to a decrease in accuracy when working at a fine-grained level of granularity. This behaviour is observed across both datasets and suggests that our approach can be generalised and adapted to different cities.

We also integrated our fine-grained geolocalisation approach into a real-time incident detection task to demonstrate its applicability. We compared geotagged and geolocalised incident-related tweets with real locations of traffic incidents. We found that users may not post incident-related content at the real locations of the incidents. We then demonstrated that our geolocalisation method could overcome this issue, and map precisely incident-related tweets at the real locations of the incidents.

This shows the power of our proposed approach in predicting geolocation of tweets, and can substantially expand the sample of geotagged data at a fine-grained level (i.e. street level or neighbourhood level), helping to a wide range of applications, including real-time event detection, topic detection and disaster and emergency analysis.

## 8. Future directions

This work has opened several interesting research questions to be investigated in the future. The first research direction is to investigate the effect of temporal aspect of tweets in our model. It is known that time is an important feature to take into account to improve geolocalisation (Dredze, Osborne, & Kambadur, 2016). Currently, our model does not take temporal characteristics into account. We, therefore, aim to incorporate such characteristics into our approach as future work.

The second research direction is to investigate the effect of data sparsity and in some cases, the cold-start problem in our model. In this work, we have shown that some users can provide more valuable information about a location than others by computing a credibility score based on their past activity. However, currently, we do not take into account the user cold-start problem; i.e. users who have posted only one tweet or few tweets. In future work, we aim to address this problem by varying the amount of data available for training, and conducting a comprehensive analysis of the minimum number of tweets per user needed to effectively calculate its credibility. In addition, in future work, we aim to investigate how the stability of our model is affected by varying the size of training data.

The third research direction aims to investigate the drawbacks of using grids in our approach. The strategy of dividing the geographical space into fixed-size cells suffers from data sparsity problem since some cells may not have sufficient data points, and thus be under-represented.

The fourth research direction is to perform a qualitative study on the difference between tweets that tend to give high weighted votes, and tweets that tend to give low weighted votes. As the vote of a tweet is given by the level of credibility computed for the user that posted the tweet, a qualitative study could reveal differences in the contents (e.g. words, location names or entities) and the context (e.g. type of event) in which high credible users are involved when posted their tweets.

The fifth research direction involves investigating the decrease of coverage observed in our results (see Section 5). This can be caused because our weighted majority voting approach does not return any prediction is there is no enough evidence (i.e. there is no a majority location). By analysing such cases, we can gain insights that help us to improve the performance of the model by increasing coverage while maintaining high accuracy.

The final research direction is to investigate the effect of location name disambiguation in our model. For example, given the word “7th avenue”, it may refer to the 7th avenue in New York or the 7th avenue in Chicago. This issue can be affecting the accuracy of our model. Therefore, it merits further investigation to evaluate the effect that disambiguation cases have on effectiveness, and incorporate methodologies into our model to alleviate this problem.

## Acknowledgement

The research leading to these results has received funding from the European Research Council under the European Union’s Seventh Framework Programme (FP7/2007-2013)/ ERC grant agreement no. 632075.

## Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.ipm.2018.03.011.

## References


Ao, J., Zhang, P., & Cao, Y. (2014). Estimating the locations of emergency events from twitter streams. *Procedia Computer Science, 31*, 731–739.

Atefeh, F., & Khreich, W. (2015). A survey of techniques for event detection in twitter. *Computational Intelligence, 31*(1), 132–164.

Baucom, E., Sanjari, A., Liu, X., & Chen, M. (2013). *Mirroring the real world in social media: twitter, geolocation, and sentiment analysis. Proceedings of the 2013 international workshop on mining unstructured big data using natural language processing*. ACM61–68.

Bivand, R., & Gebhardt, A. (2000). Implementing functions for spatial statistical analysis using the language. *Journal of Geographical Systems, 2*(3), 307–317.

Blum, A. (1996). *On-line algorithms in machine learning. In proceedings of the workshop on-line algorithms, dagstuhl*. Springer306–325.

Boyer, R. S., & Moore, J. S. (1991). Mjrty—a fast majority vote algorithm. In R. S. Boyer (Ed.). [*Automated Reasoning: Essays in Honor of Woody Bledsoe*]. Dordrecht: Springer Netherlands. http://dx.doi.org/10.1007/978-94-011-3488-0_5.

Castillo, C., Mendoza, M., & Poblete, B. (2011). *Information credibility on twitter. Proceedings of the 20th international conference on world wide webWWW '11*New York, NY, USA: ACM675–684. http://dx.doi.org/10.1145/1963405.1963500.

Chang, H.-w., Lee, D., Eltaher, M., & Lee, J. (2012). *@phillies tweeting from philly? Predicting twitter user locations with spatial word usage. Proceedings of the 2012 international conference on advances in social networks analysis and mining (asonam 2012)ASONAM '12*Washington, DC, USA: IEEE Computer Society111–118. http://dx.doi.org/10.1109/ASONAM.2012.29.

Cheng, Z., Caverlee, J., & Lee, K. (2010). *You are where you tweet: a content-based approach to geo-locating twitter users. Proceedings of the 19th ACM international conference on information and knowledge management*. ACM759–768.

Chiang, T.-H., Lo, H.-Y., & Lin, S.-D. (2012). *A ranking-based knn approach for multi-label classification. Asian conference on machine learning*81–96.

Crooks, A., Croitoru, A., Stefanidis, A., & Radzikowski, J. (2013). #earthquake: Twitter as a distributed sensor system. *Transactions in GIS, 17*(1), 124–147. http://dx.doi.org/10.1111/j.1467-9671.2012.01359.x.

Cutting, D. R., & Pedersen, J. O. (1997). *Space optimizations for total ranking. Computer-assisted information searching on internetRIAO '97*Paris, France, France: Le Centre De Hautes Etudes Internationales D'informatique Documentaire401–412.

D'Andrea, E., Ducange, P., Lazzerini, B., & Marcelloni, F. (2015). Real-time detection of traffic from twitter stream analysis. *Intelligent Transportation Systems, IEEE Transactions on, 16*(4), 2269–2283.

Dredze, M., Osborne, M., & Kambadur, P. (2016). *Geolocation for twitter: Timing matters. Hlt-naacl*1064–1069.

Eisenstein, J., O'Connor, B., Smith, N. A., & Xing, E. P. (2010). *A latent variable model for geographic lexical variation. Proceedings of the 2010 conference on empirical methods in natural language processing*. Association for Computational Linguistics1277–1287.

Flatow, D., Naaman, M., Xie, K. E., Volkovich, Y., & Kanza, Y. (2015). *On the accuracy of hyper-local geotagging of social media content. Proceedings of the eighth ACM international conference on web search and data mining*. ACM127–136.

Frias-Martinez, V., Soto, V., Hohwald, H., & Frias-Martinez, E. (2012). *Characterizing urban landscapes using geolocated tweets. Proceedings of the 2012 ASE/IEEE international conference on social computing and 2012 ASE/IEEE international conference on privacy, security, risk and trustSOCIALCOM-PASSAT '12*Washington, DC, USA: IEEE Computer Society239–248. http://dx.doi.org/10.1109/SocialCom-PASSAT.2012.19.

Gavin, D. G. (2010). K1d: Multivariate ripley's k-function for one-dimensional data. *University of Oregon, 80*.

Gonzalez Paule, J. D., Moshfeghi, Y., Jose, J. M., & Thakuriah, P. V. (2017). *On fine-grained geolocalisation of tweets. Proceedings of the ACM sigir international conference on theory of information retrievalICTIR '17*New York, NY, USA: ACM313–316. http://dx.doi.org/10.1145/3121050.3121104.

Grabovitch-Zuyev, I., Kanza, Y., Kravi, E., & Pat, B. (2007). *On the correlation between textual content and geospatial locations in microblogs. Proceedings of workshop on managing and mining enriched geo-spatial dataGeoRich'14*New York, NY, USA: ACM3:1–3:6. http://dx.doi.org/10.1145/2619112.2619115.

Graham, M., Hale, S. A., & Gaffney, D. (2014). Where in the world are you? geolocation and language identification in twitter. *The Professional Geographer, 66*(4), 568–578.

Gu, Y., Qian, Z. S., & Chen, F. (2016). From twitter to detector: Real-time traffic incident detection using social media data. *Transportation research part C: emerging technologies, 67*, 321–342.

Hakkert, A. S., & Mahalel, D. (1978). Estimating the number of accidents at intersections from a knowledge of the traffic flows on the approaches. *Accident Analysis & Prevention, 10*(1), 69–79.

Han, B., & Cook, P. (2013). *A stacking-based approach to twitter user geolocation prediction. In proceedings of the 51st annual meeting of the association for computational linguistics (acl 2013): System demonstrations*7–12.

Han, B., Cook, P., & Baldwin, T. (2014). Text-based twitter user geolocation prediction. *Journal of Artificial Intelligence Research, 49*, 451–500.

Hong, L., Ahmed, A., Gurumurthy, S., Smola, A. J., & Tsioutsiouliklis, K. (Ahmed, Gurumurthy, Smola, Tsioutsiouliklis, 2012a). *Discovering geographical topics in the twitter stream. Proceedings of the 21st international conference on world wide web*. ACM769–778.

Hong, L., Ahmed, A., Gurumurthy, S., Smola, A. J., & Tsioutsiouliklis, K. (Ahmed, Gurumurthy, Smola, Tsioutsiouliklis, 2012b). *Discovering geographical topics in the twitter stream. Proceedings of the 21st international conference on world wide webWWW '12*New York, NY, USA: ACM769–778. http://dx.doi.org/10.1145/2187836.2187940.

Hulden, M., Silfverberg, M., & Francom, J. (2015). Kernel density estimation for text-based geolocation. https://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/10034.

Imran, M., Castillo, C., Diaz, F., & Vieweg, S. (2015). Processing social media messages in mass emergency: A survey. *ACM Computing Surveys, 47*(4), 67:1–67:38. http://dx.doi.org/10.1145/2771588.

Kinsella, S., Murdock, V., & O'Hare, N. (2011). *I'm eating a sandwich in glasgow: modeling locations with tweets. Proceedings of the 3rd international workshop on search and mining user-generated contents*. ACM61–68.

Li, Y., Gao, J., Meng, C., Li, Q., Su, L., Zhao, B., et al. (2016). A survey on truth discovery. *SIGKDD Explorations Newsletter, 17*(2), 1–16. http://dx.doi.org/10.1145/2897350.2897352.

Littlestone, N., & Warmuth, M. K. (1992). The weighted majority algorithm.

Lotwick, H., & Silverman, B. (1982). Methods for analysing spatial processes of several types of points. *Journal of the Royal Statistical Society. Series B (Methodological)*, 406–413.

Mai, E., & Hranac, R. (2013). *Twitter interactions as a data source for transportation incidents. Proceedings of transportation research board 92nd ann. meeting* 13–1636

Manning, C. D., Raghavan, P., Schütze, H., et al. (2008). *Introduction to information retrieval. 1*. Cambridge university press Cambridge.

Marshall, J., Syed, M., & Wang, D. (2016). *Hardness-aware truth discovery in social sensing applications. Distributed computing in sensor systems (dcoss), 2016 international conference on*. IEEE143–152.

McCreadie, R., Macdonald, C., & Ounis, I. (2016). *Eaims: Emergency analysis identification and management system. Proceedings of the 39th international ACM sigir conference on research and development in information retrieval*. ACM1101–1104.

Mosbah, M., & Boucheham, B. (2015). *Majority voting re-ranking algorithm for content based-image retrieval. Research conference on metadata and semantics research*. Springer121–131.

Paraskevopoulos, P., & Palpanas, T. (2015). *Fine-grained geolocalisation of non-geotagged tweets. Proceedings of the 2015 IEEE/ACM international conference on advances in social networks analysis and mining 2015ASONAM '15*New York, NY, USA: ACM105–112. http://dx.doi.org/10.1145/2808797.2808869.

Perry, S. A., & Willett, P. (1983). A review of the use of inverted files for best match searching in information retrieval systems. *Journal of Information Science, 6*(2–3), 59–66. http://dx.doi.org/10.1177/016555158300600204.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program, 14*(3), 130–137.

Qian, Z. S. (2016). *Real-time incident detection using social media dataTechnical Report*.

Robusto, C. C. (1957). The cosine-haversine formula. *The American Mathematical Monthly, 64*(1), 38–40.

Rodriguez Perez, J. A., & Jose, J. M. (2015). *On microblog dimensionality and informativeness: Exploiting microblogs' structure and dimensions for ad-hoc retrieval. Proceedings of the 2015 international conference on the theory of information retrievalICTIR '15*New York, NY, USA: ACM211–220. http://dx.doi.org/10.1145/

2808194.2809466.
Rokach, L. (2010). *Pattern classification using ensemble methods. 75*. World Scientific.
Roller, S., Speriosu, M., Rallapalli, S., Wing, B., & Baldridge, J. (2012). *Supervised text-based geolocation using language models on an adaptive grid. Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*. Association for Computational Linguistics1500–1510.
Rowlingson, B. S., & Diggle, P. J. (1993). Splancs: Spatial point pattern analysis code in s-plus. *Computers & Geosciences, 19*(5), 627–655.
Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). *Earthquake shakes twitter users: Real-time event detection by social sensors. Proceedings of the 19th international conference on world wide webWWW '10*New York, NY, USA: ACM851–860. http://dx.doi.org/10.1145/1772690.1772777.
Schulz, A., Guckelsberger, C., & Janssen, F. (2017). Semantic abstraction for generalization of tweet classification: An evaluation of incident-related tweets. *Semantic Web, 8*(3), 353–372.
Schulz, A., Hadjakos, A., Paulheim, H., Nachtwey, J., & Mühlhäuser, M. (2013). *A multi-indicator approach for geolocalization of tweets. Icwsm*.
Schulz, A., Ristoski, P., & Paulheim, H. (2013). *I see a car crash: Real-time detection of small scale incidents in microblogs. The semantic web: ESWC 2013 satellite events*. Springer22–33.
Teevan, J., Ramage, D., & Morris, M. R. (2011). *# twittersearch: a comparison of microblog search and web search. Proceedings of the fourth ACM international conference on web search and data mining*. ACM35–44.
Thomas, I. (1996). Spatial data aggregation: Exploratory analysis of road accidents. *Accident Analysis & Prevention, 28*(2), 251–264.
Walther, M., & Kaisser, M. (2013). *Geo-spatial event detection in the twitter stream. Proceedings of the 35th european conference on advances in information retrievalECIR'13*Berlin, Heidelberg: Springer-Verlag356–367. http://dx.doi.org/10.1007/978-3-642-36973-5_30.
Wang, D., Marshall, J., & Huang, C. (2016). *Theme-relevant truth discovery on twitter: An estimation theoretic approach. Icwsm*408–416.
Watanabe, K., Ochi, M., Okabe, M., & Onai, R. (2011). *Jasmine: a real-time local-event detection system based on geolocation information propagated to microblogs. Proceedings of the 20th ACM international conference on information and knowledge management*. ACM2541–2544.
Wing, B. P., & Baldridge, J. (2011). *Simple supervised document geolocation with geodesic grids. Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1*. Association for Computational Linguistics955–964.
Xia, C., Schwartz, R., Xie, K., Krebs, A., Langdon, A., Ting, J., & Naaman, M. (2014). *Citybeat: Real-time social media visualization of hyper-local city data. Proceedings of the 23rd international conference on world wide webWWW '14 Companion*New York, NY, USA: ACM167–170. http://dx.doi.org/10.1145/2567948.2577020.
Yin, X., Han, J., & Yu, P. S. (2007). *Truth discovery with multiple conflicting information providers on the web. Proceedings of the 13th ACM SIGKDD international conference on knowledge discovery and data miningKDD '07*New York, NY, USA: ACM1048–1052. http://dx.doi.org/10.1145/1281192.1281309.
Zhang, C., Zhou, G., Yuan, Q., Zhuang, H., Zheng, Y., Kaplan, L., et al. (2016). *Geoburst: Real-time local event detection in geo-tagged tweet streams.* SIGIR 2016.
Zhang, D. Y., Han, R., Wang, D., & Huang, C. (2016). *On robust truth discovery in sparse social media sensing. Big data (big data), 2016 IEEE international conference on.* IEEE1076–1081.
Zheng, Y., & Xie, X. (2011). Location-based social networks: Locations. *Computing with spatial trajectories,* 277–308.