Inferring the Geolocation of Tweets at a Fine-Grained Level

Jorge David Gonzalez Paule
School of Computing Science
University of Glasgow

Submitted in fulfilment of the requirements for the Degree of
Doctor of Philosophy (PhD)

February 2019

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other University.

This dissertation is the result of my own work, under the supervision of Professor Iadh Ouinis, Dr Craig MacDonald and Dr Yashar Moshfeghi, and includes nothing which is the outcome of work done in collaboration, except where specifically indicated in the text.

Jorge David Gonzalez Paule
February, 2019

To my family,
for their infinite love, support and encouragement.

Abstract

Recently, the use of Twitter data has become important for a wide range of real-time applications, including real-time event detection, topic detection or disaster and emergency management. These applications require to know the precise location of the tweets for their analysis. However, approximately 1% of the tweets are finely-grained geotagged, which remains insufficient for such applications. To overcome this limitation, predicting the location of non-geotagged tweets, while challenging, can increase the sample of geotagged data to support the applications mentioned above. Nevertheless, existing approaches on tweet geolocalisation are mostly focusing on the geolocation of tweets at a coarse-grained level of granularity (i.e., city or country level). Thus, geolocalising tweets at a fine-grained level (i.e., street or building level) has arisen as a newly open research problem. In this thesis, we investigate the problem of inferring the geolocation of non-geotagged tweets at a fine-grained level of granularity (i.e., at most 1 km error distance). In particular, we aim to predict the geolocation where a given a tweet was generated using its text as a source of evidence.

This thesis states that the geolocalisation of non-geotagged tweets at a fine-grained level can be achieved by exploiting the characteristics of the 1% of already available individual finely-grained geotagged tweets provided by the Twitter stream. We evaluate the state-of-the-art, derive insights on their issues and propose an evolution of techniques to achieve the geolocalisation of tweets at a fine-grained level.

First, we explore the existing approaches in the literature for tweet geolocalisation and derive insights on the problems they exhibit when adapted to work at a fine-grained level. To overcome these problems, we propose a new approach that ranks individual geotagged tweets based on their content similarity to a given non-geotagged. Our experimental results show significant improvements over previous approaches.

Next, we explore the predictability of the location of a tweet at a fine-grained level in order to reduce the average error distance of the predictions. We postulate that to obtain a fine-grained prediction a correlation between similarity and geographical distance should exist, and define the boundaries were fine-grained predictions can be achieved. To do that, we incorporate a majority voting algorithm to the ranking approach that assesses if such correlation exists by exploiting the geographical evidence encoded within the Top-N most similar geotagged tweets in the ranking. We report experimental results and demonstrate that by considering this geographical evidence, we can reduce the average error distance, but with a cost in coverage (the number of tweets for which our approach can find a fine-grained geolocation).

Furthermore, we investigate whether the quality of the ranking of the Top-N geotagged tweets affects the effectiveness of fine-grained geolocalisation, and propose a new approach to improve the ranking. To this end, we adopt a learning to rank approach that re-ranks geotagged tweets based on their geographical proximity to a given non-geotagged tweet. We test different learning to rank algorithms and propose multiple features to model fine-grained geolocalisation. Moreover, we investigate the best performing combination of features for fine-grained geolocalisation.

This thesis also demonstrates the applicability and generalisation of our fine-grained geolocalisation approaches in a practical scenario related to a traffic incident detection task. We show the effectiveness of using new geolocalised incident-related tweets in detecting the geolocation of real incidents reports, and demonstrate that we can improve the overall performance of the traffic incident detection task by enhancing the already available geotagged tweets with new tweets that were geolocalised using our approach.

The key contribution of this thesis is the development of effective approaches for geolocalising tweets at a fine-grained level. The thesis provides insights on the main challenges for achieving the fine-grained geolocalisation derived from exhaustive experiments over a ground truth of geotagged tweets gathered from two different cities. Additionally, we demonstrate its effectiveness in a traffic incident detection task by geolocalising new incident-related tweets using our fine-grained geolocalisation approaches.

Acknowledgements

My research would have been impossible without the aid and support of my supervisors, Prof. Iadh Ounis, Dr Craig MacDonald and Dr Yashar Moshfeghi. I am profoundly thankful for their patience, motivation, and immense knowledge. I could not have imagined having better advisors and mentors. Thanks also to my examiners, Dr Richard McCreadie and Prof. Mohand Boughanem for their insightful comments and corrections to improve this thesis.

Special thanks to Dr Yashar Moshfeghi for being my mentor and helping me shape my research ideas. He provided me through moral and emotional support in the hardest moments, and always had the right words to guide me through this adventure. For that, I will be forever thankful.

I would also like to acknowledge my fellow doctoral students for their cooperation and friendship. With a special mention to those with whom I have shared an office. Such as David Maxwell, Stuart Mackie, Jarana Manotumruksa, Fatma Elsafoury, Colin Wilkie, Rami Alkhawaldeh, Stewart Whiting, Fajie Yuan, James McMinn and Phil McParlane. It has been a pleasure!

Also, my sincere thanks to the persons who helped me in different ways during this endeavour, and all the extraordinary people that I had the privilege to meet during my five years in Glasgow. With special mention to Jesus Rodriguez Perez for encouraging me to start this adventure, and for his invaluable help, support and the moments we shared living together.

I dedicate this thesis to my parents, Nieves Maria Paule Rodriguez and Jorge Luis Gonzalez Garcia, to my grandparents, Maria Jose Rodriguez Darias and Francisco Paule Rodriguez, and to my great grandparents in heaven, Ramon Rodriguez Rodriguez and Nieves Darias Hernandez. Without their infinite love, support and sacrifice, I would not have been to this far. I am lucky and proud of having you.

Finally, last but not least, I will always be grateful to my whole family and loved ones for always being there for me through thick and thin. To my aunts and uncles, Ana Belen Paule Rodriguez, Eduardo Sanson-Chirinos Lecuona, Maria Jose Rodriguez Darias and Venancio Nuñez Diaz. To my cousins, Victor Ruiz Paule, Raquel Moro Parrilla, Elena Sanson-Chirinos Paule and Javier Sanson-Chirinos Paule.

Contents

III Applicability 106

List of Tables

List of Figures

Part I
Introduction and Background

Chapter 1

Introduction

1.1 Introduction

Social media services enable users to connect across geographical, political or economic borders. In particular, Twitter[1] represents the most important microblog service in the world with 336 million active users as of 2018[2]. Twitter allows users to share short messages instantaneously with the community discussing a wide range of topics. In particular, through its users' messages, Twitter provides a unique perspective of events occurring in the real world (Abbasi et al., 2012) with first-hand reports of the people that are witnessing such events. Additionally, users posting from mobile devices have the option to attach geographical information to their messages in the form of GPS coordinates (longitude and latitude). These characteristics of Twitter have gained increasing popularity within several research communities, such as Computing Science and Social Science. Researchers in such communities aim to exploit Twitter data as a new source of real-time geotagged information for a broad range of applications, including real-time event detection (Atefeh and Khreich, 2015), topic detection (Hong et al., 2012b), and disaster and emergency analysis (Ao et al., 2014; Imran et al., 2015; McCreadie et al., 2016).

As location knowledge is critical for such applications, virtually all the analysis conducted in such tasks utilise geotagged Twitter data exclusively. However, since only 1% of messages in the Twitter stream contain geographical information

[1] https://twitter.com/

[2] https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/

(Graham et al., 2014), the available sample size for analysis is quite limited. Furthermore, Twitter users who publish geographical information have been found to be not representative of the broader Twitter population (Sloan and Morgan, 2015). This limitation is particularly crucial to transportation applications, where several new approaches have emerged to study transportation patterns, travel behavior and detect traffic incidents using Twitter data (Cui et al., 2014; D'Andrea et al., 2015; Gu et al., 2016; Kosala et al., 2012; Mai and Hranac, 2013; Schulz et al., 2013b; Steiger et al., 2014). Thus geolocating (or geolocalising) new non-geotagged tweets can increase the sample of geotagged data for these applications, which can lead to an improvement in their performance.

Earlier studies on the geolocalisation of tweets have limitations in the precision of the spatial resolution achieved; they are capable of geolocalise tweets at a coarse-grained level (i.e., country or city level) (Eisenstein et al., 2010a; Han and Cook, 2013; Kinsella et al., 2011; Schulz et al., 2013a). Therefore, the accuracy of existing methods remains insufficient for a wide range of applications that require highly accurate geolocated data. In this thesis, we aim to bridge this gap and investigate whether we can infer the geolocation of tweets at a fine-grained level (i.e., street or neighbourhood level). We advance the existing state-of-the-art further by developing novel fine-grained geolocalisation approaches, such that it is possible to infer the geolocation of tweets at a reasonable fine-grained level.

In particular, in this thesis, we aim to infer the most likely geolocation for a given tweet using its text as a source of evidence. It is important to note that, by doing this, we predict the geolocation encoded within the content of the tweet, which does not necessarily correlates with the geolocation where the user generated the tweet. For instance, when an event occurs, tweets describing such events can be generated by users that are physically at the location of the occurrence, or by users that are aware of such event but are physically located at another location. This issue is not relevant for the tasks we aim to assist with the methods developed in this thesis - i.e., traffic incident detection or disaster and emergency analysis-, where the ultimate goal is to detect and geolocate events regardless of the geolocation where the user generated the tweet.

The essential argument made by this thesis is that the fine-grained geolocalisation of tweets can be achieved by exploiting the characteristics of already available individual finely-grained geotagged tweets. In particular, we exploit such relation to infer the geolocation of non-geotagged tweets based on their similarity to other geotagged tweets.

We address three main issues concerning the fine-grained geolocalisation of tweets and propose an evolution of techniques to tackle them. First, we investigate the limitations of existing tweet geolocalisation approaches when working at fine-grained levels. Mainly, these approaches follow the strategy of creating a virtual document to represent an area, which is generated by aggregating the texts of the geotagged tweets belonging to that area. We show that this strategy leads to a loss of important evidence and affects the geolocalisation at a fine-grained level. To alleviate such limitations, we use individual geotagged tweets instead of an aggregation of them, and propose a new approach for fine-grained geolocalisation based on a ranking of such individual geotagged tweets. Then, we return the location of the Top-1 geotagged tweet as the predicted location coordinates.

Second, we discuss the predictability of the location of tweets at a fine-grained level. We postulate that, in order to find a fine-grained location for a given non-geotagged tweet, we should find a correlation between its content similarity and geographical distance to other geotagged tweets. To this end, we propose a new approach that uses a majority voting algorithm to find such a correlation by employing the geographical evidence encoded within the Top-N most similar geotagged tweets to a non-geotagged tweet.

Finally, we investigate the effects of the quality of the ranking of geotagged tweets on fine-grained geolocalisation. In particular, we propose a learning to rank approach to re-rank geotagged tweets based on their geographical proximity to a given non-geotagged tweet, and propose multiple features tailored for the fine-grained geolocalisation of tweets. This approach can improve the ranking of geotagged tweets and, therefore, lead to better fine-grained geolocalisation.

Additionally, this thesis investigates the applicability and generalisation of our fine-grained tweet geolocalisation approaches in a practical application related to the detection of traffic incidents, which aims to use Twitter as a data-source for detecting traffic incidents occurring in a city. Existing approaches to the task aim to detect an incident by identifying incident-related content in the geotagged tweets. Then, the predicted location of the incident is given by the location of the incident-related geotagged tweets. We show how our fine-grained geolocalisation approaches are capable of inferring the geolocation of non-geotagged incident-related tweets and effectively predict the location of the incidents. Moreover, we show how traffic incident detection is improved by adding new geolocalised incident-related tweets to the sample of already available geotagged tweets that is commonly used by existing traffic incident detection approaches.

The remainder of this chapter presents the statement and contributions of this thesis, as well as a roadmap of its structure.

1.2 Thesis Statement

This thesis states that the geolocalisation of non-geotagged tweets at a fine-grained level[1] can be achieved by exploiting the characteristics of already available individual finely-grained geotagged tweets. We assume a relationship between content similarity and geographical distance amongst tweets that are posted within an area. Thus, if two tweets are similar to each other, then they are likely to be posted within the same location. In order to validate our statement, we formulate the following four main hypotheses that will be explored in our three main contributions chapters. The first three hypothesis relates to the fine-grained geolocalisation problem. Besides, the fourth hypothesis aims to validate the applicability and generalisation of our approaches.

- **Hypothesis 1:** By considering geotagged tweets individually we can preserve the evidence lost when adapting previous approaches at a fine-grained level, and thus we can improve the performance of fine-grained geolocalisation (Chapter 3).

- **Hypothesis 2:** The predictability of the geolocation of a tweet at a fine-grained level is given by the correlation between its content similarity and geographical distance to finely-grained geotagged tweets (Chapter 4).

- **Hypothesis 3:** By improving the ranking of geotagged tweets with respect to a given non-geotagged tweet, we can increase the number of similar and

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

geographically closer geotagged tweets, and thus we can obtain a higher number of fine-grained predictions (Chapter 5).

- **Hypothesis 4:** By geolocalising non-geotagged tweets we can obtain a more representative sample of geotagged data and, therefore, improve the effectiveness of the traffic incident detection task (Chapter 6).

1.3 Contributions

The key contributions of this thesis can be summarised as follows:

- An investigation into the performance issues of existing tweet geolocalisation approaches when applied to work at a fine-grained level.
- A novel ranking approach that alleviates state-of-the-art issues and enables fine-grained geolocalisation of tweets.
- An study into what makes the geolocation of a tweet predictable at a fine-grained level. We explore the relationship between content similarity and geographical distance to derive assumptions to improve the geolocalisation.
- A new model for fine-grained geolocalisation based on a weighted majority voting that combines the geographical evidence of the most similar geotagged tweets.
- We demonstrate the effectiveness of the proposed geolocalisation approach in the traffic incident detection task. We expanded the sample of already available geotagged data and study the improvements in performance in detection rate.

1.4 Thesis Outline

In this thesis, we propose a geolocalisation approach for inferring the location of non-geotagged tweets at a fine-grained level. Initially, in the first chapters, we focus on tackling the fine-grained geolocalisation problem. Next, we evaluate the effectiveness of the proposed approach in the context of a practical application (i.e., traffic incident detection). The remainder of this thesis is organised as follows:

Part I: Introduction and Background

Chapter 2 introduces the concepts this thesis relies on. Firstly, we provide concepts from classical IR such as retrieval, indexing and approaches for weighting documents (including Vector Space Models and Probabilistic models) that we will utilise through the work of this thesis. Secondly, we provide a literature overview of previous research regarding the geolocalisation of Twitter data. This overview includes reviews of the approaches proposed to tackle the two main problems in the area: Twitter user and Tweet geolocalisation. Finally, we introduce the problem of fine-grained geolocalisation of non-geotagged tweets and motivate the limitations of previous research for tackling this task.

Part II: Fine-Grained Geolocalisation of Tweets

Chapter 3 investigates the limitations of previous tweet geolocalisation approaches when working at fine-grained levels. We show that the strategy of existing approaches of aggregating geotagged tweets to represent a location leads to a loss of important evidence for fine-grained geolocalisation. To alleviate such limitations, we propose to avoid such aggregation and propose an approach for fine-grained geolocalisation based on a ranking approach of individual geotagged tweets. Finally, we experiment to demonstrate the effectiveness of our approach and provide insights to understand the drawbacks of existing state-of-the-art works.

Chapter 4 discusses the predictability of geolocation of tweets at a fine-grained level. We postulate that such predictability is given by a correlation between content similarity and geographical distance to other geotagged tweets. We extend our ranking of individual geotagged tweets by adopting a weighted majority voting algorithm to exploit the geographical evidence encoded within the Top-N geotagged tweet in the ranking.

Chapter 5 investigates the effects of the quality of the ranking on fine-grained geolocalisation. In particular, we propose a learning to rank approach that re-ranks individual geotagged tweets based on their geographical proximity to a given non-geotagged tweet. Moreover, we propose multiple features tailored for fine-grained geolocalisation of tweets and investigate the best performing combination of them.

Part III: Applicability of The Fine-Grained Geolocalisation Approach

Chapter 6 investigates the effectiveness of our proposed fine-grained geolocalisation approaches when applied in a practical application. In particular, we study the effectiveness of geolocalised tweets in the traffic incident detection task, which aims to detect real-time traffic disruptions using messages posted in the Twitter stream. We geolocalise new non-geotagged incident-related tweets and demonstrate that, when comparing to a ground truth of real incidents, our approaches can effectively infer their location. Moreover, we show how the overall effectiveness of the traffic incident detection task is improved when expanding the sample of incident-related geotagged tweets with new geolocalised incident-related tweets, compared to the performance when using geotagged tweets alone.

Part IV: Conclusions and Future Work

Chapter 7 provides conclusion remarks of the work undertaken in this thesis and discusses the new research questions that this thesis opens to the research community, and are worth to be investigated in the future.

1.5 Origin of The Material

The research material appeared in this thesis has been published in various journal and conference papers during the course of this PhD programme:

1. (Gonzalez Paule et al., 2017) “*On fine-grained geolocalisation of tweets*”. ICTIR’17, pages 313-316.

2. (Gonzalez Paule et al., 2018b) “*On fine-grained geolocalisation of tweets and real-time traffic incident detection*”. In Information Processing & Management.

3. (Gonzalez Paule et al., 2018a) “*Learning to Geolocalise Tweets at a Fine-Grained Level*”. CIKM’18, pages 1675-1678.

4. (Gonzalez Paule et al., 2019) “*Beyond geotagged tweets: exploring the geolocalisation of tweets for transportation applications*”. In Transportation Analytics in the Era of Big Data, Springer, pages 1–21.

In addition, the work undertaken during this PhD programme has lead to the publication of other research papers that have contributed to the fields of Geographical Sciences and Social Sciences. In particular:

5 (Thakuriah et al., 2016) "*Sensing spatiotemporal patterns in urban areas: analytics and visualizations using the integrated multimedia city data platform.*" Built Environment 42.3, pages 415-429.

6 (Sun and Gonzalez Paule, 2017). "*Spatial analysis of users-generated ratings of yelp venues.*" Open Geospatial Data, Software and Standards 2.1, pages 5.

Chapter 2

Background and Related Work

2.1 Chapter Overview

In this chapter, we introduce the necessary concepts, definitions and methods that will be used later in this thesis. In particular, we provide essential background for understanding the methodologies used in Part II. Now we provide an overview of the content of this chapter.

Firstly, in Section 2.2 we introduce the field of Information Retrieval (IR), that allows users to efficiently and effectively search for relevant information within large collections of text documents by means of a query. Then, the documents are ranked by the estimated relevance with respect to the user's query. We start by describing the main components of an IR system and how text documents are processed and indexed. Lastly, we describe how relevant documents are retrieved using a retrieval model. Methods and techniques explained in this section will be used later in our experiments. Therefore, we formalise and describe in detail the state-of-the-art retrieval models that will be used in Part II of this thesis. However, while IR systems rank documents based on relevance to a given query, given the nature of our task (geolocalisation of a tweet), we aim to rank tweets based on their geographical proximity to a given tweet as a query. The behaviour of IR models in the context of this task will be explored in further experiments in this thesis.

Next, in Section 2.3 we describe the challenges arisen when dealing with Twitter data in an IR system, which is the data source used through this thesis. Twitter messages have particular characteristics, they are short documents and are normally written in formal language. Because of this, state-of-the-art IR models, that were initially tailored to work with large text documents, under-perform when dealing with Twitter posts. For this reason, we introduce how the IR community have tackled this issue and the best ways to store and process Twitter posts. These methods will be crucial for the experiments undertaken later in this thesis.

Finally, in Section 2.4 we discuss related work regarding the geolocalisation of Twitter data. We introduce the field and discuss the two main task tackled by the research community: Twitter user geolocalisation and tweet geolocalisation. Since this thesis aims to investigate the tweet geolocalisation task, we then describe the main approaches that researchers have proposed in the past to address the problem. Lastly, we motivate the problem of inferring the geolocalisation of tweets at a fine-grained level and motivate the work in this thesis.

2.2 Information Retrieval Background

The field of information retrieval (IR) deals with the representation, storage, organisation of and access to information items (Baeza-Yates et al., 1999). The discipline was developed to efficiently and effectively access information contained in large collections of text documents written in natural language. The process of IR can be summarised as follows. Firstly, a user with an information need introduces a text query using natural language into an IR system that stores a collection of text documents. The collection is stored in the IR system using a data structure called *index*, which allows efficient access to the items. Secondly, the IR system processes the query and assigns to each document a score that represents an estimation of how the document matches the information need expressed by the user in the query, this is called relevance. Finally, the system presents the documents as a ranked list ordered by their level of estimated relevance.

The concept of relevance is key in IR. A depth of understanding of the decision making processes occurring in the human brain is needed to understand user's information need fully, and users normally express this poorly in their queries.

Instead, the IR system usually estimates relevance by calculating the similarities between the content of the query and the content of the documents. This is the goal of the retrieval model, which is the core of any IR system. In this thesis, we utilise retrieval models in order to assess the similarities between tweets for geolocalisation. Thus, in this chapter, we introduce a general background of the Information Retrieval (IR) concepts required to understand the topics explored in this thesis. The rest of the chapter is organised as follows: Section 2.2.1 introduces effective indexing strategies followed by retrieval systems. Section 2.2.2 discusses the principles and formalisation of retrieval models. Section 2.2.3 details the Learning to Rank framework that uses machine learning for information retrieval.

2.2.1 Indexing

In order to effectively search for relevant items within a collection of documents, retrieval systems perform a process named *indexing*. During indexing, first text documents are transformed into a bag-of-words representation and stored into an efficient data structure called *index*. In this section, we first describe how documents are transformed in Section 2.2.1.1 and discuss how the index is constructed to efficiently search the documents in Section 2.2.1.2.

2.2.1.1 Document Transformation

The first stage of the indexing process is transforming documents into a *bag-of-words* representation in order to store them into the index. The first step of document transformation is called *tokenisation*. In this process, documents are first decomposed into terms by identifying the boundaries (or separator) between tokens (or terms). All terms are lowercased, and all the punctuation is removed at this stage. For instance, given the following document taken from Carroll (2011):

> *"Begin at the beginning", the King said gravely, "and go on till you come to the end: then stop."*

after tokenisation, the above document is transformed into:

begin at the beginning the king said gravely and go on till you come to the end then stop

According to Luhn (1957), the discriminative power of a term is normally distributed with respect to the rank of its frequency in a collection. Moreover, common terms (e.g., “the”) will occur in almost all documents. Such are typically known as *stopwords*, and their informativeness in terms of the relevance of a document is null. For this reason, *stopwords* are removed from the text.

In order to determine which terms are considered for removal, a list of pre-compiled *stopwords* is used. Several stopwords lists have been proposed in the literature (Van Rijsbergen, 1979), and typically consisting of prepositions, articles and conjunctions. Also, other terms can be extracted automatically by determining the most frequent and less informative terms within a collection (Lo et al., 2005). After stopword removal, the example document above is reduced to the following:

begin beginning king said gravely you come end then stop

Usually, the term specified in a query by the user is not in the same syntactic variation as it is present in relevant documents (Baeza-Yates et al., 1999), which prevents perfect matching between the query and the document. For example, “I am living in Scotland” is a different syntactical variation than “I have lived in Scotland”. To overcome this issue, terms are transformed and reduced to their stem using a Stemming algorithm. For example, the words *fishing*, *fished*, and *fisher* are transformed to the root word, *fish*. The first stemming algorithm was proposed by Lovins (1968), which then influenced the Porter Stemming algorithm (Porter, 1980), which is the most popular. After applying stemming the example document is reduced to the following:

begin begin king said grave you come end then stop

Finally, the resulting text is transformed into a *bag-of-words* representation by counting the number of times a term occurs in the document. This set of terms in a document with their frequencies is the final representation of a document-terms list that is stored in the index data structured. The final document-terms list for the example document is shown in Table 2.1.

Table 2.1: A bag-of-words representation of a document.

Document	
Term	**Frequency**
begin	2
king	1
said	1
grave	1
you	1
come	1
end	1
then	1
stop	1

2.2.1.2 Index Data Structures

After each document of a collection is transformed into a *bag-of-words* representation they are stored into the index as a document-terms list. However, in order to score documents with respect to the terms in a query, the retrieval system is forced to iterate through all of the document-terms list. This has a cost of $O(N)$ time complexity, where N is the total number of documents in the collection, and this is not scalable to handle large collections. In order to efficiently score documents, an alternative data structure was proposed, called *inverted index* (Van Rijsbergen, 1979). This data structure transposes the document-terms list into a term-documents list. This way, the system only scores the subset of documents (D_q) that contain the terms of the query, which reduce time complexity to $O(D_p)$.

After the collection has been indexed, documents are ready for being ranked in response to a query based on a probability score given by a retrieval model. In the next section, we describe the traditional approaches for scoring and ranking documents.

2.2.2 Retrieval Models

Given an indexed collection of documents, the primary goal of an IR system is to rank documents based on their probability of meeting the information need of the user, which is expressed as a query. In order to fully understand how humans judge relevance with respect to their information needs, it would be necessary to understand the cognitive process of decision making that occurs in the human brain. Instead, IR researchers have developed theoretical assumptions that aim to capture how documents match information need given a query. These theoretical assumptions are then formalised into a mathematical model, named *retrieval model*. In the rest of this chapter, we detail the most well-known approaches for retrieval, including the models that we will utilise in this thesis.

Fundamentally, a retrieval model estimates *relevance* as a quantification of the similarities between a document and a query. Thus, IR models assume that the most similar documents to a given query are considered to be the most relevant to the user information needs. This is typically done by a weighting the model using statistical features of the document, the query and the collection. For this reason, retrieval models are also known as *document weighting models* or *IR weighting models*.

2.2.2.1 Boolean Model

One of the first models for document retrieval is the Boolean Model (Van Rijsbergen, 1979). The boolean model is based on set theory and boolean logic. Documents are considered as sets of terms with a binary weight to represent whether they occur in the document or not. Moreover, the boolean model has no information regarding term importance in the query, document or collection. Queries are composed as a combination of terms and boolean logic operators such as *AND*, *NOT* and *OR*, which state whether the presence of a term is required or excluded in the document. Due to the boolean nature of the query, a boolean relevance score is assigned to the documents; either *TRUE* or *FALSE*. Hence, the Boolean Model is also named as *exact-match retrieval* since only documents that match the query are retrieved. Because a binary relevance score is assigned to the documents, there is no ranking per se and documents are often ordered by other metadata information such as creation date or author.

The main drawback of the Boolean Model is that there is no partial matching to the query, i.e. the model does not provide a degree of relevance. This has an impact on effectiveness which mainly depends on how well the users formulate the queries. Moreover, query formulation based on boolean logic is unnatural and presents a difficult way for the user to express their information needs (Van Rijsbergen, 1979). Despite these disadvantages, the boolean model is utilised in several applications, such as patent search (Joho et al., 2010), due to its efficiency.

2.2.2.2 Vector Space Model

The Vector Space Model (VSM) was the focus of IR research in the 1970s and was proposed to overcome the limitations of the Boolean model. The main new advantages of the VSM is to allow partial matching of the query and incorporate estimations about the relevance of the documents (Dillon, 1983; Salton et al., 1975). Therefore, the resulting list of matched documents can be ranked according to their degree of relevance to a query. In order to do that, the VSM uses a n-dimensional space of Euclidean geometry, where n is the number of terms in the index (or collection), and each dimension represents the weight of the term.

Then, in the VSM, documents and queries are represented as vectors in the above mentioned n-dimensional Euclidean space. In particular, a document d_i is represented by a vector of terms $\vec{V}(d_i) = (d_{i,1}, d_{i,2}, ..., d_{i,n})$, where $d_{i,j}$ is the weight of the j-th term in the document. Likewise, a query q is represented as a vector of terms $\vec{V}(q) = (q_1, q_2, ..., q_n)$, where q_j is the weight of the j-th term in the query. In the most simple form of the VSM, the weight of each term is the raw count or term frequency (*tf*), which term provides a measure of the importance of the term in a document. Nevertheless, other approaches for term weighting has been explored. These approaches incorporate A new statistic named Inverse Document Frequency (*idf*) that was proposed by Sparck Jones (1972). The *idf* statistic calculates the number of documents over the entire collection where the term occurs at least once, and reflects the importance of a term in the entire collection. Finally, the *TF-IDF* weighting scheme is the most commonly used for weighting the vectors, thus the *tf-idf* of the term w in a document d_i can be defined as:

$$tf_{i,w} \times idf_w = tf_{i,w} \cdot \log \frac{N}{df_w} \tag{2.1}$$

where N is the number of documents in the collection, $tf_{i,w}$ is the term frequency of the term w in the document d_i, and df_w is the number of documents in the collection where w appears at least once.

Once the document and the query vectors are constructed, the Euclidean distance can be used to compute the level of similarity (as an estimation of their relevance). However, instead of calculating the distance (or dissimilarity) a similarity measure is commonly employed to predict relevance. Therefore, documents with the highest scores are considered the most similar and, therefore, should be ranked at the top of the list.

Several similarity measures have been proposed in the literature (Van Rijsbergen, 1979). The most popular is known as the *cosine similarity*, which we utilise in this thesis. The cosine similarity computes the cosine of the angle θ between two vectors. Thus, the similarity between the document d_i and the query q is calculated as the cosine of the angle θ between the document vector $\vec{V}(d_i)$ and the query vector $\vec{V}(q)$ defined as:

$$similarity(d_i, q) = cosine\theta_{d_i,q} = \frac{\vec{V}(q) \cdot \vec{V}(d_i)}{|\vec{V}(q)| \cdot |\vec{V}(d_i)|} \tag{2.2}$$

2.2.2.3 Probabilistic Models: BM25

Previous retrieval models assessed relevance in different ways. The Boolean model determined relevance by a binary decision of the existence of the query terms in the document. Then, in the Vector Space Model relevance is determined by the cosine similarity of two weighted vectors in a Euclidean space representing the document and the query. However, relevance can also be quantified as a value that measures the level of uncertainty that the content of a document is relevant to the user's information need. This is the basic principle of Probabilistic Retrieval Models, that are rooted by the *Probability Ranking Principle* (PRP) (Cooper, 1971; Robertson, 1977) and is based on the foundations of probability theory. The PRP is stated as:

> *"If a reference retrieval system's response to each request is ranking of the documents in the collections in order of decreasing probability of relevance to the*

user who submitted the request, where the probabilities are estimated as accurately as possible on the basis whatever data have been made available of the system for this purpose, the overall effectiveness of the system to its user will be the best that is obtainable on the basis of those data."

The PRP assumes that the probability of relevance of a document to a query is independent of other documents. Based on this assumption, and applying the Bayes Theorem, a new probabilistic weighting model for retrieval can be derived. The most notable model is the Okapi BM25 (Robertson et al., 1995), which will be used in this thesis. In the Okapi BM25 the probability of a document d to be relevant to a given query q is defined as follows:

$$\mathrm{P}(rel|d,q) \propto \sum_{t \in q} \log \left(\frac{N - df_i + 0.5}{df_i + 0.5} \right) \cdot \frac{(k_1 + 1) \cdot tf_i}{k1((1-b) + b\frac{dl}{\mathrm{avgdl}}) + tf_i}$$

where tf_i represents the frequency of the term in the document, df_i is the document frequency of the term, and document length is represented as dl. Document length is normalised by dividing the length of the document by the average document length of the collection $avgdl$. The model is tuned using two parameters; $k1$ and b. By adjusting $k1$ we control the influence of term frequency tf_i in the final score, whereas adjusting b varies the influence of document length normalisation $\frac{dl}{\mathrm{avgdl}}$.

2.2.2.4 Language Modelling

Statistical Language modelling has been applied to predict the next term given an observed sequence of terms. Thus, a language model is a probability distribution over sequences of terms (Manning et al., 1999). In the context of IR, a language model represents, in essence, the probability of observing a term in a document. Language modelling was introduced as a ranking approach in the late 1990s (Berger and Lafferty, 2017; Hiemstra, 1998; Miller et al., 1999b; Ponte and Croft, 1998). From a statistical perspective, language models (LM) are fundamentally different to probabilistic models (PM) in Section 2.2.2.3. Probabilistic models determine relevance for a document given a query, whereas language models calculate the probability of a query of being generated by a document.

The Language Modelling (LM) approach attempts to model the process of generating of a query (Ponte and Croft, 1998) given a document. The approach assumes that a query q is generated by a probabilistic model based on observations of terms in a document d. Thus, we aim to calculate the conditional probability $P(d/q)$. By applying Bayes' rule we obtain:

$$P(d|q) = \frac{p(q|d)p(d)}{p(q)} \propto p(q|d)p(d) \tag{2.3}$$

where $p(d)$ is the prior belief that the document is relevant to any query, and $p(q|d)$ is the query likelihood given the document. Note that $p(q)$ is ignored as it is the same for every document in the collection, and therefore does not affect the ranking of the documents in response to a query. The prior $p(d)$ is mostly assumed to be uniformly distributed (Berger and Lafferty, 2017; Hiemstra, 1998; Ponte and Croft, 1998), but many alternative priors has been also investigated in the past (Miller et al., 1999a). In this thesis, we assume a uniform prior distribution. After this simplification, the model is reduced to the task of estimating $p(q|d)$, the probability of observing the query q given the document d. Thus, using a multinominal unigram language model, the probability of generating the query terms using document d is formalised as:

$$Score_{QLM}(q,d) = p(q|d) \propto \prod_{t \in q} p(t|\theta_d)^{tf_{t,q}} \tag{2.4}$$

where $p(t|d)$ is the probability of observing a term t of the query given the language model θ_d for document d, and $tf_{t,q}$ denotes the term frequency of the term t in the query q. Note that, in order to calculate $p(t|\theta_d)$, a sparsity problem appears as a term t in a query may not be present in the document d. This is called the zero probability problem. To tackle the problem of zero probabilities for unseen terms, the language model of the document is complemented with the collection model, which has knowledge of any term in the entire collection. This technique is known as *smoothing*, and various strategies for doing so have been proposed in the literature (Zhai and Lafferty, 2017): Jelinek-Mercer, Dirichlet and Absolute discounting.

In this thesis, we will deal with Twitter data as we explain later in Section 2.3. Experiments realised by Zhai and Lafferty (2017) showed that Dirichlet smoothing performs the best when using title queries, that are short queries containing mostly two or three keywords. This is in line with the average query length in Twitter search reported by Teevan et al. (2011) (1.64 words per query). Additionally, language models with Dirichlet smoothing have been used as the baseline retrieval models for the 2013 and 2014 instances of the microblog search tracks (Lin and Efron, 2013; Lin et al., 2014) that we introduce in detail later in Section 2.3. For these reasons, in our experiments, we apply the language model approach that applies Dirichlet smoothing, which we will describe next.

Dirichlet Smoothing. For any language model, the general form for smoothing is given by:

$$P(t|d) = \begin{cases} p_s(t|d) & \text{if term t is seen,} \\ \alpha_d p(t|C) & \text{otherwise} \end{cases} \tag{2.5}$$

where $p_s(t|d)$ is the probability of a term in the document d, $\alpha_p(t|C)$ is the probability of a term in the entire collection C and α_d is a coefficient that controls the probability assigned to unseen terms. In the Language Model with Dirichlet smoothing, the prior distribution of terms in the collection is given by a Dirichlet distribution with parameters $(\mu p(t_1|C), \mu p(t_2|C), ..., \mu p(t_n|C))$. Thus, the model is given by:

$$p(t|d) = \frac{tf_{t,d} + \mu p(t|C)}{\sum_t tf_{w,d} + \mu} \tag{2.6}$$

where $tf_{t,d}$ is the frequency of the term t in the document d, and μ is the controlling coefficient for the smoothing.

2.2.2.5 Divergence From Randomness

Divergence From Randomness (DFR) is a probabilistic approach that works under the assumption that the more the content of a document diverges from a random distribution, the more informative the document is Amati (2003). Therefore, the most informative terms are distributed over an elite set of documents, whereas the non-informative terms are randomly distributed over the entire collection (Bookstein and Swanson, 1974; Damerau, 1965; Harter, 1975a,b). The underlying hypothesis of DFR models is:

> *"The informative content of a term can be measured by examining how much the term frequency distribution departs from [...] the distribution described by a random process"* (Amati, 2003)

Thus, to compute the importance of a given term t in a document d, the DFR models calculate the distribution of its term frequency tf in the documents, and compute its divergence from a distribution generated through a random process. The standard DFR model, given a query q and a document d, is defined as:

$$Score_{DFR}(d,q) = \sum_{t \in q} w_{t,q} w_{t,d} \tag{2.7}$$

where $w_{t,q}$ is the normalised frequency of term t in the query q, given by:

$$w_{t,q} = \frac{tf_{t,q}}{\max_{t_i \in q} tf_{t_i,q}} \tag{2.8}$$

and $w_{t,d}$ is the weight of a term t in a document d is given by:

$$w_{t,d} = inf_1 inf_2 \tag{2.9}$$

The frequency of the term in the document $w_{t,d}$ is composed by $inf_1 = -\log_2 p_1(t|C)$ and $inf_2 = 1 - p_2(t|d)$, which defines the informativeness of the term t in the entire collection C and in a document d that contains the term, respectively.

The component $p_1(t|C)$ is named the *basic randomness model* of the distribution of term t in the entire collection C. The most used basic models are the following[1] (Amati, 2003): divergence approximation of the binomial (D), approximation of the binomial (P), Bose-einstein distribution (B_e), geometric approximation of the Bose-einstein (G), inverse document frequency model ($I(n)$), inverse term-frequency model ($I(F)$), and inverse expected document frequency model ($I(n_e)$).

On the other hand, the $p_2(t|d)$ component defines the *information gain* of observing the term t in the document d. This can be computed using two models: Laplace (L) model:

$$L = \left(\frac{1}{tf_{t,d} + 1} \right) \tag{2.10}$$

and the ratio of two Bernoulli's process (B):

$$B = \left(\frac{F}{df_{t,c}(tf_{t,d} + 1)} \right) \tag{2.11}$$

However, a third component is needed for DFR models. Because the amount of information in a document is in proportion to its length, a document length normalisation is needed, called $Normalisation2$, as defined bellow:

$$tf_n = tf \cdot \log \left(1 + c \cdot \frac{avgdl}{dl} \right) \tag{2.12}$$

In this thesis, we experiment with different combinations of the components mentioned above, which configure different DFR models. We now briefly introduce them as described in (Amati, 2003):

InB2: Inverse Document Frequent model with Bernoulli after-effect and normalisation 2.

$$w_{t,d} = \frac{F+1}{n_t \cdot (tfn + 1)} (tfn \cdot \log_2 \frac{N+1}{n_t + 0.5}) \tag{2.13}$$

[1] As described in the Terrier IR platform (Ounis et al., 2006) (http://terrier.org/docs/v3.5/dfr_description.html)

IneB2: Inverse Expected Document Frequent model with Bernoulli after-effect and normalisation 2.

$$w_{t,d} = \frac{F+1}{n_t \cdot (tfn+1)} (tfn \cdot \log_2 \frac{N+1}{n_e+0.5}) \tag{2.14}$$

IFB2: Inverse Term Frequency model with Bernoulli after-effect and normalisation 2.

$$w_{t,d} = \frac{F+1}{n_t \cdot (tfn+1)} (tfn \cdot \log_2 \frac{N+1}{F+0.5}) \tag{2.15}$$

InL2: Inverse Document Frequency model with Laplace after-effect and normalisation 2.

$$w_{t,d} = \frac{1}{tfn+1} (tfn \cdot \log_2 \frac{N+1}{n_t+0.5}) \tag{2.16}$$

PL2: Poisson model with Laplace after-effect and normalisation 2.

$$w_{t,d} = \frac{1}{tfn+1} \big(tfn \cdot \log_2 \frac{tfn}{\lambda} + (\lambda - tfn) \cdot \log_2 e + 0.5 \cdot \log_2 (2\pi \cdot tfn)\big) \\ \text{with } \lambda = \frac{F}{N} \text{ and } F \ll N \tag{2.17}$$

where tf is the within-document frequency of t in d, n_t is the document frequency of t, F is the term frequency of t in the whole collection, N is the number of document in the whole collection, n_e is the number of expected documents containing the term according to the binomial law (See Amati (2003); Section 4.5.2).

2.2.3 Learning to Rank

The ranking approaches described in previous sections aim to capture the relevance of a document for a given query. However, such models can be effective in specific search scenarios, but it is improbable that they can also be effective in all search scenarios Zhai and Fang (2013). This issue is particularly true for Web retrieval, due to the diversity in size and content of web pages and the variability and complexity of the information needs of the users that search the web (Liu et al., 2009). However, any single of such models can capture different aspects of the relevance of a document. Thus, by combining them as multiple features in a machine-learned ranking function, we can potentially provide effective results in multiple search scenarios. This is the aim of Learning To Rank (L2R) approaches.

In the basic form of L2R approaches, features are extracted from a set of ranked documents to a given query to train a ranking function. This function is then applied to re-rank unseen document rankings and increase the desire ranking criteria (i.e., relevance in web retrieval) at the top documents in the list. Figure 2.1 present the general framework of learning to rank. As defined by (Liu et al., 2009), in order to train the ranking function the learning to rank approach uses an initial *sample* of ranked documents with respect to a query, called training set. This sample of query-document pairs should have high-recall and should have many relevant documents (Macdonald et al., 2013).

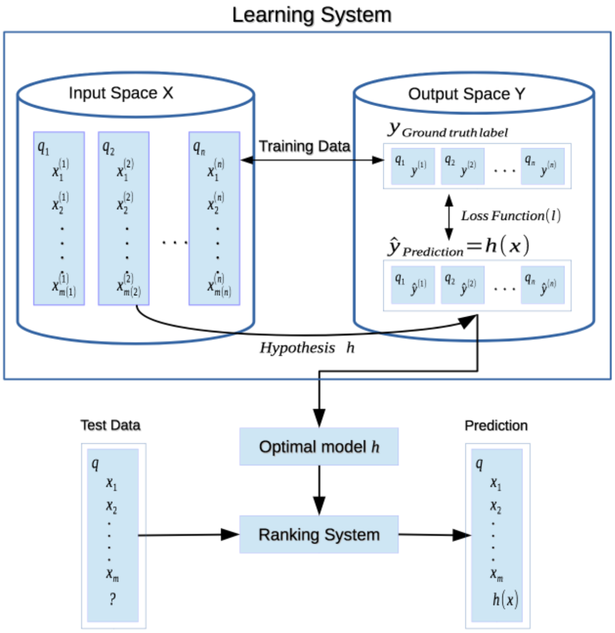
Learning System
Input Space X
q_1 $x_1^{(1)}$ $x_2^{(1)}$ $x_{m(1)}^{(1)}$
q_2 $x_1^{(2)}$ $x_2^{(2)}$ $x_{m(2)}^{(2)}$
q_n $x_1^{(n)}$ $x_2^{(n)}$ $x_{m(n)}^{(n)}$
Training Data
Output Space Y
y Ground truth label
q_1 $y^{(1)}$ q_2 $y^{(2)}$ q_n $y^{(n)}$
Loss Function(l)
$\hat{y}$ Prediction $=h(x)$
q_1 $\hat{y}^{(1)}$ q_2 $\hat{y}^{(2)}$ q_n $\hat{y}^{(n)}$
Hypothesis h
Test Data
q x_1 x_2 x_m ?
Optimal model h
Ranking System
Prediction
q x_1 x_2 x_m $h(x)$

Figure 2.1: Learning to Rank Framework (Liu et al., 2009).

As a ranking function, many algorithms have been proposed in the literature and can be categorised in three groups based on the number of documents that are taken into account for learning: point-wise approaches (Breiman, 2001; Friedman, 2001), pair-wise approaches (Burges et al., 2007; Wu et al., 2010) and list-wise approaches (Metzler and Croft, 2007; Xu and Li, 2007). Point-wise approaches consider each document in the sample ranking independently, pair-wise approaches consider pairs of documents and list-wise approaches optimise an information retrieval measure and consider the entire ranking list at one time. Prior works have shown that list-wise approaches are the most effective (Liu et al., 2009). However, this performance has not been investigated in the specific task of this work (tweet geolocalisation). In Chapter 5, we experiment with several types of L2R approaches and identify the best performing ones on our task.

2.3 Information Retrieval for Twitter

State-of-the-art Information Retrieval (IR) models are mainly tailored to provide a relevance estimation score to large text documents. The most common application of IR models is the search of web documents (Arasu et al., 2001), where the issues of IR models to work with the specific characteristics of web pages have been widely studied (Croft et al., 2010). However, with the appearance of Twitter, it also appeared the necessity of searching for information in Twitter posts. The task of information retrieval in the context of Twitter, where users issues textual queries to a search engine to find relevant previously published tweets, is named "*Microblog Retrieval*" in the literature.

Due to the social characteristics of the Twitter content, the way how and why users search in Twitter differs from how users search the Web (Teevan et al., 2011). Users' queries in Twitter search are shorter and are more repetitive to track specific results about social events. On the other hand, Web queries are more changing, and users develop more queries in order to find more information about a topic.

In this section, we first discuss the specific structural characteristics of Twitter documents and how they are preprocessed and indexed into the retrieval system. Next, we discuss recent research on the applicability of IR models in the context of a microblog search task.

2.3.1 Tweet Indexing

Twitter documents differ from web and traditional documents in many ways. First, they are short documents with an initial maximum length of 140 characters as 2016, but it was increased to 280 in 2017. In this thesis, we will experiment with Twitter messages posted during 2016, so we will use 140 characters long documents. Second, Twitter messages can contain informal and slang language in their text; such as abbreviations (e.g., BRB for *Be Right Back*, or LOL for *Laughing Out Loud*) or spelling errors. Third, Twitter provides users with ways to interact with other users and propagate their messages into a topic discussion. Users can use the called *mentions*, which are ways to mention other users in their text. A mention consist of the character @ followed by a user name (e.g., *@Salias* for mentioning the user "Salias"). Moreover, users have the possibility of adding *hashtags*, consisting of the character # followed by a keyword, to specify the topic of their tweet (e.g., *#indyref* for a tweet about the Scottish independence referendum). Finally, Twitter messages can contain URLs or hyperlinks to external websites (e.g., http://www.dcs.gla.ac.uk).

2.3.1.1 Tweet Document Transformation

Identical to the indexing process explained before in Section 2.2.1, tweets are transformed into a *bag-of-words* representation before storing them in the index structure. However, due to the singularities of Twitter data, extra preprocessing steps are needed in order to remove the specific characteristics described above to obtain a final representation.

- **Emotion Removal:** Remove words or symbols that express feeling or emotions, such as *lol*, *haha* or *xoxo*
- **Stopwords Removal:** Due to the informal language of tweets, an extended stopword list is needed for this process. This extended list should contain, apart from the common stopwords discussed in Section 2.2.1, informal version of them such as *gonna* or *ain't*.
- **HashTag, Mention and HyperLink Removal:** Remove username mentions, hashtags and links to external websites appeared in the text.

The effects on retrieval effectiveness of different combinations of the preprocessing steps described above have been studied by Thomas (2012). The study concluded that the best performing is achieved when all the preprocessing steps were applied. Nevertheless, some specific combinations might be beneficial depending on the final objective of the task. Therefore, relevant information is preserved. For example, avoiding emotion removal is essential for the sentiment analysis task (Agarwal et al., 2011; Baucom et al., 2013; Kouloumpis et al., 2011; Pak and Paroubek, 2010). In further experiments, we will explain and motivate the most suitable preprocessing steps for the research undertaken in this thesis.

2.3.2 Microblog Retrieval

Due to the rising importance of Twitter documents, IR researchers have investigated the challenges of searching Twitter posts. Since 2011, the TREC[1] conference, sponsored by the National Institute of Technology (NIST) and the U.S. Department of Defense, have organised a number of *Microblog Retrieval Tracks* (Lin and Efron, 2013; Lin et al., 2014, 2015; Ounis et al., 2011; Soboroff et al., 2012) to gather the IR research community and together address the problem. Consequently, several participants attempted to improve the retrieval performance by submitting their adapted retrieval techniques to the track; including document expansion (Jabeur et al., 2013), query expansion (Aboulnaga et al., 2012; Rodriguez Perez et al., 2013; Yang et al., 2013) and learning to rank (L2R) (Gao et al., 2013; Zhu et al., 2013).

The solutions proposed on the *Microblog Retrieval Tracks* focused on increasing the performance of the retrieval of Twitter posts. However, they do not provide an in-depth study of the behaviour of the state-of-the-art retrieval models in the context of microblog search (described in Section 2.2.2). This has been the focus of recent research that has identified the main problems affecting retrieval models in Twitter search. For example, Ferguson et al. (2012) and Naveed et al. (2011) found that, due to the short length of tweets, using document normalisation will affect the performance of the task negatively, and the benefits of applying term frequency weighting are minor.

[1] http://trec.nist.gov/

Moreover, more recently Rodriguez Perez and Jose (2015) and Rodriguez Perez (2018) confirmed these findings and performed an exhaustive investigation of the problems of the state-of-the-art retrieval models in microblog search. Their findings showed that models relying on term frequency and document normalisation performed poorly compared to models relying only on document frequency information. These observations are crucial to understanding the results obtained in further experiments in this thesis.

In the remainder of this chapter, we discuss related literature regarding the geolocalisation of Twitter data and motivate the work of this thesis.

2.4 Geolocalisation of Twitter Data

In recent years, social media services have gained increasing popularity within the research community. Specifically, Twitter has become very popular since their data is generated in real-time and geographical information is attached to the posts. Such characteristics have provided new opportunities for a broad range of real-time applications, such as real-time event detection (Atefeh and Khreich, 2015; Crooks et al., 2013; Sakaki et al., 2010; Walther and Kaisser, 2013; Watanabe et al., 2011; Xia et al., 2014; Zhang et al., 2016a),, that exploits such combination of textual and geotagged information for their analysis. Geographical information is attached to tweets in two ways: (i) the exact longitude and latitude if the GPS location of the user device is activated; and (ii) as a suggested area from a list that can be extrapolated to a polygon, that is available to the users when sending a tweet. Despite such options being available, only a very small sample of messages (around 1%) in the Twitter stream contains geographical information (Graham et al., 2014). In order to increase this sample, researchers have tackled the challenge of inferring the geolocation of Twitter data.

There are two main objectives in the literature regarding geolocalisation on Twitter data. First, some approaches have aimed to infer the home location of Twitter users (Chang et al., 2012; Cheng et al., 2010; Eisenstein et al., 2010b; Han and Cook, 2013), whereas other approaches aimed to infer the location where the user posted a tweet, or the location the users are tweeting about (i.e., the location of an individual tweet). This differentiation is important depending on the use case of the Twitter data. For example, for market research, the home location of a user is important. On the other hand, in other applications, such as emergency management (Ao et al., 2014; Imran et al., 2015; McCreadie et al., 2016), the location of each individual tweet is relevant. In this thesis, we focus on the geolocalisation of individual tweets and explore their applicability in the traffic incident detection task, which aims to use Twitter as a data source for detecting traffic incidents occurring in a city (see Chapter 6). In the next sections, we provide an overview of the current state-of-the-art in tweet geolocalisation.

2.4.1 Tweet Geolocalisation

Many researchers have tackled the problem of geolocalising individual tweets in the past. In order to infer the location of tweets, researchers have mainly exploited the evidence gathered from the text of the tweet and its metadata. In order to obtain a predicted location, three main strategies have been adopted in the literature. Firstly, in Section 2.4.1.1 we describe existing approaches that rely on external geographical databases, called *gazetteer*, in order to obtain the location of place names mentioned in the text and the user profile. Second, in Section 2.4.1.2 we describe more recent approaches that exploit the text of the 1% geotagged tweets available in the Twitter stream for geolocalising, which is the strategy we follow in this thesis. Finally, in Section 2.4.1.3, we describe recent work that uses neural networks for predicting the location label (i.e., country or city) of tweets.

2.4.1.1 Using External Sources (Gazetteer)

Schulz et al. (2013a) extracted different spatial indicators from the text and the user profile. These spatial indicators are mapped into different databases containing geospatial information using different methods, such as *DBpedia Spotlight*[1] or *Geonames*[2]. Each of the methods produces a polygon that represents a geographical area. Moreover, each method is associated with a confidence level that is then added to the polygon as a third dimension (height of the polygon). Finally, all the 3-D polygons obtained for a tweet are combined using a stacking algorithm

[1] https://wiki.dbpedia.org/
[2] http://www.geonames.org/

that returns the overlapping area with the highest confidence as the predicted location.

The drawback of the above work is that the information in the user profile is not always accurate. For instance, it is known that 34% of the users report fake locations in their user profile as the location field (Hecht et al., 2011), which can produce misleading predictions. On the other hand, Schulz et al. (2013a) looked for place names in the text that matched with an entry in a geographical database (*gazzeter*). However, place names in the text can be ambiguous. For example, *Glasgow* may refer to a city in the UK or a city in the USA. The same way, people do not always mention places using their formal names, for example, the city of *Barcelona* (Spain) is also referred as *Barna*. The problem of resolving this ambiguity is known as *toponym recognition*, and several models have been developed to solve it (Ji et al., 2016; Li and Sun, 2014, 2017).

2.4.1.2 Exploiting Geotagged Tweets

Due to the ambiguity problem occurring when matching the text of a tweet with geographical databases, it seems more convenient to do so using another geotagged dataset that shares the same characteristics. Therefore, recent work has used the small percentage of tweets that are already geotagged in the Twitter stream (Graham et al., 2014) as training documents for their models. In order to do that, these works have followed two different approaches for dividing the geographical space and mapping the textual information of the geotagged tweets.

The first approach used in the literature opted for representing the geographical space by clusters based on the density of geotagged tweets in different areas. Firstly, Eisenstein et al. (2010b) and Priedhorsky et al. (2014) applied Gaussian Mixture Models (GMM) to generate geographic density estimates for all the n-grams contained in the tweet. Then, the predicted location is given by the weighted sum of all the density estimates obtained for a tweet. Lastly, Flatow et al. (2015) adopted an iterative process that fits a Gaussian model for a given n-gram, using the coordinate points of the geotagged tweets that contain the n-gram. An n-gram is geospecific if we can create an ellipse (using the Gaussian model) that covers a predefined maximum area and contains at least a certain ratio of the total tweets. Then, the centre of the ellipse of the longest geospecific n-gram contained in the test tweet is returned as the predicted location.

The last and most popular approach in the literature divided the geographical space as a grid of predefined areas of a given size, and then modelled the language for each area to perform the prediction (Hulden et al., 2015; Kinsella et al., 2011; Paraskevopoulos and Palpanas, 2015; Roller et al., 2012; Wing and Baldridge, 2011). Next, these approaches calculate the probability of a tweet to be generated in an area based on its similarity to the geotagged tweets in the area and return the most similar area as the predicted location. In order to obtain the most likely area, these approaches have adopted classification and information retrieval techniques.

An example of these works is the approach proposed by Hulden et al. (2015). The authors divided the geographical area of the earth using a grid structure of squared cells of size length 1^o (≈111 kilometres). After aggregating the texts of the tweets in each cell, they used a Multinomial Naive Bayes and Kullback-Leibler divergence functions and incorporated words counts as features. Additionally, the authors extended these functions by adding to each cell, instead of word counts, a density measure estimated using a Gaussian Kernel.

On the other hand, Roller et al. (2012) used language models using an adaptive grid that is created from the geotagged tweets using a *kd*-tree algorithm. The *kd*-tree algorithm generates cells with size computed according to the density of geotagged tweets in the area. This provides a finer granularity in dense regions and coarse granularity elsewhere. Additionally, Kinsella et al. (2011) also used language models to compute the probability of a tweet being generated in a geolocation. However, they divided the space into zip codes instead of squared cells of a grid. Finally, Paraskevopoulos and Palpanas (2015) used a Vector Space Model (VSM) with TF-IDF weighting to rank locations based on their content similarity.

2.4.1.3 Using Neural Networks

More recently, another set of works used neural networks to predict the location of a tweet. First, Huang and Carley (2017) proposed a Convolutional Neural Network (CNN) using features extracted from the content and the user profile of the tweet. They trained the CNN to obtain high-level text features for predicting the location label (country or city) of tweets. Their approach achieved a 52% and 92% of accuracy on city-level and country-level prediction, respectively. More recently, Kumar and Singh (2019) proposed to use CNN to extract location information from the text of the text, such as place names or city names. However, their method did not provide a predicted geographical location (i.e., longitude/latitude coordinate or geographical polygon).

On the other hand, another set of works have adopted word embeddings for predicting the city of tweets. For instance, Miura et al. (2016) proposed an ensemble approach that created vector representations of the words in the text, location field or user description, that are then concatenated into a single vector to compound a full tweet representation. A softmax function is finally used to select the most likely class. The authors evaluated their model in the context of the W-NUT Twitter Geolocation Prediction Shared Task (Han et al., 2016), achieving an accuracy of 47.6%. More recently (Miura et al., 2017), the authors refined their approach by unifying the same vector representations through an attention mechanism to avoid ensemble methods, increasing the accuracy in the W-NUT dataset up to 56.7%.

2.4.2 From Coarse-Grained Level to Fine-Grained Level

Previous studies inferred the geolocation of tweet 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 thesis is the geolocalisation of Twitter posts at a fine-grained level – i.e. street or neighbourhood level. This is important for tasks that require fine-grained geolocated data, such as emergency management or the traffic incident detection task that we explore in Chapter 6. To this end, recent work has attempted to tackle fine-grained geolocalisation by adapting previous approaches to work at that level of granularity. To do so, they reduced the granularity of the cells of the grid that divides the geographical space. Firstly, Kinsella et al. (2011) reduced each cell of the grid to a zip code area. Then, Paraskevopoulos and Palpanas (2015) refined the work by Kinsella et al. (2011) by dividing the geographical space into fine-grained squares of size 1 km. However, their results showed that reducing granularity also decreases accuracy and their approaches demonstrated to be limited in the context of fine-grained geolocalisation. In this thesis, we aim to improve the state-of-the-art further and enable effective fine-grained geolocalisation of tweets.

More recent work has been developed in parallel to this thesis. In their work Ozdikis et al. (2018b) used Ripley's K function to find the co-occurrence distributions of pairs of terms or bigrams, and compare them to the spatial patterns of their unigrams to then identify clustering or dispersion tendencies between them. Then, bigrams with a spatially significant pattern with respect to their unigrams are added as features for classifying the most likely location, which is represented as cells of a grid. Another work by Ozdikis et al. (2018a) used Kernel Density Estimations to obtain probability distributions of terms. Then, the probability distributions of all the terms in a tweet are combined in an obtain the cell that maximises the cumulative probability. Another work by Bakerman et al. (2018) uses Gaussian Mixture Models and refined the work by Priedhorsky et al. (2014) by combining textual features and information about the Twitter network.

Finally, Table 2.2 shows a summary of the existing approaches described in this section. For each reference, we report their algorithmic technique (*Inference Model*) and their strategy to represent the geographical space that corresponds to:

- *Grid* for the approaches that divides the area into a grid,
- *Density* for models that use estimators to obtain a density area, or
- *Gazzeter* for models that utilise external geographical databases.

Also, we report the minimum granularity reported by the authors.

2.4.3 Tweet Geolocation Datasets

Several datasets from the literature have been published online for research purposes. For instance, Eisenstein et al. (2010a) released their GEOTEXT[1] dataset which contains 377,616 messages collected over one week of March 2010 within the

[1] http://www.cs.cmu.edu/~ark/GeoText/

Table 2.2: Summary of the state-of-the-art Tweet geolocalisation approaches.

Reference	Year	Geo. Division	Inference Model	Granularity
Eisenstein et al. (2010b)	2010	Density	Topic Modelling	Country and City level
Kinsella et al. (2011)	2011	Grid	Language Model	Zip Code and City level
Wing and Baldridge (2011)	2011	Grid	Naive Bayes & Kullback-Leibler	0.1° (≈ 11.13 km)
Roller et al. (2012)	2012	Adaptative Grid	Language Model	0.1° (≈ 11.13 km)
Schulz et al. (2013a)	2013	Gazzeter	Polygon Stacking	City level
Priedhorsky et al. (2014)	2014	Density	Gaussian Mixture Model	City level
Hulden et al. (2015)	2015	Grid+Density	Naive Bayes & Kullback-Leibler	1° (≈ 111.31 km)
Paraskevopoulos and Palpanas (2015)	2015	Grid	VSM (TF-IDF weighting)	1 km
Flatow et al. (2015)	2015	Density	Gaussian model	2 km
Ozdikis et al. (2018b)	2018	Grid	Multinomial Naive Bayes	1 km
Ozdikis et al. (2018b)	2018	Grid	Kernel Density Mixture	1 km
Bakerman et al. (2018)	2018	Density	Gaussian mixture models	Country and City level

United States. Same way, Roller et al. (2012) published the UTGeo2011[1] composed of 390 million tweets collected worldwide during September and November 2011. Moreover, Han et al. (2012a) used a similar dataset[2] with 26 million geotagged tweet that covers the entire globe collected during January 2014. More recently, Hulden et al. (2015) released the WORLDTWEET[3] dataset containing over 4 million geotagged tweets distributed worldwide and generated during January 2014. Finally, the W-NUT 2016 tweet geolocation shared task (Han et al., 2016) made available a global dataset of approximately 12.8 million geotagged tweets collected from 2013 to 2016.

The works mentioned used the Twitter Public Stream[4] to collect real-time tweets, which provides a 0.95% sample of the complete public tweets (Wang et al., 2015). Previous works collected these datasets for inferring the location of a tweet at a coarse-grained level of granularity (i.e., country or city level) and, thus, the authors used a broad spatial filter[5] on the Twitter stream to obtain tweets from all over the globe, or from a specific country. However, in this work, we aim to infer the locations of tweets at a fine-grained level and, thus, we need a representative sample of geotagged tweets belonging to a smaller region (i.e., city or metropolitan area). For this reason, we collect our datasets by applying a

[1] https://github.com/utcompling/textgrounder/wiki/RollerEtAl_EMNLP2012

[2] https://sites.google.com/a/student.unimelb.edu.au/hanb/research

[3] http://geoloc-kde.googlecode.com

[4] https://dev.twitter.com/streaming/public

[5] https://developer.twitter.com/en/docs/tweets/filter-realtime/guides/basic-stream-parameters#locations

small spatial filter covering two main USA cities. We describe our datasets later in Chapter 3, Section 3.3.1.

2.5 Conclusions

In this chapter, we have presented a general background of the Information Retrieval (IR) field and techniques. First, we described how text documents are efficiently indexed to then be retrieved using a retrieval model that ranks documents in relevant order to the user query. We focused on describing the retrieval models used in this thesis. Finally, we introduced how machine learning is applied to IR in order to increase the effectiveness of a given retrieved ranking. This set of techniques are called *Learning to Rank*.

Next, we discussed the current literature of microblog search where, due to the inherent characteristics of Twitter posts, researchers have aimed to tackle the problem of finding approaches for searching Twitter posts effectively. We discussed the *TREC Microblog Search Tracks* and more recent investigations of the behaviour of state-of-the-art retrieval models in the context of Twitter search, which are key for the understanding of the experiments undertaken later in this thesis.

Finally, we presented an overview of the existing literature on the problem of inferring the geolocation of Twitter data. We focused on the tweet geolocalisation problem, which aims to infer the location where individual Twitter messages were posted, or where such places are mentioned in the text. We described the three main strategies followed in the literature to map the information in the tweet with geographical information. Firstly, in Section 2.4.1.1, we described work that finds specific n-grams in the text and maps them with external databases to obtain an associated geographical area. We discussed the ambiguity problem that these approaches are facing and introduced the second strategy in the literature that overcomes this issue. In the second strategy (Section 2.4.1.2), previous works exploited the similarities between a given tweet and the text of the geotagged tweets available in the Twitter stream. The geographical space is divided into discrete areas, and then each geotagged tweet is associated with their corresponding area.

Additionally, we describe the techniques used in the literature to compute the similarity between a tweet and the geotagged tweets in an area. Finally, we discussed that when applied to work at a fine-grained level, these approaches showed a decrease in accuracy, which motivates the work in this thesis. Lastly, in Section 2.4.1.3, we discussed recent work that adopted neural networks (i.e., word embedding and deep learning) to predict the location label of tweets. Besides, in Section 2.4.3, we provided an overview of the research datasets available in the literature.

The remainder of this thesis is structured as follows. Firstly, Chapter 3 studies the limits of the tweet geolocalisation models when applied to work at a fine-grained level, and propose a novel approach based on a ranking of geotagged tweets using IR retrieval models. Secondly, in Chapter 4 we improve the accuracy of the geolocalisation further by exploring the geographical evidence encoded within the Top-N most similar geotagged tweets ranked using the approach introduced in Chapter 3. Thirdly, in Chapter 5 we explore whether increasing the effectiveness of the ranking can also improve geolocalisation at a fine-grained level, and propose a learning to rank approach to re-rank geotagged tweets based on their geographical proximity. Finally, in Chapter 6 we explore the applicability of our tweet geolocalisation approaches and study the effectiveness of geolocated tweets in a real-world practical scenario – the traffic incident detection task.

Part II

Fine-Grained Geolocalisation of Tweets

Chapter 3

Enabling Fine-Grained Geolocalisation

3.1 Introduction

As introduced in Chapter 2, Section 2.2, geolocalisation has been mainly addressed in the literature by exploiting the similarity between the content of non-geotagged tweets and geotagged tweets, which are already available in the Twitter stream. The first research efforts achieved geolocalisation at a coarse-grained level of granularity (i.e., zip codes to cities or countries). Examples of such works are Hulden et al. (2015), Roller et al. (2012) and Wing and Baldridge (2011) were they represent areas by dividing the geographical space into predefined coarse-grained areas, and then concatenating (or aggregating) the texts of the tweets posted within that area into a single document. After the aggregation process, each area is represented as a bag-of-words vector extracted from the aggregated document. Finally, a matching function returns the most likely area as the predicted location by computing the content similarity of each area with respect to a given non-geotagged tweet.

On the other hand, more recent research attempted to achieve fine-grained geolocalisation (i.e., street or neighbourhood level) by adopting the approach mentioned above. To this end, the authors reduced the size of the predefined areas and represented them as zip code areas, in work by Kinsella et al. (2011), and as a grid of squared areas of size length 1 km, in work by Paraskevopoulos and Palpanas (2015). Moreover, as the matching function Kinsella et al. (2011) adopted a language model approach and Paraskevopoulos and Palpanas (2015) opted for TF-IDF weighting model. However, when adapting such approach to provide fine-grained locations with the overall performance of the geotagging system decreases compared to the performance at coarse-grained level.

In this chapter, we investigate the performance issues exhibited by existing approaches in the literature and propose a solution that enables the geolocalisation at fine-grained levels. Due to the morphology of Twitter documents (short texts limited to 140 characters (Teevan et al., 2011)) when aggregating the tweets into a single document, relevant information about discriminative words representative of fine-grained locations is lost, thus affecting the performance of geolocalisation at fine-grained levels of granularity. Therefore, The central hypothesis of this chapter is that by considering geotagged tweets individually we can preserve the evidence lost when adapting previous approaches at a fine-grained level, and thus we can improve the performance of fine-grained geolocalisation (see **Hypothesis 1** in Section 1.2). To this end, we propose a new strategy which avoids the aggregation of tweets, and utilises individual tweet documents, thus preserving evidence otherwise lost in the aggregation process.

3.1.1 Research Questions

The main research goals in this chapter are to understand how the approach of aggregating the tweets in an area is affecting the geotagging accuracy performance, evaluate our proposed solution to alleviate the problem and enable the fine-grained geolocalisation of tweets. We contextualise the work of this chapter in terms of the following research questions :

- **RQ-3.1:** Does consider geotagged tweets individually improve the performance of fine-grained geolocalisation?

- **RQ-3.2:** What is the effect of aggregating tweets within a predefined area on accuracy when geolocalising tweets at a fine-grained level?

In order to answer these research questions, we experiment to understand the behaviour of aggregated and individual approaches utilising state-of-the-art retrieval models.

The rest of the chapter is organised as follows. In Section 3.2 we describe our two approaches to modelling the fine-grained geolocalisation task. In Section 3.3

we describe our experimental setup. Finally, Section 3.4 presents the experimental results of our fine-grained geolocalisation methods. We conclude with a discussion of our main findings in Section 3.5.

3.2 Modelling Tweet Geolocalisation

Given a set of already available geotagged tweets and a non-geotagged tweet, we model tweet geolocalisation in two main components. First, we represent candidate locations using the text of the geotagged tweets belonging to that location. Second, we select the most likely location based on their similarity to the content of a given a non-geotagged tweet.

3.2.1 Representing Candidate Locations

We represent candidate locations in two ways. The first approach follows state-of-the-art to represent a location as a vector that aggregates the texts of the geotagged tweets posted within a predefined geographical area (**Aggregated**). We consider this approach as our baseline in our experiments in Section 3.3. In the second approach, we propose to represent a location as a vector that contains the text of an individual geotagged tweet (**Individual**).

3.2.1.1 Aggregated Approach

In this approach, we represent candidate locations as a set of predefined areas that are obtained by creating a grid that divides the geographical space of interest into squares or cells. The size of the squares defines the granularity of the grid. As mentioned in Section 3.1, this way of representing the geographical space is widely adopted in state-of-the-art approaches for tweet geolocalisation (Hulden et al., 2015; Kinsella et al., 2011; Paraskevopoulos and Palpanas, 2015). In order to work at fine-grained levels of granularity, we create predefined areas of size length 1 km, following the work by Paraskevopoulos and Palpanas (2015).

Next, we associate each of the geotagged tweets with its corresponding area based on its longitude and latitude coordinates. To represent a candidate location, we generate a bag-of-words vector by concatenating (or aggregating) the texts of the geotagged tweets associated with a given squared area.

3.2.1.2 Individual Approach

In our second approach, instead of dividing the geographical space into predefined areas we utilise longitude and latitude coordinates attached to the already available geotagged tweet as locations. Then, instead of aggregating the texts of geotagged tweets we treat each geotagged tweet individually. This way, candidate locations are represented as single documents containing the text of individual tweets. Then, a bag-of-words vector is created from each document.

3.2.2 Predicting a Geolocation

Once we have obtained the vectors of the candidate locations, we can then estimate the probability of a non-geotagged tweet being posted in a location by computing its content-based similarity to each vector. The most likely location is then selected as the predicted location. There are two ways of approaching this process: using IR techniques for ranking the locations Kinsella et al. (2011); Paraskevopoulos and Palpanas (2015), or as a classification task Hulden et al. (2015). In this thesis, we use a ranking approach for selecting the predicted location. We obtain the Top-N most likely candidate locations using a ranking function and retrieve the most likely area (Top-1) as the predicted location. We utilise several state-of-the-art retrieval models in our ranking function, which are introduced further in Section 3.3.3.

Note that, depending on the approach, the predicted location is returned as a longitude and latitude position representing: either the centroid of a squared area, which is returned in Aggregated approach or the location of a geotagged tweet, which is returned in Individual approach.

3.3 Experimental Setting

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

3.3.1 Data

Previous studies have shown that geotagged and non-geotagged data have the same characteristics (Han et al., 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) location of a tweet is known and focus on detecting their fine-grained geolocation[1]. Therefore, we experimented over a ground truth sample of English geotagged tweets.

Table 3.1: North-east (NE) and south-west (SW) longitude/latitude coordinates of the bounding boxes of the Chicago and New York datasets.

	Longitude/Latitude Coordinates	
	NE	*SW*
Chicago	-87.523661, 42.023131	-87.940267, 41.644335
New York	-73.700171, 40.917577	-74.259090, 40.477399

In Section 2.4.3, we describe other datasets from the literature that are available online. However, they were collected to evaluate coarse-grained geolocalisation methods using a wide spatial filter on the Twitter Stream, that covers global or country areas. Due to this filtering approach, they do not provide a representative sample of small geographical areas and, therefore, we collect our datasets for evaluating the fine-grained geolocalisation task. In total, we collect two datasets containing geotagged tweets located in two different cities from the Twitter Public stream[2]. The tweets were posted on March 2016 in Chicago and New York (USA) containing 131,757 and 153,540 geotagged tweets respectively.

We use a spatial filter to collect geotagged tweets posted within an area delimited by a bounding box that covers the metropolitan areas of the city. We create bounding boxes for Chicago and New York cities, which are defined with a pair of longitude/latitude coordinates that represents the north-east (NE) and south-west (SW) corners of the box (see Table 3.1).

The geographical distribution of geotagged tweets over the target cities is not uniform. Figure 3.1 shows the distributions of geotagged tweets for the Chicago and New York cities. We observe that some areas, such as the outlying districts of the cities, contains low-density of tweets and thus are underrepresented. Besides, other areas, such as the metropolitan area, contains high-density of tweets and

[1] The city-level location of tweets can be inferred by using approaches from previous works (Cheng et al., 2010; Kinsella et al., 2011; Schulz et al., 2013a).

[2] `https://dev.twitter.com/streaming/public`

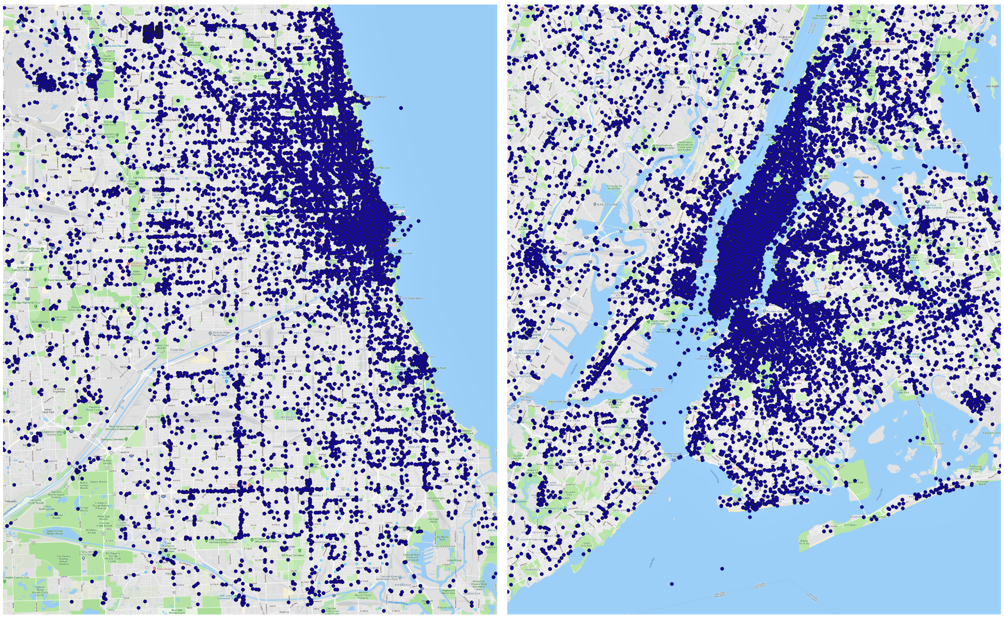

Figure 3.1: Geographical distribution of geotagged tweets in Chicago (left) and New York (right) during March 2016.

are overrepresented. It is important to note that, due to a bias towards the high-density areas, this variation in the geographical distribution may affect the inference.

3.3.1.1 Training, Testing and Validation Sets

To evaluate our approach, we divide each dataset into three subsets. We use 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 divide the last week data into validation and test sets to ensure that they have similar characteristics. Table 3.2 describes the distribution of tweets for the three datasets.

Table 3.2: Number of geotagged tweets distributed between training, validation and testing sets of the Chicago and New York datasets.

		Number of Geotagged Tweets		
Dataset	**Collection Time**	*Training*	*Validation*	*Testing*
Chicago	March 2016	99,418	16,061	16,278
New York	March 2016	111,577	20,886	21,077

3.3.2 Preprocessing and Indexing

As a preprocessing step, for each tweet, we remove punctuations, hyperlinks, stop-words, tokenise (1-gram) and apply Porter Stemmer (Porter, 1980). Moreover, we preserve retweets, usernames and hashtags as tokens in the dataset. The reason behind preserving retweets is that when a user retweets a content, the geolocation of the original tweets is not necessarily preserved.

Moreover, the similarity between a tweet and its retweet is high. Therefore we can assign the location of the original tweet to the retweet. Finally, we index every geotagged tweet in the training set using the Lucene platform[1].

3.3.3 Models

In this section, we describe the baseline models, as well as the different configurations of our approach (*Individual*) that we use in our experiments.

3.3.3.1 Aggregated

We consider the models that use the Aggregated approach for representing candidate locations, described in 3.2.1.1, as the baselines models for our experiments and comparison to our proposed approach, Individual. We implement two categories of approaches that use the aggregation of tweets that differ in the way they obtain the predicted geolocation for a given tweet, described in Section 3.2.2. First, we implement the work by Kinsella et al. (2011) and Paraskevopoulos and Palpanas (2015), that use a ranking approach. On the other hand, we adopt the work by Hulden et al. (2015), that uses a classification approach.

Moreover, the approaches by Kinsella et al. (2011) and Hulden et al. (2015) work at a coarse-grained level of granularity, therefore, to adapt them to the task of fine-grained geolocalisation, for each city mentioned in Section 3.3.1, we create a grid structure of squared areas with a side length of 1 km. For each of the areas, we concatenate the text of the tweets associated with that area into a document and index the document (see Section 3.3.1) which represents that area. After indexing the documents, for each non-geo-tagged tweets, we retrieve

[1] https://lucene.apache.org/

the most content-based similar document (Top-1) through a ranking function to follow (Paraskevopoulos and Palpanas, 2015) and (Kinsella et al., 2011).

On the other hand, we follow the work by Hulden et al. (2015). Following authors approach, we model each cell of the 1 km grid using Multinomial Naive Bayes (denoted by *NB*) and Kullback-Leibler divergence (denoted by *KL*) using words as features. We also report standard Naive Bayes and Kullback-Leiber versions using kernel density estimation, denoted as *NB+KDE* and *KL+KDE* respectively.

3.3.3.2 Individual

In this model, we implement the approach introduced in Section 3.2.1.2, where a single geotagged tweet represents each location. Thus, we index each tweet as a single document. We preprocess each tweet following the same steps explained in Section 3.3.1. After indexing the tweets, we obtain the Top-N content-based most similar geotagged tweets for each non-geotagged tweet using a ranking function. We experiment with the same five retrieval models utilised in Approach 1 in our ranking function. Finally, we return the longitude and latitude coordinates of the Top-1 tweet as the predicted location.

3.3.4 Ranking Functions

For the models described above, **Aggregated** and **Individual**, we experimented with the following retrieval models as the ranking function:

Vector Space Models (Dillon, 1983; Salton and Buckley, 1988; Salton et al., 1975).

- **TF-IDF** weighting as described in Section 2.2.2.2.
- **IDF** weighting as described in Section 2.2.2.2.

Probabilistic Models

- **LMD** (Zhai and Lafferty, 2017): Language Model with Dirichlet Smoothing.

- **DFR** (Amati and Van Rijsbergen, 2002): Divergence From Randomness Framework, as introduced in Section 2.2.2.5. We utilise different configurations of the framework as described later in Section 3.3.5.
- **BM25** (Robertson et al., 1995): As introduced in Section 2.2.2.3.

3.3.5 Parameter Tuning

We tune the parameters of the ranking function for *Aggregated* and *Individual* approaches to optimise the average error distance (see Section 3.3.6) utilising the validation sets for Chicago and New York described in Section 3.3.1. Note that TF-IDF and IDF are parameter free, thus we optimise parameters for BM25, LMD and DFR.

BM25: We experiment with a range of values for parameter k (0.0, 0.2, 0.5, 0.7, 1.0, 1.2, 1.5, 1.7, 2.0), and values for parameter b (0.0, 0.2, 0.5, 0.7, 1.0).

LMD: On the other hand, for LMD we experiment with values of μ (1, 5, 20, 50, 100, 500, 1000, 2500).

DFR: We test the following configurations of the DFR framework:

1. *InB2*: Inverse Document Frequent model with Bernoulli after-effect and normalisation 2.
2. *IneB2*: Inverse Expected Document Frequent model with Bernoulli after-effect and normalisation 2.
3. *IFB2*: Inverse Term Frequency model with Bernoulli after-effect and normalisation 2.
4. *InL2*: Inverse Document Frequency model with Laplace after-effect and normalisation 2.
5. *PL2*: Poisson model with Laplace after-effect and normalisation 2.

The final optimised parameters for *Aggregated* and *Individual* on our two datasets, Chicago and New York, are reported in Table 3.3.

Table 3.3: Optimised parameters for the ranking functions used in *Aggregated* and *Individual* approaches on our two datasets, Chicago and New York.

	Chicago			**New York**		
	LMD	**BM25**	**DFR**	**LMD**	**BM25**	**DFR**
Aggregated	μ=2500	$k = 1.2$ $b = 0.0$	InB2	μ=2500	$k = 0.5$, $b = 0.2$	InB2
Individual	μ=1	$k = 0.5$, $b = 0.7$	IFB2	μ=1	$k = 1.5$, $b = 0.5$	InB2

3.3.6 Evaluation Metrics

Following previous works in the literature (Flatow et al., 2015; Kinsella et al., 2011; Paraskevopoulos and Palpanas, 2015), to evaluate the effectiveness of the approaches over the tweets in the test set T_{test} the following metrics are reported:

Error distance (*km*): The fundamental measure for evaluating tweet geolocalisation approaches is the distance error $d(\hat{l}_i, l_i)$ between the predicted location $\hat{l}_i$ and the real coordinates l_i of the tweet in the test set $t_i \in T_{test}$. To this end, we use the Haversine distance (Robusto, 1957), which calculates distances on Earth, to compute the error. For this metric, lower values represent better performance.

As described in Section 3.3.3, the output of our models can be either a tweet or a squared area. When our prediction is a single tweet (*Individual* approach), we compute the distance between two coordinates; when our prediction is an area (*Aggregated* approach), the distance between the ground truth coordinate and the centroid of the area is calculated. Moreover, to describe the distribution of the error committed by the models we report the **average** and **median** error distance.

Accuracy@1km: In this thesis, we aim to geolocalise tweets at a fine-grained level (1 kilometre error or less). Therefore, we compute accuracy of the model by determining the fraction of predicted locations that lie within a radius of 1 kilometre from the real location. For this metric, higher values represents better performance. Accuracy@1km is formalised as follows:

$$Accuracy@1km = \frac{|\{t_i \in GeoTweets \mid d(\hat{l}_i, l_i) \leq 1km\}|}{|GeoTweets|} \tag{3.1}$$

where $GeoTweets$ is the set of tweets in the test set for which the model finds a geolocation, $\hat{l_i}$ is the predicted location and l_i is the real location of the test tweet $t_i \in T_{test}$.

Coverage: We consider Coverage as the fraction of tweets in the test set T_{test} from which the model finds a geolocation regardless of the distance error. Coverage can be formalised as follows:

$$Coverage = \frac{|GeoTweets|}{|T_{test}|} \tag{3.2}$$

3.4 Results and Discussions

This Section presents our evaluation results and discusses the effectiveness of the proposed approaches on fine-grained geolocalisation. In particular, Tables 3.4 and 3.5 provide experimental results on the Chicago and New York datasets respectively. We report results for the baseline models that performs an aggregation of tweets (*Aggregated*) and our proposed approach that uses individual tweets (*Individual*) compared to each other. Moreover, we present results when suing different functions for selecting the predicted location, described in Section 3.3.3. In each table, we report the metrics described in Section 3.3.6: average error distance (*AED*), median error distance (*MED*), as well as accuracy at 1 kilometre (*Acc@1km*).

Next, in each table, a paired t-test is used to assess if the difference in effectiveness is statistically significant, and are denoted by * when a result is significantly different ($p<0.01$) different to the best baseline (*Aggregated using LMD*). Finally, the best performing fine-grained geolocalisation approach for each measure is highlighted in bold.

In the following subsections, we address the research questions formulated in Section 3.1.1. Particularly, Subsection 3.4.1 tackles **RQ-3.1** and discusses the effectiveness of the *Aggregated* and *Individual* approaches for representing candidate location on fine-grained geolicalisation; Subsection 3.4.2 addresses **RQ-3.2** and derives conclusions on why the aggregation of tweets underperforms with respect to treating tweets individually.

3.4.1 Aggregated Versus Individual

Comparing the performance across the two datasets presented in Tables 3.4 and Table 3.5, we observe that the *Individual* approach generally outperforms *Aggregated* in all the metrics reported, when using any of the prediction functions. For instance, in the Chicago dataset *Individual* models significantly (statistically) improve performance with respect to the best performing baseline (*Aggregated* using LMD); accuracy is increased from 46.97% to 55.20% using TF-IDF, and average error distance (*AED*) is reduced from 6.162 km to 4.694 km using IDF. Additionally, median error distance (*MED*) is substantially reduced in all cases, which explains the increment on accuracy as a higher number of tweets are predicted at fine-grained level (i.e., 1 km distance) using *Individual*.

Table 3.4: Evaluation results for the Chicago dataset. The table presents the Average Error Distance in kilometres (AED), Median Error Distance in kilometres (MED), Accuracy at 1 kilometre (Acc@1km) and Coverage. Significant (statistically) differences with respect to the best Baseline (Aggregated using BM25) are denoted by $*$ ($p<0.01$).

Chicago Dataset					
Model	**Function**	**AED(km)↓**	**MED(km)↓**	**Acc@1km↑**	**Coverage↑**
Aggregated	*NB+KDE*	7.340	2.445	29.63%	**100.00%**
Aggregated	*KL+KDE*	7.501	2.828	25.24%	**100.00%**
Aggregated	*NB*	6.233	0.817	50.79%	**100.00%**
Aggregated	*KL*	7.051	1.351	48.15%	**100.00%**
Aggregated	*IDF*	13.439	13.705	14.02%	99.40%
Aggregated	*TF-IDF*	8.040	3.402	41.82%	99.40%
Aggregated	*DFR*	6.250	1.333	47.06%	99.40%
Aggregated	*LMD*	5.998	1.194	47.64%	99.40%
Aggregated	*BM25*	4.806	0.906	50.67%	99.40%
Individual	*IDF*	**4.693**∗	0.100∗	55.13%∗	99.40%
Individual	*TF-IDF*	4.714∗	**0.080**∗	**55.20%**∗	99.40%
Individual	*DFR*	4.802∗	0.138∗	54.58%∗	99.40%
Individual	*LMD*	4.853∗	0.181∗	54.10%∗	99.40%
Individual	*BM25*	4.923∗	0.465∗	52.74%∗	99.40%

Lastly, we discuss the performance of the ranking functions against the classification approached for selecting the most likely location, described in Section

Table 3.5: Evaluation results for the New York dataset. The table presents the Average Error Distance in kilometres (AED), Median Error Distance in kilometres (MED), Accuracy at 1 kilometre (Acc@1km) and Coverage. Significant (statistically) differences with respect to the best Baseline (Aggregated using BM25) denoted by $*$ (p<0.01).

New York Dataset					
Model	**Function**	**AED(km)↓**	**MED(km)↓**	**Acc@1km↑**	**Coverage↑**
Aggregated	*NB+KDE*	6.627	2.595	23.62%	**100.00%**
Aggregated	*KL+KDE*	6.628	2.703	20.03%	**100.00%**
Aggregated	*NB*	6.318	1.951	43.67%	**100.00%**
Aggregated	*KL*	7.119	2.497	41.54%	**100.00%**
Aggregated	*IDF*	12.536	11.842	13.82%	99.98%
Aggregated	*TF-IDF*	7.308	2.620	41.08%	99.98%
Aggregated	*DFR*	6.499	2.415	42.21%	99.98%
Aggregated	*LMD*	6.873	2.873	42.03%	99.98%
Aggregated	*BM25*	**4.862**	1.547	45.40%	99.98%
Individual	*IDF*	5.041*	1.325*	47.98%*	99.98%*
Individual	*TF-IDF*	4.972*	**1.251***	**48.46%***	99.98%*
Individual	*DFR*	5.826*	2.769*	39.79%*	99.98%*
Individual	*LMD*	5.118*	1.377*	47.77%*	99.98%*
Individual	*BM25*	5.642*	1.936*	44.23%*	99.98%*

3.2.1.1. We observe that using a ranking approach performs better than using classification in most of the cases in terms of average error distance, independently of using *Aggregated* or *Individual*. However, using *Aggregated* with IDF and TF-IDF exhibits worst performance, which suggests that document frequency information is not that informative when aggregation the tweets, as we will discuss later in Section 3.4.2. Moreover, it is interesting to note that classification approaches provides 100% *Coverage* compared to 99.40% and 99.98% of the ranking approaches in Chicago and New York, respectively. This difference in *Coverage* can be explained because classification approaches provide inference for all the tweets in the test set, whereas ranking approaches are not capable of finding similar geotagged tweets for some test tweets.

These results support the hypothesis **RQ-3.1** introduced in Section 3.1, which proposes using individual tweets instead of aggregated tweets within an area would result in better performance for fine-grained geolocalisation of tweets.

3.4.1.1 The BM25 case

Previously, we concluded before that treating tweets individually using our *Individual* approach is the best performing strategy for fine-grained geolocalisation when using any of utilised retrieval models, however, we observe an interesting behaviour when comparing *BM25* in both *Individual* and *Aggregated* approaches in Tables 3.4 and 3.5. Despite *Individual* is still the best performing, we note there is not a high difference in the metrics. In particular, in the Chicago dataset, we obtain an average error distance (*AED*) of 4.806 km using *Aggregated* approach and 4.923 km using *Individual* approach, which represents a difference of 0.117 km.

The reason behind the similar performance of *Individual* and *Aggregated* using *BM25* can be explained by the inherent characteristics of the *BM25* (Robertson et al., 1995) model. The similarity of a document d to the query q is formalised as follows:

$$\text{BM25}(q,d) = \sum_{t \in q} \log\left(\frac{N - df_i + 0.5}{df_i + 0.5}\right) \cdot \frac{(k_1 + 1) \cdot tf_i}{k1((1-b) + b\frac{dl}{\text{avgdl}}) + tf_i} \tag{3.3}$$

where tf_i represents the frequency of the term in the document, df_i is the document frequency of the term, and document length is represented as dl. Document length is normalised by dividing by the average document length of the collection $avgdl$. The model is tuned using two parameters; $k1$ and b. By adjusting $k1$ we control the influence of term frequency tf_i in the final score, whereas adjusting b varies the influence of document length normalisation $\frac{dl}{\text{avgdl}}$.

In previous research, Ferguson et al. (2012) demonstrated that when $k1$ and b parameters are close to zero, the performance of retrieval over Twitter documents improves. This is due to the nature of tweets, which are short documents, and the evidence encoded in terms of document length, and term frequency is lower than longer documents (i.e., web documents). In Section 3.3.5 we the parameters $k1$ and b are adjusted to the characteristics of short documents in the *Individual* approach and long documents in the *Aggregated*, and therefore leads to similar performance on fine-grained geolocalisation. This behaviour suggests that document frequency provides the strongest evidence for fine-grained geolocalisation in contrast to term frequency or document length. In the next Subsection 3.4.2 will address **RQ-3.2** and derive an explanation of the effects that aggregating tweets have on the evidence in terms of document frequency, which is affecting geolocalisation at a fine-grained level.

3.4.2 Effect of Tweet Aggregation

In order to show the importance of document frequency for fine-grained geolocalisation, we compute the distribution of the error distance over the similarity scores given by the retrieval model to the document that represents the predicted location (Top-1). Figure 3.2 shows the distribution of error distance for all the models. We observe that generally, as the similarity increases, the error distance of the predicted location decreases. However, *Individual* models show the lowest error distances across all the values of similarity score. As indicated in the figure, the best performing configuration is *Individual_IDF*. This observation is consistent with the behaviour described before in Subsection 3.4.1.1 which shows that the importance of document frequency over term frequency and document length is higher when treating with short documents (*Individual*). On the other hand, when dealing with long documents (*Aggregated*), IDF performs the worst and models that utilise in term frequency, and document length (*LMD, DFR and BM25 optimised*) perform better, but still underperforms *Individual* models.

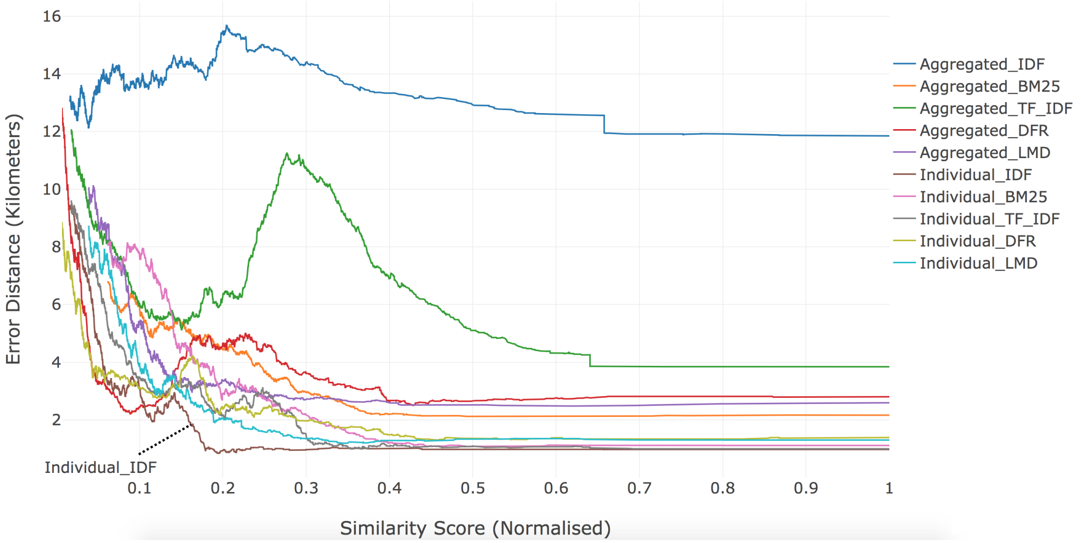

Figure 3.2: Distribution of error distance (y-axis) against similarity score (x-axis) for the Chicago dataset.

Additionally, we statistically compare the error distance against the similarity score by computing correlation coefficients in terms of K.Tau, SP.Rho and Pearson. Table 3.6 presents the correlation coefficients for all the geolocalisation models. We observe that the best coefficient is achieved by *Individual_IDF*, which shows a significant negative Pearson correlation of -0.350, K.Tau of -0.362 and SP.Rho of -0.504. On the contrary, we note that *Aggregated_IDF* shows to be the model with the lowest correlation. This suggests the document frequency information is not discriminative enough when tweets are aggregated, but becomes the most important evidence when tweets are treated individually.

In order to address our research question **RQ-3.2** described in Section 3.1.1, we now present a theoretical explanation of the effects of aggregating tweets on fine-grained geolocalisation, supported by the results obtained before. Based on the results presented in Table 3.6 and Figure 3.2, we postulate that discriminative information about the query terms that manifests in the way of document frequency when using individual tweets, is then transferred into term frequency

Table 3.6: Correlations between error distance and retrieval score. Significant (statistically) differences are denoted by $*$ (p<0.01).

Model	**K.Tau**	**SP.Rho**	**Pearson**
Aggregated_IDF	0.028*	0.041*	-0.024*
Aggregated_TF-IDF	-0.127*	-0.190*	-0.125*
Aggregated_DFR	-0.253*	-0.380*	-0.214*
Aggregated_LMD	-0.250*	-0.361*	-0.241*
Aggregated_BM25	-0.128*	-0.189*	-0.175*
Individual_IDF	**-0.362***	**-0.504***	**-0.350***
Individual_TF-IDF	-0.361*	-0.501*	-0.348*
Individual_DFR	-0.293*	-0.406*	-0.258*
Individual_LMD	-0.300*	-0.415*	-0.267*
Individual_BM25	-0.297*	-0.412*	-0.267*

information when tweets are aggregated into a single document. Therefore, retrieval models that rely on document frequency capture strong evidence for geolocalisation and perform the best using the *Individual* approach, whereas retrieval models that rely on term frequency still capture more of that evidence when using the *Aggregated* approach. Nevertheless, the performance is still worst compared to the *Individual* approach, which suggests that some evidence is lost in the aggregation or retrieval models are not capable of capture such evidence.

The results presented in this section, related to the importance of document frequency in fine-grained geolocalisation, are in line with previous findings in microblog search, introduced in Section 2.3. First, Ferguson et al. (2012) and Naveed et al. (2011) observed that term-frequency information have little impact on retrieval effectiveness and document length normalisation have a negative effect, as we also observed previously in Section 3.4.1.1. Finally, in Rodriguez Perez (2018); Rodriguez Perez and Jose (2015) the authors performed an exhaustive investigation of the problem of retrieval models in microblog search. Their work confirmed previous findings (Ferguson et al., 2012; Naveed et al., 2011) and, in line to our work, they observed that models relying on document frequency performed significantly better than others.

3.5 Chapter Summary

Existing geolocalisation models in the literature utilised the content of already available geotagged tweets to represent locations. Next, these models ranked candidate locations based on their similarity to a given non-geotagged tweet using a ranking function and returning as a prediction the most similar location (Top-1). The first attempts to geolocalise tweets divided the geographical area into predefined coarse-grained areas (i.e., country or city level), and represented each area as a single document containing the texts of the geotagged tweets belonging to that area. More recent works adapted the existing approach to work at a fine-grained level (i.e., squared areas of length size 1 km), which resulted in a decrease in performance compared to coarse-grained predictions.

In this chapter, we hypothesised that by aggregating the texts of the geotagged tweets to represent a location, important information about discriminative words that are representative of a fine-grained location is lost, thus affecting the performance of the geolocalisation. Based on this assumption, we postulated that by representing locations as a single vector containing the text of individual geotagged tweets, the performance of fine-grained geolocalisation would improve.

The experiments in this chapter were focused to answer **RQ-3.1** and **RQ-3.2** introduced in Section 3.1.1. To conduct our experiments, we collected two datasets of English geotagged tweets located in two major cities in USA, Chicago and New York. Next, we proposed a new approach that treats tweets individually, named *Individual*, and compared against the state-of-the-art approach that aggregates the texts within predefined areas as the baseline, named *Aggregated*. As the ranking function, we experimented with IR retrieval models including Vector Space Models using TF-IDF and IDF weighting, and probabilistic models such as BM25, DFR Framework and Language Model with Dirichlet smoothing (LMD). We optimised the parameter of each retrieval model for both approaches, *Individual* and *Aggregated*.

Our first experimental results showed that representing locations as individual tweets significantly (statistically) outperforms state-of-the-art strategies of aggregating the tweets of an area (see Tables 3.4 and 3.5), which address **RQ-3.1** and support our hypothesis that treating tweets individually will perform better for fine-grained geolocalisation.

Secondly, we addressed **RQ-3.2** by analysing the performance of the different retrieval models when used in *Individual* and *Aggregated* approaches. We observed that BM25 showed similar performance in both approaches (see Section 3.4.1.1). This is because the parameters of the BM25, $k1$ and b, controls the influence of term frequency and document length information and were adjusted to work with short documents (*Individual*) and long documents (*Aggregated*). Based on the BM25 formulation, this suggested us that evidence in terms of document frequency is a strong signal for fine-grained geolocalisation.

Inspired by the previous observation, we then addressed **RQ-3.2** and derived a theoretical explanation of the effects that aggregating tweets have on the evidence in terms of document frequency, which is affecting geolocalisation at a fine-grained level. To this end, we computed the distribution of error distance committed by *Individual* and *Aggregated* approaches against the similarity score given by the different retrieval models utilised (see Section 3.4.2). We identified from Table 3.6 and Figure 3.2 that retrieval models that relies on term frequency and document length (BM25, DFR and LMD) performed the worst when using *Individual*, and performed the best when using *Aggregated*. On the other hand, we noted that retrieval models that rely on document frequency (IDF and TF-IDF) performed the best when using the *Individual* approach, and performed the worst when using the *Aggregated* approach. This suggested us that document frequency information is not discriminative enough when tweets are aggregated, but becomes the most important evidence then tweets are treated individually. Additionally, the fact the models relying on term frequency performed the best when aggregating the tweets, suggested us that the evidence lost as document frequency information is transformed into term frequency information and still captured by such models.

In this chapter, we demonstrated that our proposed approach of treating tweets individually *Individual* is the best strategy for the fine-grained geolocalisation task. In particular, our experimental results showed that evidence in the form of document frequency information is the most discriminative. For this reason, the IDF weighting model showed to be the best ranking function. Additionally, we provided a theoretical explanation of why aggregating the tweets is not convenient for fine-grained geolocalisation. However, the average error distance achieved by our best performing model (4.693 km by *Individual* using IDF) is still insufficient for tasks that require fine-grained geolocation levels defined as the objective of this thesis work (i.e., located at 1 km or less).

So far, our approach returns the most similar location (Top-1) as the returned prediction. However, having a Top-N ranking of individual geotagged tweets as evidence allows us to explore ways to improve the performance of fine-grained geolocalisation further. In Chapter 4, we propose a new approach for fine-grained geolocalisation to increase the number of tweets geolocalised at 1 km error distance.

Chapter 4

Majority Voting For Fine-Grained Geolocalisation

4.1 Introduction

In Chapter 3 we enabled state-of-the-art methods to work at fine-grained levels by proposing a new approach that represents locations as documents generated from individual geotagged tweets, instead of an aggregation of them. Then, a ranked list of the most likely candidate locations is retrieved based on the similarity of a given non-geotagged tweet to the documents. Thus, the most likely candidate location (Top-1) is returned as the predicted location. However, the average error distance (AED) of the predictions returned by such approach is still not sufficient to reliably enable tasks that require high accurate geolocated data, such as the traffic incident detection we will explore in 6 – the best average error distance is 4.693 km (Chicago) which represents a confidence area of 69.19 km^2. In this thesis, we aim to reduce the average error to 1 km which represents a confidence area of 3.14 km^2.

The main drawback of the approach derived in Chapter 3 is that only the similarity dimension is contemplated to perform a prediction of the geographical location of a tweet. Thus, the approach returns always the location of the most similar geotagged tweet (Top-1). However, the similarity between two tweets is not always sufficient evidence of their geographical location, and thus it can be challenging to return a prediction in such cases. For example, two tweets about a topic that is not related to any geolocation (i.e., a new album released by a famous singer) are highly similar, but they are not necessarily be posted in the same location. Thus the predictability of the geolocation of such tweets is low. In contrast, two tweets about an event occurring at specific geolocation (i.e., a traffic incident) are likely to be generated within the same area or include the location name in the text. Thus the predictability of their location is high.

In this chapter, we hypothesise that the predictability of the geolocation of tweets at a fine-grained level is given by the correlation between their similarity and geographical distance to other finely-grained geotagged tweets (see **Hypothesis 2** in Section 1.2). We postulate that in some cases the similarity of the tweets does not always correlate with geographical distance. Therefore, there may not be sufficient evidence to return a fine-grained prediction in such cases. We believe that by identifying such cases, we can increase the number of predictions at a fine-grained level.

In Figure 4.1, we illustrate the correlation between the content similarity and the geographical distance of a tweet to other geotagged tweets. Red areas represent high correlation whereas blue areas represent low correlation. In this figure, there are four areas of interest as we observe the corners. Firstly, the top left corner represents the area of high similarity and low geographical distance. This area is the most correlated with the hypothesis which links distance with content similarity. Secondly, the top right corner represents an area of high similarity yet high geographical distance. Thirdly, the bottom left stands for an area of low similarity and low geographical distance. This area is not in line with the hypothesis mentioned above, yet it is of interest as potentially any prediction in this area should produce good results. Finally, the bottom right corner describes an area of low similarity and high geographical distance. Consequently, the area through the middle connecting the top left and bottom right corners embodies the hypothesis linking similarity to geographical distance.

On the other hand, in Figure 4.2 we present the utility area for fine-grained predictions that is enclosed at the left of the line in the graph. The closer the line is to the left, the lower is the geographical distance and, therefore, the better the predictions. Note that this is happening regardless of the level of content similarity. In line with this assumption, by analysing how dispersed in space are the Top-N most similar geotagged tweets in the rank provides, we can obtain valuable evidence of the geographical distance. Thus, we can identify the predictions that fall at the left of the line and reduce the average error distance (i.e., 1 km) for better fine-grained geolocalisation.

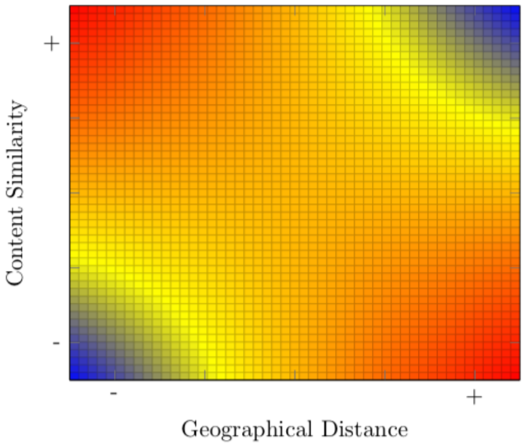

Figure 4.1: The figure presents the correlation between the content similarity and the geographical distance of a tweet to a set of Top-N geotagged tweets. Red areas represent high correlation whereas blue areas represent low correlation.

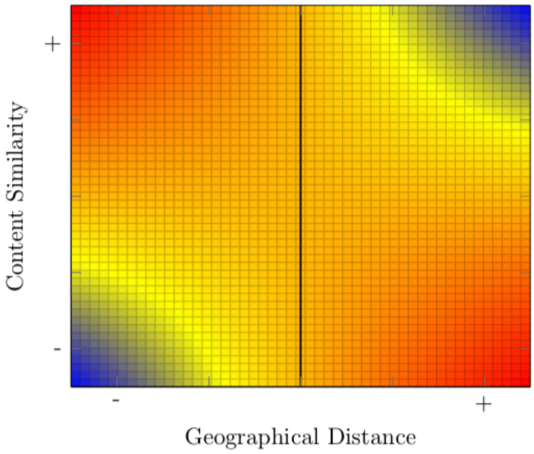

Figure 4.2: Regardless of the content similarity, the space at the left of the line represents the utility area for fine-grained geolocalisation — the closer the line to the left, the lower the geographical distance and the better the predictions.

In this chapter, we aim to explore the geographical evidence encoded within the Top-N most similar geotagged tweets in order to obtain more reliable predictions. To combine evidence from the Top-N elements, we propose to model fine-grained geolocalisation as a voting process, where each candidate location is represented as a set of geotagged tweets. Using a ranked list of retrieved geotagged tweets for a non-geotagged tweet, we propose to adopt a majority voting algorithm to estimate the geographical location by collecting the geolocation votes of the geotagged tweets in the rank. In the case that the voting process finds a location with a majority of the votes, it is indicative of low geographical distance, regardless of the content similarity, and we consider that there is sufficient evidence for a fine-grained prediction.

Additionally, we weighted the majority voting algorithm to alleviate the restrictive power of the voting process. The weight of each vote is calculated based on the credibility of the user of the geotagged tweet and the degree of content similarity 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.

4.1.1 Research Questions

Based on previous assumptions introduced before, in this chapter we aim to address the following research questions:

- **RQ-4.1:** Can we obtain fine-grained predictions based on the geographical evidence between the Top-N most similar geotagged tweets?
- **RQ-4.2:** What is the percentage of tweets we can predict at a fine-grained level?

The rest of the chapter is organised as follows. In Section 4.3 we describe our majority voting model for fine-grained geolocalisation of non-geotagged tweets. Section 4.4 presents our experimental setup, followed by results and discussing in Section 4.5. Finally, we provide concluding remarks and details of contributions in Section 4.6.

4.2 Related Work

Obtaining The Most Fine-Grained Predictions Previous approaches in the literature for fine-grained geolocalisation approaches also considered to return prediction only if there is sufficient evidence. For example, Flatow et al.

(2015) first identified geospecific n-grams in the dataset by applying a clustering approach using a Gaussian Kernel. The approach creates an ellipse that covers the locations in which the n-gram appears is there are clustered in space. Then, if a given non-geotagged tweet contains any of the geospecific n-grams, the centre of the generated ellipse is returned as a prediction. In contrast, the authors considered that tweets that do not contain any of the geospecific n-grams are not predictable. Another example of such works is the approach by Paraskevopoulos and Palpanas (2015), which reports their metrics (precision) based on the number of tweets in the test set their approach managed to geolocalise. This number corresponds to coverage but is not reported by the authors.

Majority Voting The majority voting algorithm is a well known, fast and effective strategy widely adopted for prediction and re-ranking tasks (Chiang et al., 2012; Mosbah and 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. 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 et al., 2007). In this work, we have decided to apply a voting approach due to its simplicity and effectiveness.

Credibility of Twitter Users Some works have attempted to measure the veracity/credibility of the information derived from social media (Marshall et al., 2016; Wang et al., 2016; Zhang et al., 2016b), and specifically for event detection and disaster and emergency management (Castillo et al., 2011; McCreadie et al., 2016). For example, McCreadie et al. (2016) considered the idea of assigning a credibility score to measure the veracity of a tweet in the context of a 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.

In this chapter, we define credibility differently to previous work. We aim to assign a score to Twitter users that provide a measure of their trustworthiness/credibility for fine-grained geolocalisation. To do that, we analyse their past activity in Twitter and calculate how usually these users post similar content related to other tweets in the same geolocation. The procedure to compute this score is detailed later in Section 1. Moreover, in contrast to McCreadie et al. (2016), we incorporate this score as a weight for each vote in our adopted majority voting approach.

4.3 Voting For Fine-Grained Geolocalisation

Our proposed approach consists of three stages. First, following previous approaches in Chapter 3 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. Second, we obtain the Top-N content-based most similar geotagged tweets to each non-geotagged tweet using a retrieval model (see Section 3.3.4). For the ranking task, we follow the *Individual* approach proposed in Section 3.3.3, that considers geotagged tweets as individual documents.

Finally, we combine the evidence gathered from the Top-N geotagged tweets mentioned above by adopting a weighted majority voting algorithm, which we introduce in the next Section.

4.3.1 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 and Moore, 1991; Chiang et al., 2012; Littlestone and Warmuth, 1992; Mosbah and 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 the majority voting as follows:

$$Location(t_{ng}) = \underset{l_j \in L}{\arg\max} \left(\sum_{i=1}^{N} Vote(t_i^{l_i}, l_j) \right) \tag{4.1}$$

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 ranking. Then, a vote is given to the location l_j by the tweet t_i as follows:

$$Vote(t_i^{l_i}, l_j) = \begin{cases} 1 & t_i^{l_i} = l_j \\ 0 & t_i^{l_i} \neq l_j \end{cases} \tag{4.2}$$

4.3.2 Weighted Majority Voting

In addition to Equation 4.1, were the votes are considered equally, we consider a weighted version of the majority voting formalised as follows:

$$Location(t_{ng}) = \arg\max_{l_j \in L} \left(\sum_{i=1}^{N} W_{t_i}(\alpha, t_{ng}) * Vote(t_i^{l_i}, l_j) \right) \tag{4.3}$$

were the vote from tweet t_i is weighted by:

$$W_{t_i}(\alpha, t_{ng}) = \alpha \cdot Credibility(u_i) + (1 - \alpha) \cdot Sim(t_i, t_{ng}) \tag{4.4}$$

where $\alpha \in [0, 1]$, and $Credibility(u_i)$ is the credibility of user u_i that posted the tweet t_i (see Section 4.3.2.1). $Sim(t_i, t_{ng})$ 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 4.3.2.2). Finally, the location l_j that obtains the highest number of weighted votes is returned as the final predicted geolocation for a given non-geotagged tweet.

We chose to use a linear combination as our weighting function (Equation 4.4) in order to study the effectiveness of each of the components ($Credibility(u_i)$ and $Sim(t_i, t_{ng})$) together and separately. Therefore, when using $\alpha = 1$ only $Credibility(u_i)$ is considered, whereas $Sim(t_i, t_{ng})$ is only considered when $\alpha = 0$. Likewise, when $\alpha = 0.5$ both components are considered equally. Lastly, the functions requires to normalise the values of the $Sim(t_i, t_{ng})$ component between 0 and 1 to be equivalent to the values of the $Credibility(u_i)$ component.

4.3.2.1 Extracting Credibility from a Tweet's User

We believe that some Twitter users tend to describe, more than others, the events occurring in the geographical locations they visit. This means that the geotagged tweets posted by such users are a valuable source of information for fine-grained geolocalisation. We aim to exploit this by computing, for each user, a score based on their posting activity. Finally, we utilise this score to weight the vote of a tweet in our adapted majority voting algorithm, as discussed above in Section 4.3.2. As discussed in Section 4.2, our concept of credibility differs from previous literature (Castillo et al., 2011; McCreadie et al., 2016), which aimed to measure the veracity of the information encoded in a tweet instead of the trustworthiness of the user.

To obtain the credibility score, we use the training and validation sets introduced in Section 3.3.1 of Chapter 3. The procedure to compute the score is detailed in Algorithm 1 and works as follows. For each user u_i in the training set T, we compute the credibility score as follows. First, for each tweet in the validation set (t_{v_i}) we obtain the Top-N most similar geotagged tweets (Top) from the training set (T) using a ranking approach. We collect the tweets (t_{u_i}) in the Top-N that were generated by the user u_i, along with their corresponding t_{v_i}, into a set named TN. After all the tweets in the validation set are processed, 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.

Algorithm 1: Computes the credibility score for a user u_i

```
Credibility ($u_i, N$)
    Data: Validation set $V$; Training set $T$
    Inputs : A user $u_i \in T$; Values of $N$ for the Top-N ranking obtained
             by the ranking function $rank$.
    Output: The credibility score for user $u_i$.

    $TN \leftarrow \emptyset$;
    foreach $t_{vi} \in V$ do
        $Top \leftarrow rank(t_{vi}, T, N)$;
        $TN \leftarrow \{(t_{u_i}, t_{v_i}) \in Top\}$
    end
    $C = \frac{|\{t_{ui} \in TN \mid distance(t_{ui}, t_{vi}) \leq 1km\}|}{|TN|}$;
    return $C$
```

Figure 4.3 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 3.3.1 of Chapter 3). As can be observed, an important chunk of the user population exhibit a low ratio (≤ 0.01). We consider that this set of users are less likely to post valuable information for geolocalisation, the votes of their tweets will be less contributive. 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. This is the set of users that are most likely to post valuable information, and their votes will be more discriminative. We observe similar patterns in all the cities considered in Section 3.3.1 of Chapter 3.

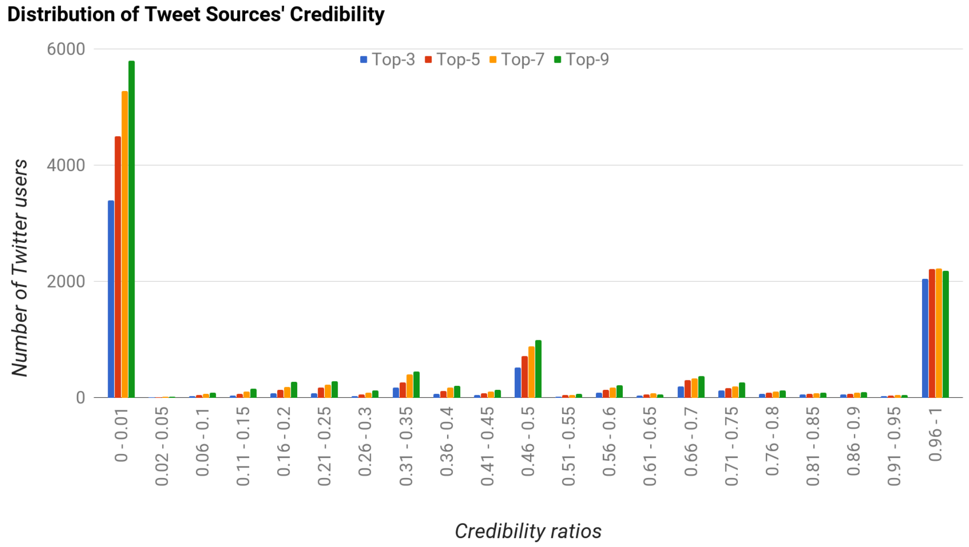

Figure 4.3: Distribution of Tweet Users' Credibility. The Figure presents the number of Twitter users (y-axis) distributed over different values of credibility ratios (x-axis).

4.3.2.2 Similarity Score and Tweet Geolocation

Previous research (Grabovitch-Zuyev et al., 2007; Hong et al., 2012a) has shown the correlation between the content of the tweets and their geographical location. This is because highly similar tweets are often 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*) and both contain two informative terms: *birthday* and *7th avenue*. Therefore, they will be associated with a high similarity score. Assuming there is a significant number of birthdays parties occurring in different areas, then it is very likely that both tweets were posted in the same geographical locations.

However, we can observe 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 in a nightclub at 7th avenue*", their similarity score will be lower as both tweets contain only 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.

The intuition behind is that the vote given by a high similar geotagged tweets contribute more in order to discriminate between locations. To this end, we introduce the similarity score $Sim(t_i, t_{ng})$ in Equation 4.4 in Section 4.3.2.

The contribution of the similarity component is adjusted by the value of an α parameter. In particular, the lower the value of α the higher the contribution of the content-based similarity score to the total weighting of each tweet vote.

4.4 Experimental Setting

In this section, we describe the experimental setup that supports the evaluation of our proposed approach for fine-grained geolocalisation tweets. We utilise the same experimental settings described previously in Chapter 3:

- We experiment over the two datasets of geotagged tweets described in Section 3.3.1 (Chicago and New York).
- We preprocess and index each geotagged tweet following Section 3.3.2.
- We utilise the retrieval models in Section 3.3.4 as ranking functions for the models detailed in Section 4.4.1.
- We report the metrics described in Section 3.3.6.

4.4.1 Models

In this section, we describe the baseline models, as well as the different configurations of our approach utilised in our experiments.

4.4.1.1 Baseline Models

We compare the performance of our majority voting model with the *Aggregated* and *Individual* approaches explored before in Chapter 3, which perform fine-grained geolocalisation by always returning the most similar document to a given non-geotagged tweet. The detailed implementations of the baselines are described in Section 3.3.3. We select the best performing configurations obtained in Tables 3.4 and 3.5 for the *Chicago* and *New York* datasets respectively.

4.4.1.2 Majority Voting Model

We implement our proposed approach explained in Section 4.3 (denoted by "WMV"). We use the same squared areas of the fine-grained grid defined for the baseline models. However, in WMV model, each of these defined squared areas is represented as multiple bag-of-word vectors where each vector represents a single geotagged tweet associated with that area. By doing this, we index each tweet as a single document for the retrieval task. We preprocess all tweets following the same step explained in section 3.3.2. After indexing the tweets, we perform a retrieval task to obtain the Top-N content-based most similar geotagged tweets for each non-geotagged tweet using the *Individual* approach proposed in Chapter 3, configured to use the *IDF* weighting model. Finally, we use a majority voting algorithm to return the final predicted location as the predefined area that obtains the majority of the votes. We build two majority voting models according to the way of weighting the votes:

WMV: We apply our weighted majority voting algorithm on top of the retrieval task, as described in Equation 4.3. The weight of each vote is given by Equation 4.4. In our experimental evaluation we considered the Top-N content-based most similar tweets obtained from the retrieval task with values of $N \in \{3, 5, 7, 9, ..., 49\}$, and different values of α (0.0, 0.25, 0.50, 0.75, 1.0) for the weighting function. The components in the weighting function (See equation 4.4) are normalised using min-max normalisation.

MV: We apply the majority voting version that does not weight the votes, as described in Equation 4.1. In our experimental evaluation, we considered the Top-N content-based most similar tweets obtained from the retrieval task with $N \in \{3, 5, 7, 9, ..., 49\}$.

4.5 Experimental Results

To assess the proposed weighted majority voting approach for fine-grained geolocalisation, we evaluate the models described in Section 4.4.1 using different values of N for the Top-N ranked geotagged tweets that are fed into the majority voting algorithm. Moreover, we provide results varying the values of α, that controls the influence of the similarity and user credibility components incorporated in our weighting function (See Equation 4.4).

In particular, Tables 4.1 and Table 4.2 provide experimental result on the *Chicago* and *New York* datasets respectively for each of the different settings for the weighted majority voting models (*WMV@Top-N* and α values). The tables also present, as a remainder, results of the best performing baseline approaches explored in Chapter 3 (*Aggregated* and *Individual*) that always return the Top-1 tweet as the predicted location. In each table, we report the following metrics (see Section 3.3.6): average error distance (*AED*), median error distance (*MED*), accuracy (*Acc@1km*), and coverage (*Coverage*). Lastly, for each measure and geolocalisation model setting, the best performing approach in each column is highlighted in bold. Due to the differences in *Coverage* obtained by different configurations of MV and WMV models, the models returns a prediction for different subsets of the tweets in the test set. For this reason, we do not compare (statistically) these results against our baselines.

In the following sections, we address the different research questions relating to the experimental results in the tables. Section 4.5.1 analyses the performance and effectiveness of the weighted majority voting approach using different values of N and α; Section 4.5.2 discusses the contribution to geolocalisation of each of the components in the weighting function formalised in Equation 4.4; Finally, we provide concluding remarks in Section 4.6.

4.5.1 Performance of Fine-Grained Geolocalisation

We observe in Tables 4.1 and 4.2 that our weighted majority voting approach (WMV@Top-N) obtained better predictions than the baselines 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 by 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 average error distance (*AED*), higher accuracy (*Acc@1km*), but lower coverage (*Coverage*). This suggest that our majority voting approach is capable of identifying fine-grained predictions, according to Figure 4.1 and 4.2, which address the first research question (**RQ-4.1**) described in Section 4.1.1. This observation is in line with the hypothesis, introduced in Section 4.1, that in some cases the similarity of the tweets does not always correlate with the geographical distance.

In both datasets, the best performing configuration in terms of accuracy (*Acc@1km*) and average error distance (*AED*), is obtained using the Top-9 tweets in the ranking and a value of $\alpha = 0.0$ (WMV@Top-9, $\alpha = 0.0$). On the other hand, the best performing configuration regarding coverage is obtained using the Top-3 tweets in the ranking and a value of $\alpha = 1.0$ (WMV@Top-3, $\alpha = 1.0$). Therefore, we observe that the goal of maximising coverage conflicts with reducing the average error distance. This set of observations answer the second research question (**RQ-4.2**) introduced in Section 4.1.1.

Table 4.1: Results for the Chicago dataset. The table presents the Average Error Distance in kilometres (*AED*), Median of Error distance (*MDE*), Accuracy at 1 kilometre (*A@1km*) and *Coverage* for our proposed approach (*WMV*) using the Top-N (*@TopN*) elements in the rank and values of α, against the baselines. Additionally, we present results of the best performing models of Chapter 3, *Aggregated* and *Individual*.

		Chicago			
Model	***Config***	***AED(km)***↓	***MED(km)***↓	***Acc@1km***↑	***Coverage***↑
Aggregated	*BM25*	4.806	0.906	50.37%	**99.40%**
Individual	*IDF*	**4.694**	**0.100**	**54.80%**	**99.40%**

		Chicago			
Model	***Config***	***AED(km)***↓	***MED(km)***↓	***Acc@1km***↑	***Coverage***↑
MV@Top-3	*No Weight*	2.484	**0.471**	76.09%	63.15%
MV@Top-5	*No Weight*	1.907	**0.471**	81.47%	54.53%
MV@Top-7	*No Weight*	1.702	**0.471**	83.51%	49.99%
MV@Top-9	*No Weight*	1.639	**0.471**	84.00%	46.87%
WMV@Top-3	$\alpha = 0.0$	3.488	**0.471**	67.21%	74.82%
WMV@Top-3	$\alpha = 0.25$	3.549	0.473	66.51%	75.95%
WMV@Top-3	$\alpha = 0.5$	3.692	0.481	65.20%	77.67%
WMV@Top-3	$\alpha = 0.75$	4.020	0.503	62.01%	81.83%
WMV@Top-3	$\alpha = 1.0$	4.365	0.532	58.88%	**84.15%**
WMV@Top-5	$\alpha = 0.0$	2.134	**0.471**	79.68%	59.75%
WMV@Top-5	$\alpha = 0.25$	2.178	**0.471**	79.21%	60.20%
WMV@Top-5	$\alpha = 0.5$	2.310	**0.471**	77.79%	61.30%
WMV@Top-5	$\alpha = 0.75$	2.709	**0.471**	73.69%	64.71%
WMV@Top-5	$\alpha = 1.0$	3.829	0.498	63.24%	75.41%
WMV@Top-7	$\alpha = 0.0$	1.748	**0.471**	83.65%	54.44%
WMV@Top-7	$\alpha = 0.25$	1.767	**0.471**	83.34%	54.55%
WMV@Top-7	$\alpha = 0.5$	1.863	**0.471**	82.25%	54.99%
WMV@Top-7	$\alpha = 0.75$	2.128	**0.471**	79.54%	56.87%
WMV@Top-7	$\alpha = 1.0$	3.117	**0.471**	69.80%	64.58%
WMV@Top-9	$\alpha = 0.0$	**1.602**	**0.471**	**85.14%**	51.39%
WMV@Top-9	$\alpha = 0.25$	1.647	**0.471**	84.60%	51.58%
WMV@Top-9	$\alpha = 0.5$	1.712	**0.471**	83.90%	51.86%
WMV@Top-9	$\alpha = 0.75$	1.897	**0.471**	81.92%	52.84%
WMV@Top-9	$\alpha = 1.0$	2.730	**0.471**	73.50%	58.24%

Table 4.2: Results for the New York dataset. The table presents the Average Error Distance in kilometres (*AED*), Median of Error distance (*MDE*), Accuracy at 1 kilometre (*A@1km*) and *Coverage* for our proposed approach (*WMV*) using the Top-N (*@TopN*) elements in the rank and values of α, against the baselines. Additionally, we present results of the best performing models of Chapter 3, *Aggregated* and *Individual*.

		New York			
Model	***Config***	***AED(km)***↓	***MED(km)***↓	***Acc@1km***↑	***Coverage***↑
Aggregated	*BM25*	**4.862**	1.547	45.40%	**99.98%**
Individual	*TF-IDF*	4.972	**1.251**	**48.46%**	**99.98%**

		New York			
Model	***Config***	***AED(km)***↓	***MED(km)***↓	***Acc@1km***↑	***Coverage***↑
MV@Top-3	*No Weight*	2.522	0.461	72.76%	55.31%
MV@Top-5	*No Weight*	1.878	0.428	79.85%	46.20%
MV@Top-7	*No Weight*	1.610	0.412	82.52%	41.67%
MV@Top-9	*No Weight*	**1.448**	**0.405**	84.00%	38.70%
WMV@Top-3	$\alpha = 0.0$	3.949	0.556	58.86%	72.22%
WMV@Top-3	$\alpha = 0.25$	4.011	0.567	58.13%	73.87%
WMV@Top-3	$\alpha = 0.5$	4.174	0.602	56.69%	76.25%
WMV@Top-3	$\alpha = 0.75$	4.459	0.668	53.89%	**80.93%**
WMV@Top-3	$\alpha = 1.0$	4.567	0.703	52.56%	79.71%
WMV@Top-5	$\alpha = 0.0$	2.264	0.444	76.03%	52.71%
WMV@Top-5	$\alpha = 0.25$	2.310	0.447	75.30%	53.46%
WMV@Top-5	$\alpha = 0.5$	2.504	0.457	73.33%	55.41%
WMV@Top-5	$\alpha = 0.75$	3.127	0.485	66.85%	61.52%
WMV@Top-5	$\alpha = 1.0$	4.392	0.642	55.00%	74.92%
WMV@Top-7	$\alpha = 0.0$	1.687	0.417	81.73%	46.35%
WMV@Top-7	$\alpha = 0.25$	1.712	0.418	81.35%	46.73%
WMV@Top-7	$\alpha = 0.5$	1.817	0.424	80.22%	47.56%
WMV@Top-7	$\alpha = 0.75$	2.209	0.441	75.75%	50.78%
WMV@Top-7	$\alpha = 1.0$	3.931	0.545	59.53%	65.67%
WMV@Top-9	$\alpha = 0.0$	1.490	0.412	**84.09%**	43.36%
WMV@Top-9	$\alpha = 0.25$	1.499	0.412	83.63%	43.56%
WMV@Top-9	$\alpha = 0.5$	1.586	0.412	82.90%	44.24%
WMV@Top-9	$\alpha = 0.75$	1.889	0.424	79.90%	46.08%
WMV@Top-9	$\alpha = 1.0$	3.229	0.488	65.57%	56.57%

Additionally, Figures 4.4 and 4.5 show the distribution of the average error distance (*AED*) and *Coverage*, respectively, in the Chicago across values of $N \in \{3, 5, 7, 9, ..., 49\}$ for the Top-N for any of the α values considered (0.0, 0.25, 0.50, 0.75, 1.0). We observe a logarithmic decay of the error distance as the values of N for the Top-N increases. Moreover, we identify a big jump when considering the Top-5 and Top-10 tweets. This suggests that the Top-10 geotagged tweets are the most informative concerning geographical evidence for geolocalisation. When values of N are higher than 10, we observe that the decrease is gradually smaller, which suggest that the rest of the tweets in the ranking are less informative.

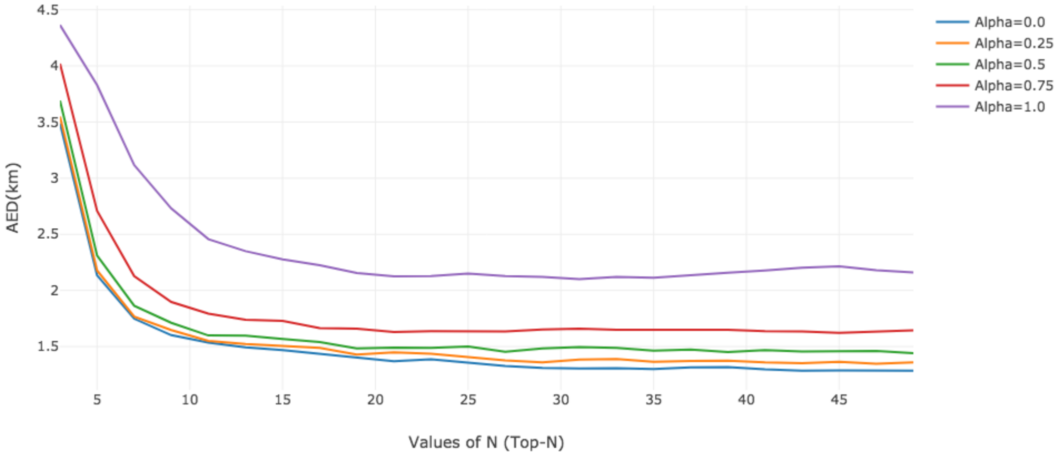

Figure 4.4: Distribution of the Average Error Distance (*AED*) in the Chicago dataset when considering values of $N \in \{3, 5, 7, 9, ..., 49\}$ for the Top-N most similar geotagged tweets.

4.5.2 Effect of Weighting The Votes

The effects of Equation 4.4 of the weighted majority voting models (*WMV*) can be observed in Table 4.1, Table 4.2 and Figure 4.4. As the values of alpha decrease, our approach achieves higher accuracy (*Acc@1km*), and reduce the average error distance (*AED*). This pattern can be observed for any of the investigated values of N for the Top-N tweets in the rank. Additionally, compared to the majority voting models (*MV*), the weights of the votes is capable of alleviating the decrease

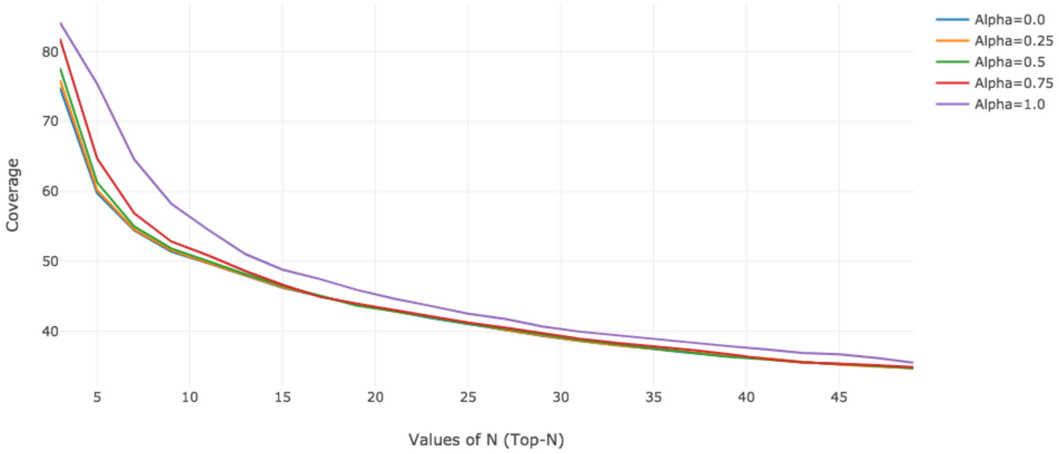

Figure 4.5: Distribution of the Coverage in the Chicago dataset when considering values of $N \in \{3, 5, 7, 9, ..., 49\}$ for the Top-N most similar geotagged tweets.

of coverage. In particular, *WMV@Top-9* achieves better average error distance while maintaining higher coverage than the best *MV* model (*MV@Top-9*).

This observation suggests that, in some cases, the majority voting alone does not return predictions if any location within the Top-N geotagged tweets does not accumulate the critic number of votes. Therefore, by pondering the importance of the votes, the weighted majority voting is capable of discriminate a location and find a fine-grained prediction for such cases.

4.5.3 Comparing Behaviour Across Datasets

In this section, we explore the similarities and differences in behaviour of our approach across both datasets, Chicago and New York. In particular, our approach exhibit two main patterns in both datasets.

First, as the number of values of N for the Top-N tweets increases, we observe that the average error distance (*AED*) decreases, accuracy at 1 km (*Acc@1km*) increases, and coverage (*Coverage*) decreases. Additionally, in Tables 4.1 and 4.2 we observe that, when considering the *WMV@Top-9* and $\alpha = 0.0$, our approach is capable reducing *AED* and increasing *Acc@1km* while increasing the coverage with respect to the best baseline (MV@Top-9). Second, we identified in both datasets the pattern regarding the weighting of the votes, discussed in Section 4.5.2. As values of α for the weighting function (See Equation 4.4) are closer to 0.0, we observe a decrease in AED, and increase in *Acc@1km* and a drop in Coverage. On the other hand, we notice that, overall, we achieve higher coverage and lower average error distance in the Chicago dataset compared to the New York dataset. This variation can be explained by the difference in size between datasets, reported in Section 3.3.1. Due to this, the ratio of tweets that our approach is capable of finding a prediction is lower in the New York dataset than the Chicago dataset and, therefore, we obtain lower coverage and average error distance.

Despite their geographical and cultural differences, our approach performs similarly across the two cities investigated in our experiments, Chicago and New York. Moreover, our approach is data-driven and does not require specific information of the city to perform, such as location or place names. Therefore, the consistency of behaviour observed in our two used datasets (Chicago and New York) appear to support the generalisation of our approach and suggests that our approach can be generalised and adapted to different cities.

4.6 Chapter Summary

In Chapter 3 we demonstrated that fine-grained geolocalisation could be achieved by representing candidate locations as individual tweets. However, the average error distance of the predictions returned by such approach is still not sufficient to reliably enable tasks that require fine-grained geolocated data, such as traffic incident detection – the best average error distance is 4.693 km (Chicago) which represents a confidence area of 69.19 km^2, whereas we aimed to reduce the average error to 1 km which represents a confidence area of 3.14 km^2. To achieve that, we proposed a new approach to reduce the average error distance returned by the geolocalisation method.

In this chapter, we hypothesised that in some cases the similarity of the tweets does not always correlate with their geographical distance. Therefore, there may not be sufficient evidence to return a fine-grained prediction in such cases. These cases are being considered by approaches derived in Chapter 3 as they always return the location of the most similar geotagged tweet (Top-1). We believe that by identifying such cases, we can increase the quality of the predictions at a fine-grained level. Additionally, based on this assumption we developed a theoretical framework illustrated in Figure 4.1, which presents the correlation between the content similarity and the geographical distance of a tweet to other geotagged tweets. Next, we identified the utility areas in Figure 4.2 that we targeted in this chapter to obtain the most fine-grained predictions, and concluded that we could achieve that by exploring evidence of the geographical distance between the Top-N geotagged tweets, regardless of their content similarity to the non-geotagged tweet.

To combine evidence from the Top-N geotagged tweets, we proposed to model fine-grained geolocalisation as a voting process, where each candidate location is represented as a set of geotagged tweets. We adopted a majority voting algorithm to estimate the geographical location by collecting the geolocation votes of the geotagged tweets in the rank. In the case that the voting process finds a location with a majority of the votes, it is indicative of low geographical distance, and we consider that there is sufficient evidence for a fine-grained prediction.

We contextualised the work in this chapter into two research questions, introduced in Section 4.1.1. In order to address them, we experimented with a set of geotagged tweets collected from Chicago and New York, using the experimental settings utilised in Chapter 3 (Section 3.3). Results were presented in Tables 4.1 and 4.2. Firstly, our experimental results showed that our weighted majority voting is capable of increasing the performance regarding accuracy (*Acc@1km*) and average error distance (*AED*), in both cities, across all the investigated values of N for the Top-N tweets. We identified that the best performing configuration in terms of accuracy and error distance is obtained using the Top-9 tweets in the ranking and a value of $\alpha = 0.0$. This observation addressed research question **RQ-4.1**.

Moreover, we observed that as the number of voting candidates (i.e., Top-N) increases, our approach achieved lower error distance, higher accuracy but lower coverage. We identified that the best performing configuration in terms of coverage is obtained using the Top-3 tweets in the ranking and a value of $\alpha = 1.0$. Therefore, we observed that the goal of maximising coverage conflicts with reducing the average error distance. These results addressed the research question (**RQ-4.2**).

Finally, we analysed the effect of the weighting in the majority voting algorithm. We weighted the votes of each geotagged tweet in the Top-N using information about the credibility of the user that posted the tweet (See 4.3.2.1), and the content similarity to the non-geotagged tweet (See 4.3.2.2). We combined the weights using Equation 4.4, and controlled the influence of the credibility and the similarity by a parameter $\alpha \in [0, 1]$. We observed that by weighting the majority voting, we alleviated the decrease of coverage.

So far, our work is generating a ranking list of candidate locations using retrieval models, being IDF weighting the best performing one. However, the quality of the Top-N ranked elements can be improved further and thus the performance of geolocalisation. In Chapter 5 we will explore how to integrate Learning to Rank techniques into the fine-grained geolocalisation task to improve the ranking and, therefore, the performance of the task.

Chapter 5

Learning to Geolocalise

5.1 Introduction

Previously, in Chapter 3, we introduced a ranking approach for fine-grained geolocalisation and demonstrated that, in contrast to existing works in the literature, considering geotagged tweets as individual documents lead to better performance, regardless of the retrieval model utilised for ranking the documents. Also, we observed that document frequency is the most discriminative feature and compared the performance of different retrieval models. Among the tested models, the IDF weighting model showed to be the best for geolocalisation (see Section 3.4.1). Next, in Chapter 4, we explored the geographical evidence encoded within the Top-N most similar geotagged tweets by adopting a weighted majority voting algorithm that collects the geolocation votes of the tweets in the ranking. We achieved an average error distance of 1.602 km, which represents a confidence area of 8.06 km^2 (See *WMV@Top-9,* $\alpha = 0.0$ in Table 3.4). Nevertheless, as introduced in Chapter 1 (Section 1.2), in this thesis we aim to obtain fine-grained predictions with an average error distance of 1 km, which represents a confidence area of 3.14 km^2.

The approaches explored before in this thesis obtained the Top-N most similar tweets using a retrieval model which computes the similarity based on document frequency information (IDF weighting). However, considering only document frequency to perform the ranking can limit the quality of the Top-N geotagged tweets. In this chapter, we postulate that by improving the ranking component of previous approaches will lead to an improvement in the performance of the fine-grained geolocalisation. According to Figure 4.2 in Chapter 4, which presents the correlation between similarity and geographical distance between tweets, the set of fine-grained predictions will fall within the area where high similarity yet low geographical distance. Thus, we hypothesise that by improving the ranking of geotagged tweets with respect to a given non-geotagged tweet, we can obtain more similar and geographically closer geotagged tweets, and thus we can obtain a higher number of other fine-grained predictions (see **Hypothesis 3** in Section 1.2).

In order to improve the ranking, instead of only considering document frequency information, we aim to combine multiple indicators from the tweets to learn a new ranking function. As introduced in Chapter 2 (Section 2.2.3), learning to rank approaches have the capacity of doing so by using machine learning techniques in order to learn a more effective ranking function. Learning to rank approaches have demonstrated to benefit effectiveness in several retrieval tasks using web documents or large text documents (Liu et al., 2009). Also, previous research has demonstrated improvements in retrieval tasks using short documents, such as Twitter posts (Cheng et al., 2012). Therefore, in this thesis, we adopt a learning to rank approach to rank geotagged tweets for fine-grained geolocalisation.

Our approach learns from the characteristics of pairs of geotagged tweets posted within the same fine-grained area (i.e., squared areas of length size 1 km), and re-ranks geotagged tweets based on their geographical proximity. We propose multiple types of features for geolocalisation and evaluate our proposed approach using a ground truth of geotagged tweets gathered from two different cities. Additionally, we investigate the best type of features for fine-grained geolocalisation. We focus the work in this chapter towards addressing two research questions:

- **RQ-5.1:** What is the best performing learning to rank algorithm to improve the ranking?
- **RQ-5.2:** Does improving the ranking of the geotagged tweets lead to better fine-grained geolocalisation?
- **RQ-5.3:** What set of features contributes the most to improve the accuracy of fine-grained geolocalisation?

The remainder of the chapter is organised as follows. Section 5.2 introduces our learning to rank approach for fine-grained geolocalisation of tweets. Section 5.3 describes our experimental setup and the evaluation of our proposed approach. Section 5.4 presents and discusses our results. Lastly, we provide concluding remarks in Section 5.5.

5.2 Learning to Geolocalise

As introduced in Section 5.1, we aim to use a learning to rank approach to improving the ranking of the Top-N most content-based similar geotagged tweets (denoted as a **doc-tweet**) to a given non-geotagged tweet (denoted as a **query-tweet**), and thus improve the effectiveness of fine-grained geolocalisation. To this end, we aim to learn a ranking function to re-rank doc-tweets based on their geographical proximity to the query-tweet. We experiment with different learning to rank algorithms in Section 5.3.2. Also, we propose a set of features to learn our function. We extract these features from pairs of geotagged tweets posted within the same fine-grained area (i.e., 1 km squared area).

Our proposed approach consists of two main components. First, we use our learned ranking function to re-rank doc-tweets based on their probability of being posted in the same area as the query-tweet. Finally, we feed the Top-N doc-tweets into a majority voting algorithm (as described in Section 4.3.1) to select the predicted location - a squared area of size 1km - within the Top-N doc-tweets.

Next, in Section 5.3.2, we describe in detail our proposed features for features.

5.2.1 Feature Set For Fine-Grained Tweet Geolocalisation

For each pair of geotagged tweets, we extract a set of features to model fine-grained geolocalisation. We compute document features extracted from the doc-tweet, query features extracted from the query-tweet, as well as query-dependent features to model the relationship between query-tweets and doc-tweets. In total we extracted 28 features, presented in Table 5.1. We describe and motivate each feature next.

Table 5.1: Features extracted for fine-grained geolocalisation of tweets.

Features	**Description**	**Total**
Query Features and Document Features		
Hashtags	Number of hashtags in the text.	2
Mentions	Number of mentions in the text.	2
Urls	Number of urls in the text.	2
Entities	Number of entities in the text.	2
Verbs	Number of verbs in the text.	2
Adverbs	Number of adverbs in the text.	2
Adjectives	Number of adjectives in the text.	2
Checkin	Whether the tweet is a Foursquare checkin.	2
Hour	The hour of the day (0 to 24h) that the tweet was posted.	2
Weekday	The day of the week (Monday to Sunday) that the tweet was posted.	2
User Ratio	User credibility ratio (See Section 4.3.2.1).	2
Query-dependent Features		
Hashtags	Shared number of Hashtags.	1
Mentions	Shared number of Mentions.	1
User	Whether both tweets belong to the same user.	1
Hour	Whether both tweets are posted the same hour of the day (0h to 24h).	1
Weekday	Whether both tweets are posted same day of the week (Monday to Sunday).	1
IDF	Similarity score given by the IDF weighting.	
Total Features		28

5.2.1.1 Query Features and Document Features

We extract features from the query-tweet and the doc-tweet independently. We categorise these features in two groups: content quality and geospecific features.

Content Quality Features. The more quality of the content of a tweet is, the more valuable information it provides. Previous research has shown the usefulness of content quality features of a tweet for learning to rank (Cheng et al., 2012; Damak et al., 2013; Duan et al., 2010; Han et al., 2012b). Inspired by these works, we modelled the quality of a tweet by extracting indicators of the richness of its text. We extract a total of 8 different features. First, we exploit the characteristics of the Twitter social network by counting the number of hashtags, the number of mentions and number of URLs of the tweet. Second, we utilise natural language techniques to count the number of entities, verbs, adjectives, nouns and adverbs in the text.

Geospecific Features. In addition to previous state-of-the-art features, we added new features as signals for geolocalisation by extracting geospecific information contained within the query-tweet and the doc-tweet. We compute a total of 4 different features. First, we check if the tweet corresponds to a Foursquare[1] check-in. Foursquare is a social media network in which users can do check-ins at venues when they visit them. Users have the option of generating a tweet sharing this information with their followers along with the geolocation of the venue.

Second, following the weighting majority voting approach introduced in Chapter 3 (Section 3.2), we compute a credibility score for the tweet which represents the posting activity of the user that generated the tweet. A tweet posted by a user with a high score is more likely to be indicative of a geolocalisation. The credibility score is based on the ratio of tweets posted by a user at a fine-grained distance (1 km) to other similar tweets (Top-N). We utilise the training and validation set described in Section 5.3.1 to compute the score.

Finally, different types of events tend to occur at different hours of the day or days of the week. For instance, people usually visit clubs at nights and weekends. Thus, if two tweets were posted in the same time frame, their content is likely to be related to the same type of events that are recurrent in the same location. To model that, we add the hour of the day (0 to 24 hours) and the day of the week (Monday to Sunday) as features.

5.2.1.2 Query-Dependent Features.

Query-dependent features aim to model the relationship between the query-tweet and the doc-tweet. These set of features are presented in Table 5.1. The intuition behind these features is that when people visit a certain location, they make use of social media to describe their surroundings or events occurring in the location. This means that many of the generated tweets will share the same characteristics. Therefore, the similarities between the two tweets are a strong indicator of their geolocalisation. Firstly, we model the similarities between the query-tweet and the doc-tweet by computing their IDF similarity score. Second, we count the number of common entities, mentions and hashtags, and check if the same user posted both tweets. Finally, we calculate if the query-tweet and the doc-tweet were generated in the same hour of the day or on the same day of the week.

[1] http://www.foursquare.com

5.3 Experiments

In this section, we describe the experimental setup that supports the evaluation of our proposed learning to rank approach for fine-grained geolocalisation tweets. In order to compare results with previous approaches, we utilise the same experimental settings described previously in Chapter 3:

- We experiment using the two datasets of geotagged tweets described in Section 3.3.1 (Chicago and New York).
- We preprocess and index each geotagged tweet following Section 3.3.2.
- We report the same metrics described in Section 3.3.6, namely average error distance (AED), median error distance (MED), accuracy at 1 km (Acc@1km) and coverage (*Coverage*).

5.3.1 Creating Training and Testing Sets for Learning to Rank

In order to evaluate our learning to rank approach, we generate training and a testing set for each of our datasets (Chicago and New York). First, we divide the dataset following Section 3.3.1 and create three subsets. The first set (named document set) contains the geotagged tweets from the first three weeks of March 2016, resulting in 100,176 geotagged tweets for Chicago and 111,292 for New York. Second, we randomly divide the last week of March into background-queries set and test-queries set to ensure the same characteristics. The background-queries set consists of 16,262 geotagged tweets for Chicago, and 20,982 geotagged tweets for New York. Finally, the test-queries set contains 16,313 geotagged tweets for Chicago and 20,870 geotagged tweets for New York. It is important to note that we preserve the same testing and training/document tweets from Chapter 3 and Chapter 4 for comparison.

Training and Test: After dividing the datasets, we create our training set and test sets for learning to rank as follows. First, we perform a retrieval task (using IDF weighting model) with the geotagged tweets in the background-queries set as query-tweets and the geotagged tweets in the document set as doc-tweets. We use the generated pairs of query-tweet and doc-tweet as a training set to train our learning to rank approach. Finally, we perform the same task but using the tweets in the test-queries set as query-tweets to build the test set for evaluating our learning to rank approach.

Labelling: We label pairs of query-tweet and doc-tweet in the training and test sets described above. As explained in Section 5.2 we re-rank doc-tweets based on their geographical proximity to the query-tweet. Therefore, we first divide the geographical space of interest into a grid of fine-grained squared areas of size 1 km and associate each geotagged query-tweet and doc-tweet to their corresponding area based on their longitude/latitude location. Then, pairs of tweets posted in the same area (i.e. distance 1 km or less) are labelled as positive. On the other hand, pairs of tweets posted in different areas (i.e. distance more than 1 km) are labelled as negative.

5.3.2 Models

In total, we implement four version of our learning to rank approaches using four different subsets of features. As a baseline, we use the best performing ranking approach (*Individual*) proposed in Chapter 3, and the majority voting version (*MV*) proposed in Chapter 4.

L2Geo and L2Geo+MV models

We implement our proposed learning to rank approach, described in Section 5.2. We experiment with different learning to rank algorithms as ranking functions for our approach. We configure the ranking functions to re-rank the Top-100 most similar geotagged tweets obtained by the baseline (IDF weighting), and optimise NDCG@N with N values of 3, 5, 10, 20, 30, 40 and 50 during the training process. After the ranking process, we return a predicted location in two ways:

- **L2Geo:** In this model we return the longitude/latitude of the Top-1 geotagged tweet re-ranked by our learning to rank algorithm as the predicted location, following the approach in Chapter 3.

- **L2Geo+MV:** In this model, we feed the Top-N most similar geotagged tweets into the majority voting algorithm described in Section 4.3.1.

Additionally, in order to assess the best set of features for fine-grained geolocalisation, we built nine different versions of our approaches that use different combinations of the features described in Section 5.2.1, denoted by:

- **All:** This model incorporates all the features.
- **Common:** This model uses only the set of query-dependent features.
- **Query:** This model incorporates features extracted only from the query-tweet.
- **Doc:** This model incorporates features extracted only from the doc-tweet.
- **Query_Doc:** This model combines features extracted from the query-tweet and the doc-tweet.
- **Query_Common:** This model uses the query-dependent features along with features extracted from the query-tweet.
- **Doc_Common:** This model uses the query-dependent features along with features extracted from the doc-tweet.
- **Query_Content:** This model utilises the set of content quality features extracted only from the query-tweet.
- **Doc_Content:** This model utilises the set of content quality features extracted only from the doc-tweet.
- **Query_Geo:** This model uses the set of geospecific features extracted only from the query-tweet.
- **Doc_Geo:** This model uses the set of geospecific features extracted only from the doc-tweet.

5.4 Experimental Results

In this section, we present results fine-grained geolocalisation as follows. First, we evaluate the effectiveness of different learning to rank algorithms in Subsection 5.4.1, and assess if the learning to rank approach is improving the ranking of geotagged tweets at the Top-N positions compared to the baseline (*IDF* weighting).

Tables 5.2 (Chicago dataset) and 5.3 (New York dataset) compares the performance of the ranking generated by different learning to rank algorithms. We compare the ranking performance against the baseline (*IDF*). We train the learning to rank algorithms to optimise NDCG@3, @5, @10, @20, @30, @40 and @50. Finally, for each algorithm, we report NDCG@1, @3, @5 and @10. This results aim to address research question **RQ-5.1**. Additionally, we conduct a Randomised permutation test to asses statistical differences ($p \leq 0.01$). In particular, we use:

- $\gg$ results that are significantly better than the baseline (*IDF weighting*),
- $\ll$ to denote measures that are significantly worse than the baseline, and
- $=$ to denote no statistical differences.

Second, we evaluate whether improving the ranking leads to an increase in performance of the fine-grained geolocalisation. First, in Section 5.4.2, we use the *L2Geo* model, described in Section 5.3.2, which returns always the Top-1 geotagged tweets as the predicted location. Moreover, we evaluate the quality the geographical evidence encoded within the Top-N geotagged tweets obtained by our learning to rank approach by applying the majority voting algorithm (*L2Geo+MV*). We report these results using the best performing learning to rank algorithm from Section 5.4.1.

Tables 5.4 (Chicago dataset) and 5.5 (New York dataset) present the results on fine-grained geolocalisation for our learning to rank models using a ranking approach returning the Top-1 geotagged tweets a the predicted location (*L2Geo*). We compare this model against the best performing approach explored in Chapter 3 (*Individual*). Moreover, Tables 5.6 and 5.7 present the performance of our learning to rank models with the majority voting algorithm applied on the Top-N most similar geotagged tweets (*L2Geo+MV*). In each of the tables mentioned above, we report the metrics described in Section 3.3.6, namely; average error distance (*AED*), median error distance (*MED*), accuracy (*Acc@1km*), and coverage (*Coverage*). Due to the differences in *Coverage* obtained by different configurations of *L2Geo+MV* models, the models return a prediction for different subsets of the tweets in the test set. For this reason, we do not compare (statistically) these results against our baselines. Finally, for each measure, we denote in bold the best performing approach. This set o results aim to address research question **RQ-5.2**, introduced in Section 5.1. Finally, we discuss the effects of the different types of features proposed in Section 5.2.1, which aims to answer the research question **RQ-5.3**. Now we describe the presentation of the tables before the analysis of the results.

The remainder of this section is as follows. In Section 5.4.1 we address **RQ-5.1** and compare different learning to rank algorithms. Next, in Section 5.4.2 we address **RQ-5.2** and discuss the improvement of the ranking using learning to rank and the impact on the effectiveness of fine-grained geolocalisation. Section 5.4.2.1 analyses results when applying the majority voting algorithm to consider geographical evidence within the Top-N most similar geotagged tweets. Moreover, we address **RQ-5.3** and discuss the best type of features for fine-grained geolocalisation. Finally, we provide concluding remarks in Section 5.5.

5.4.1 Performance of Learning to Rank Algorithms

We first compare the ranking performance of state-of-the-art learning to rank algorithms in order to whether we can improve the ranking of geotagged tweets compared to the baseline (IDF weighting). We use the *L2Geo* model which incorporates all the features *All*, described in Section 5.2.1. As introduced in Section 2.2.3, learning to rank algorithms can be categorised in three groups: point-wise, pair-wise and list-wise approaches. In total, we compare the performance of six different algorithms representing the three mentioned groups, namely:

- **Point-wise:** MART (Friedman, 2001) and Random Forests (Breiman, 2001).
- **Pair-wise:** RankNet (Burges et al., 2005).

- **List-wise:** AdaRank (Xu and Li, 2007) and ListNet (Cao et al., 2007).
- **Pair-wise/List-wise:** LambdaMART (Wu et al., 2010) [1].

Tables 5.2 and 5.3 shows the performance for the Chicago and New York datasets respectively. First, we do not observe different behaviour when using point-wise, pair-wise or list-wise algorithms. However, we observe that LambdaMART shows the best performance overall, and significant (statistical) improve the baseline *IDF*. These results suggest that the LambdaMART algorithm is the most suitable algorithm to improve the ranking, which answers **RQ-5.1**.

Also, we identify the best performance is obtained at the Top-1 geotagged tweet (NDCG@1), but this performance decreases as more tweets of the Top-N are considered (up to NDCG@10). On the other hand, for training the LambdaMART algorithm, we identify that the best optimisation metric is NDCG@10 for the Chicago dataset and NDCG@30 for the New York dataset.

In this section, we demonstrate that our learning to rank approach can improve the ranking over the baseline (IDF weighting). Next, we aim to asses whether this improvement leads to better performance in fine-grained geolocalisation. In the next section, we only report experimental results on fine-grained geolocalisation using the best performing configurations for each dataset. These configurations are:

- LambdaMART optimising NCDG@10 for Chicago, and
- LambdaMART optimising NDCG@30 for New York.

5.4.2 Effectiveness on Fine-Grained Geolocalisation

In order to assess the effectiveness on fine-grained geolocalisation, we first perform predictions returning always the Top-1 most similar geotagged tweets (*L2Geo*), following the approach in Chapter 3. We compare the performance against the best ranking approach in Chapter 3, *Individual*. The results are presented in Tables 5.4 and 5.5 for the Chicago and New York dataset respectively.

[1] According to Wu et al. (2010), LambdaMART is both pair-wise and list-wise

Table 5.2: Ranking performance for the Chicago dataset. The table presents NDCG@1, @3, @5 and @10 for the learning to rank algorithms and the baseline (*IDF* weighting). We run our experiment using all the features (*All*). A randomized permutation test was conducted to show significant differences with respect to the baseline (IDF), denoted by $\gg$ (p$<$0.01) for better performance, $\ll$ for worse performance and = for no statistical difference.

		NDCG			
Ranking	**Optimisation**	**@1**	**@3**	**@5**	**@10**
IDF	*N/A*	0.5513	0.5261	0.5136	0.5010
Mart	*NDCG@3*	0.5787$^{\gg}$	0.5631$^{\gg}$	0.553$^{\gg}$	0.5424$^{\gg}$
Mart	*NDCG@5*	0.5787$^{\gg}$	0.5631$^{\gg}$	0.553$^{\gg}$	0.5424$^{\gg}$
Mart	*NDCG@10*	0.5787$^{\gg}$	0.5631$^{\gg}$	0.553$^{\gg}$	0.5424$^{\gg}$
Mart	*NDCG@20*	0.5787$^{\gg}$	0.5631$^{\gg}$	0.553$^{\gg}$	0.5424$^{\gg}$
Mart	*NDCG@30*	0.5787$^{\gg}$	0.5631$^{\gg}$	0.553$^{\gg}$	0.5424$^{\gg}$
Mart	*NDCG@40*	0.5787$^{\gg}$	0.5631$^{\gg}$	0.553$^{\gg}$	0.5424$^{\gg}$
Mart	*NDCG@50*	0.5787$^{\gg}$	0.5631$^{\gg}$	0.553$^{\gg}$	0.5424$^{\gg}$
Randomforest	*NDCG@3*	0.5469$^{\gg}$	0.5338$^{\gg}$	0.5259$^{\gg}$	0.5176$^{\gg}$
Randomforest	*NDCG@5*	0.5453$^{\gg}$	0.5323$^{\gg}$	0.5244$^{\gg}$	0.5164$^{\gg}$
Randomforest	*NDCG@10*	0.5408$^{\gg}$	0.5307$^{\gg}$	0.5231$^{\gg}$	0.5151$^{\gg}$
Randomforest	*NDCG@20*	0.5438$^{\gg}$	0.5312$^{\gg}$	0.5235$^{\gg}$	0.516$^{\gg}$
Randomforest	*NDCG@30*	0.544$^{\gg}$	0.532$^{\gg}$	0.5247$^{\gg}$	0.5168$^{\gg}$
Randomforest	*NDCG@40*	0.5442$^{\gg}$	0.5318$^{\gg}$	0.5243$^{\gg}$	0.5163$^{\gg}$
Randomforest	*NDCG@50*	0.5431$^{\gg}$	0.5328$^{\gg}$	0.5247$^{\gg}$	0.5169$^{\gg}$
Ranknet	*NDCG@3*	0.5521$^{\gg}$	0.5261$^{\gg}$	0.5131$^{\gg}$	0.5001$^{\ll}$
Ranknet	*NDCG@5*	0.5521$^{\gg}$	0.5261$^{\gg}$	0.5131$^{\gg}$	0.5001$^{\ll}$
Ranknet	*NDCG@10*	0.5521$^{\gg}$	0.5263$^{\gg}$	0.5132$^{\gg}$	0.5003$^{\ll}$
Ranknet	*NDCG@20*	0.552$^{\gg}$	0.5261$^{\gg}$	0.5131$^{\gg}$	0.5002$^{\ll}$
Ranknet	*NDCG@30*	0.5521$^{\gg}$	0.5262$^{\gg}$	0.5131$^{\gg}$	0.5002$^{\ll}$
Ranknet	*NDCG@40*	0.5521$^{\gg}$	0.5261$^{\gg}$	0.5131$^{\gg}$	0.5001$^{\ll}$
Ranknet	*NDCG@50*	0.552$^{\gg}$	0.5261$^{\gg}$	0.5131$^{\gg}$	0.5002$^{\ll}$
Lambda	*NDCG@3*	0.6272$^{\gg}$	0.6026$^{\gg}$	0.589$^{\gg}$	0.5732$^{\gg}$
Lambda	*NDCG@5*	**0.6274**$^{\gg}$	0.6039$^{\gg}$	0.5908$^{\gg}$	0.5749$^{\gg}$
Lambda	*NDCG@10*	0.6273$^{\gg}$	0.6045$^{\gg}$	**0.5915**$^{\gg}$	**0.5757**$^{\gg}$
Lambda	*NDCG@20*	0.6268$^{\gg}$	0.6037$^{\gg}$	0.5906$^{\gg}$	0.5756$^{\gg}$
Lambda	*NDCG@30*	0.6268$^{\gg}$	0.6046$^{\gg}$	0.5906$^{\gg}$	0.5754$^{\gg}$
Lambda	*NDCG@40*	**0.6274**$^{\gg}$	0.6039$^{\gg}$	0.5904$^{\gg}$	0.5755$^{\gg}$
Lambda	*NDCG@50*	0.6263$^{\gg}$	**0.6050**$^{\gg}$	0.5913$^{\gg}$	0.5755$^{\gg}$
Adarank	*NDCG@3*	0.5851$^{\gg}$	0.5616$^{\gg}$	0.5499$^{\gg}$	0.5371$^{\gg}$
Adarank	*NDCG@5*	0.5928$^{\gg}$	0.5686$^{\gg}$	0.5554$^{\gg}$	0.5423$^{\gg}$
Adarank	*NDCG@10*	0.587$^{\gg}$	0.5642$^{\gg}$	0.5527$^{\gg}$	0.5395$^{\gg}$
Adarank	*NDCG@20*	0.5864$^{\gg}$	0.5635$^{\gg}$	0.5519$^{\gg}$	0.5389$^{\gg}$
Adarank	*NDCG@30*	0.5865$^{\gg}$	0.5635$^{\gg}$	0.5518$^{\gg}$	0.5389$^{\gg}$
Adarank	*NDCG@40*	0.5865$^{\gg}$	0.5635$^{\gg}$	0.5518$^{\gg}$	0.5389$^{\gg}$
Adarank	*NDCG@50*	0.5865$^{\gg}$	0.5635$^{\gg}$	0.5518$^{\gg}$	0.5389$^{\gg}$
Listnet	*NDCG@3*	0.5525$^{\gg}$	0.5274$^{\gg}$	0.5142$^{\gg}$	0.5020$^{\gg}$
Listnet	*NDCG@5*	0.5524$^{\gg}$	0.5273$^{\gg}$	0.5148$^{\gg}$	0.5019$^{\gg}$
Listnet	*NDCG@10*	0.5524$^{\gg}$	0.5274$^{\gg}$	0.5147$^{\gg}$	0.5019$^{\gg}$
Listnet	*NDCG@20*	0.5527$^{\gg}$	0.5273$^{\gg}$	0.5148$^{\gg}$	0.5017$^{\gg}$
Listnet	*NDCG@30*	0.5783$^{\gg}$	0.5595$^{\gg}$	0.5146$^{\gg}$	0.5376$^{\gg}$
Listnet	*NDCG@40*	0.5525$^{\gg}$	0.5274$^{\gg}$	0.5148$^{\gg}$	0.502$^{\gg}$
Listnet	*NDCG@50*	0.5524$^{\gg}$	0.5273$^{\gg}$	0.5148$^{\gg}$	0.5019$^{\gg}$

Table 5.3: Ranking performance for the New York dataset. The table presents NDCG@1, @3, @5 and @10 for the learning to rank algorithms and the baseline (*IDF* weighting). We run our experiment using all the features (*All*). A randomized permutation test was conducted to show significant differences with respect to the baseline (IDF), denoted by $\gg$ (p<0.01) for better performance, $\ll$ for worse performance and = for no statistical difference.

		NDCG			
Ranking	**Optimisation**	**@1**	**@3**	**@5**	**@10**
IDF	*N/A*	0.4798	0.4613	0.4520	0.4458
Mart	*NDCG@3*	$0.4935^{\gg}$	$0.4763^{\gg}$	$0.4704^{\gg}$	$0.4656^{\gg}$
Mart	*NDCG@5*	$0.4935^{\gg}$	$0.4763^{\gg}$	$0.4704^{\gg}$	$0.4656^{\gg}$
Mart	*NDCG@10*	$0.4935^{\gg}$	$0.4763^{\gg}$	$0.4704^{\gg}$	$0.4656^{\gg}$
Mart	*NDCG@20*	$0.4935^{\gg}$	$0.4763^{\gg}$	$0.4704^{\gg}$	$0.4656^{\gg}$
Mart	*NDCG@30*	$0.4935^{\gg}$	$0.4763^{\gg}$	$0.4704^{\gg}$	$0.4656^{\gg}$
Mart	*NDCG@40*	$0.4935^{\gg}$	$0.4763^{\gg}$	$0.4704^{\gg}$	$0.4656^{\gg}$
Mart	*NDCG@50*	$0.4935^{\gg}$	$0.4763^{\gg}$	$0.4704^{\gg}$	$0.4656^{\gg}$
Randomforest	*NDCG@3*	$0.4665^{\gg}$	$0.4514^{\gg}$	$0.4461^{\gg}$	$0.4428^{\ll}$
Randomforest	*NDCG@5*	$0.4676^{\gg}$	$0.4527^{\gg}$	$0.4475^{\gg}$	$0.444^{\ll}$
Randomforest	*NDCG@10*	$0.4649^{\gg}$	$0.4505^{\gg}$	$0.4456^{\ll}$	$0.4424^{\ll}$
Randomforest	*NDCG@20*	$0.4667^{\gg}$	$0.4521^{\gg}$	$0.4471^{\gg}$	$0.4433^{\ll}$
Randomforest	*NDCG@30*	$0.4655^{\gg}$	$0.4512^{\gg}$	$0.446^{\gg}$	$0.4428^{\ll}$
Randomforest	*NDCG@40*	$0.466^{\gg}$	$0.452^{\gg}$	$0.4466^{\gg}$	$0.4432^{\ll}$
Randomforest	*NDCG@50*	$0.4659^{\gg}$	$0.4515^{\gg}$	$0.4465^{\gg}$	$0.4433^{\ll}$
Ranknet	*NDCG@3*	$0.4821^{\gg}$	$0.4628^{\gg}$	$0.4533^{\gg}$	$0.4469^{\gg}$
Ranknet	*NDCG@5*	$0.4821^{\gg}$	$0.4628^{\gg}$	$0.4533^{\gg}$	$0.4469^{\gg}$
Ranknet	*NDCG@10*	$0.4821^{\gg}$	$0.4628^{\gg}$	$0.4533^{\gg}$	$0.4469^{\gg}$
Ranknet	*NDCG@20*	$0.4821^{\gg}$	$0.4628^{\gg}$	$0.4533^{\gg}$	$0.4469^{\gg}$
Ranknet	*NDCG@30*	$0.4821^{\gg}$	$0.4628^{\gg}$	$0.4533^{\gg}$	$0.4469^{\gg}$
Ranknet	*NDCG@40*	$0.4821^{\gg}$	$0.4628^{\gg}$	$0.4533^{\gg}$	$0.4469^{\gg}$
Ranknet	*NDCG@50*	$0.4821^{\gg}$	$0.4628^{\gg}$	$0.4533^{\gg}$	$0.4469^{\gg}$
Lambda	*NDCG@3*	$0.5424^{\gg}$	$0.5216^{\gg}$	$0.5100^{\gg}$	$0.4997^{\gg}$
Lambda	*NDCG@5*	$0.5450^{\gg}$	$0.5242^{\gg}$	$0.5129^{\gg}$	$0.5026^{\gg}$
Lambda	*NDCG@10*	$0.5461^{\gg}$	$0.5265^{\gg}$	$0.5151^{\gg}$	$0.5053^{\gg}$
Lambda	*NDCG@20*	$0.5474^{\gg}$	$0.5267^{\gg}$	$0.5160^{\gg}$	$0.5063^{\gg}$
Lambda	*NDCG@30*	$\mathbf{0.5478}^{\gg}$	$\mathbf{0.5274}^{\gg}$	$\mathbf{0.5164}^{\gg}$	$\mathbf{0.5068}^{\gg}$
Lambda	*NDCG@40*	$0.5454^{\gg}$	$0.5257^{\gg}$	$0.5147^{\gg}$	$0.5057^{\gg}$
Lambda	*NDCG@50*	$0.5469^{\gg}$	$0.5266^{\gg}$	$0.5157^{\gg}$	$0.5058^{\gg}$
Adarank	*NDCG@3*	$0.5136^{\gg}$	$0.492^{\gg}$	$0.4832^{\gg}$	$0.4758^{\gg}$
Adarank	*NDCG@5*	$0.5114^{\gg}$	$0.4888^{\gg}$	$0.48^{\gg}$	$0.4725^{\gg}$
Adarank	*NDCG@10*	$0.5098^{\gg}$	$0.487^{\gg}$	$0.4784^{\gg}$	$0.4714^{\gg}$
Adarank	*NDCG@20*	$0.5018^{\gg}$	$0.4812^{\gg}$	$0.4736^{\gg}$	$0.4674^{\gg}$
Adarank	*NDCG@30*	$0.5013^{\gg}$	$0.4809^{\gg}$	$0.4735^{\gg}$	$0.4673^{\gg}$
Adarank	*NDCG@40*	$0.5016^{\gg}$	$0.4811^{\gg}$	$0.4736^{\gg}$	$0.4675^{\gg}$
Adarank	*NDCG@50*	$0.507^{\gg}$	$0.4864^{\gg}$	$0.4786^{\gg}$	$0.4721^{\gg}$
Listnet	*NDCG@3*	$0.4824^{\gg}$	$0.4641^{\gg}$	$0.4549^{\gg}$	$0.4487^{\gg}$
Listnet	*NDCG@5*	$0.4826^{\gg}$	$0.4642^{\gg}$	$0.4549^{\gg}$	$0.4485^{\gg}$
Listnet	*NDCG@10*	$0.4828^{\gg}$	$0.4645^{\gg}$	$0.4552^{\gg}$	$0.4488^{\gg}$
Listnet	*NDCG@20*	$0.4827^{\gg}$	$0.4645^{\gg}$	$0.4553^{\gg}$	$0.4489^{\gg}$
Listnet	*NDCG@30*	$0.4827^{\gg}$	$0.4642^{\gg}$	$0.4549^{\gg}$	$0.4484^{\gg}$
Listnet	*NDCG@40*	$0.5164^{\gg}$	$0.4966^{\gg}$	$0.4876^{\gg}$	$0.4791^{\gg}$
Listnet	*NDCG@50*	$0.4827^{\gg}$	$0.4646^{\gg}$	$0.4553^{\gg}$	$0.4489^{\gg}$

Table 5.4: Fine-grained geolocalisation results for the Chicago dataset considering only the Top-1. We compare our learning to rank approach (*L2Geo*) against the baseline (*Individual* using *IDF*). We report average error distance (*AED*), median error distance (*MED*), accuracy at 1 km (*Acc@1km*) and coverage (*Coverage*). We use a paired t-test to assess significant differences, denoted by $\gg$ for better performance, $\ll$ for worse performance and = for no statistical difference.

Chicago					
Model	***Features***	***AED(km)***↓	***MED(km)***↓	***Acc@1km***↑	***Coverage***↑
Individual	*IDF*	4.694	**0.100**	54.80%	99.40%
L2Geo	*All*	**3.835**$^{\gg}$	0.514$^{\ll}$	**62.27%**$^{\gg}$	99.40%$^{=}$
L2Geo	*Common*	4.173$^{\gg}$	0.519$^{\ll}$	60.51%$^{\gg}$	99.40%$^{=}$
L2Geo	*Query*	4.893$^{\ll}$	0.594$^{\ll}$	54.90%$^{\gg}$	99.40%$^{=}$
L2Geo	*Doc*	6.462$^{\ll}$	3.126$^{\ll}$	38.77%$^{\ll}$	99.40%$^{=}$
L2Geo	*Query_Doc*	5.562$^{\ll}$	1.426$^{\ll}$	47.07%$^{\ll}$	99.40%$^{=}$
L2Geo	*Query_Common*	4.157$^{\gg}$	0.518$^{\ll}$	60.58%$^{\gg}$	99.40%$^{=}$
L2Geo	*Doc_Common*	3.847$^{\gg}$	0.516$^{\ll}$	62.01%$^{\gg}$	99.40%$^{=}$
L2Geo	*Geo_Query*	4.893$^{\ll}$	0.594$^{\ll}$	54.90%$^{\gg}$	99.40%$^{=}$
L2Geo	*Geo_Doc*	6.782$^{\ll}$	3.609$^{\ll}$	36.96%$^{\ll}$	99.40%$^{=}$
L2Geo	*Content_Query*	4.893$^{\ll}$	0.594$^{\ll}$	54.90%$^{\gg}$	99.40%$^{=}$
L2Geo	*Content_Doc*	6.897$^{\ll}$	3.949$^{\ll}$	36.05%$^{\ll}$	99.40%$^{=}$

Table 5.5: Fine-grained geolocalisation results for the New York dataset considering only the Top-1. We compare our learning to rank approach (*L2Geo*) against the baseline (*Individual* using *TF-IDF*). We report average error distance (*AED*), median error distance (*MED*), accuracy at 1 km (*Acc@1km*) and coverage (*Coverage*). We use a paired t-test to assess significant differences, denoted by $\gg$ for better performance, $\ll$ for worse performance and = for no statistical difference.

New York					
Model	***Features***	***AED(km)***↓	***MED(km)***↓	***Acc@1km***↑	***Coverage***↑
Individual	*TF-IDF*	4.972	1.251	48.46%	99.98%
L2Geo	*All*	4.207$^{\gg}$	**0.668**$^{\gg}$	**53.58%**$^{\gg}$	99.98%$^{=}$
L2Geo	*Common*	4.694$^{\gg}$	0.760$^{\gg}$	51.74%$^{\gg}$	99.98%$^{=}$
L2Geo	*Query*	5.192$^{\ll}$	1.356$^{\ll}$	47.09%$^{\ll}$	99.98%$^{=}$
L2Geo	*Doc*	6.287$^{\ll}$	3.378$^{\ll}$	34.20%$^{\ll}$	99.98%$^{=}$
L2Geo	*Query_Doc*	5.797$^{\ll}$	2.603$^{\ll}$	38.73%$^{\ll}$	99.98%$^{=}$
L2Geo	*Query_Common*	4.645$^{\gg}$	0.727$^{\gg}$	52.08%$^{\gg}$	99.98%$^{=}$
L2Geo	*Doc_Common*	**4.199**$^{\gg}$	0.671$^{\gg}$	53.19%$^{\gg}$	99.98%$^{=}$
L2Geo	*Geo_Query*	5.192$^{\ll}$	1.356$^{\ll}$	47.09%$^{\ll}$	99.98%$^{=}$
L2Geo	*Geo_Doc*	6.416$^{\ll}$	3.562$^{\ll}$	32.77%$^{\ll}$	99.98%$^{=}$
L2Geo	*Content_Query*	5.192$^{\ll}$	1.356$^{\ll}$	47.09%$^{\ll}$	99.98%$^{=}$
L2Geo	*Content_Doc*	6.986$^{\ll}$	4.281$^{\ll}$	29.84%$^{\ll}$	99.98%$^{=}$

We observe that the *L2Geo* model that incorporates all the features (*All*) shows the best performance in terms of average error distance (*AED*) in the Chicago dataset, achieving 3.835 km error. Similarly, the model that uses the *Doc_ Common* features shows to be the second best performing model with 4.157 km. On the other hand, in the New York dataset we observe the same behaviour but, in this case, the model using *Doc_ Common* features presents better performance than the model using *All* features; 4.199 km and 4.207 km respectively. Additionally, we also identify significant improvements in every model that incorporates query-dependent features (*Common*, *Query_ Common* and *Doc_ Common*). On the other hand, features extracted from the query-tweet (*Query*) shows better performance than features extracted from the query-doc (*Doc*).

Regarding the specific subsets of query-tweet and doc-tweet features, we observe that either geospecific features *Geo* and content quality features *Content* show better average error distance (*AED*) when they are considered at query-tweet level (Geo_Query and *Content_ Query*) than when they are extracted at doc-tweet level (Geo_Doc and *Content_ Doc*). This is consistent with the previous observation of the higher impact of query-tweet features over doc-tweet features. Next, we evaluate the performance when using the majority voting algorithm for selecting a predicted geolocation (*L2Geo+MV*).

5.4.2.1 Applying Majority Voting

In the previous section, we analysed the performance of the learning to rank approach on fine-grained geolocalisation when considering only the Top-1 most similar geotagged tweets. However, as we described in Chapter 4, it is beneficial to exploit the correlation between similarity and geographical distance in order to obtain better fine-grained predictions. Thus we consider the geographical evidence encoded within the Top-N geotagged tweets. Now, we apply the majority voting algorithm described in Chapter 4, Section 4.3.1, on the Top-N geotagged tweets re-ranked by our learning to rank approach. Results are presented in Tables 5.6 (Chicago) and 5.7 (New York).

Out first observation is that our approach (*L2Geo+MV*) is capable of reducing the average error distance (*AED*) while maintaining coverage (*Coverage*) when using query-dependent features (*Common*). For instance, in the Chicago

Table 5.6: Fine-grained geolocalisation results for the Chicago dataset using the majority voting algorithm. We compare our approach (*L2Geo+MV*) against the baseline (*MV*) considering the Top-3, -5, -7 and -9 most similar geotagged tweets. Table reports average error distance (*AED*), median error distance (*MED*), accuracy at 1 km (*Acc@1km*) and coverage (*Coverage*).

Chicago					
Model	***Features***	***AED(km)***↓	***MED(km)***↓	***Acc@1km***↑	***Coverage***↑
MV@Top-3	*N/A*	2.484	**0.471**	76.09%	63.15%
MV@Top-5	*N/A*	1.907	**0.471**	81.47%	54.53%
MV@Top-7	*N/A*	1.702	**0.471**	83.51%	49.99%
MV@Top-9	*N/A*	1.639	**0.471**	84.00%	46.87%
L2Geo+MV@Top-3	*All*	2.340	**0.471**	76.91%	**72.24%**
L2Geo+MV@Top-5	*All*	1.973	**0.471**	80.92%	63.81%
L2Geo+MV@Top-7	*All*	1.910	**0.471**	81.93%	59.32%
L2Geo+MV@Top-9	*All*	1.829	**0.471**	82.13%	55.28%
L2Geo+MV@Top-3	*Query*	2.484	**0.471**	76.09%	63.15%
L2Geo+MV@Top-5	*Query*	1.907	**0.471**	81.47%	54.53%
L2Geo+MV@Top-7	*Query*	1.702	**0.471**	83.51%	49.99%
L2Geo+MV@Top-9	*Query*	1.639	**0.471**	84.00%	46.87%
L2Geo+MV@Top-3	*Doc*	4.691	0.660	55.39%	62.72%
L2Geo+MV@Top-5	*Doc*	4.088	0.544	61.91%	53.19%
L2Geo+MV@Top-7	*Doc*	3.724	0.498	65.18%	48.65%
L2Geo+MV@Top-9	*Doc*	3.515	**0.471**	67.92%	45.00%
L2Geo+MV@Top-3	*Common*	2.192	**0.471**	78.65%	67.50%
L2Geo+MV@Top-5	*Common*	1.702	**0.471**	83.56%	58.58%
L2Geo+MV@Top-7	*Common*	1.519	**0.471**	85.53%	53.95%
L2Geo+MV@Top-9	*Common*	1.484	**0.471**	86.13%	50.21%
L2Geo+MV@Top-3	*Query_Doc*	3.273	0.489	67.90%	61.95%
L2Geo+MV@Top-5	*Query_Doc*	2.667	**0.471**	73.77%	53.14%
L2Geo+MV@Top-7	*Query_Doc*	2.420	**0.471**	76.99%	48.86%
L2Geo+MV@Top-9	*Query_Doc*	2.321	**0.471**	78.00%	45.88%
L2Geo+MV@Top-3	*Query_Common*	2.137	**0.471**	79.30%	67.47%
L2Geo+MV@Top-5	*Query_Common*	1.657	**0.471**	84.14%	58.80%
L2Geo+MV@Top-7	*Query_Common*	1.483	**0.471**	86.05%	54.08%
L2Geo+MV@Top-9	*Query_Common*	**1.451**	**0.471**	**86.38%**	50.47%
L2Geo+MV@Top-3	*Doc_Common*	2.364	**0.471**	76.50%	72.16%
L2Geo+MV@Top-5	*Doc_Common*	2.033	**0.471**	80.30%	63.73%
L2Geo+MV@Top-7	*Doc_Common*	1.957	**0.471**	81.41%	59.18%
L2Geo+MV@Top-9	*Doc_Common*	1.892	**0.471**	81.70%	55.31%
L2Geo+MV@Top-3	*Geo_Query*	2.484	**0.471**	76.09%	63.15%
L2Geo+MV@Top-5	*Geo_Query*	1.907	**0.471**	81.47%	54.53%
L2Geo+MV@Top-7	*Geo_Query*	1.702	**0.471**	83.51%	49.99%
L2Geo+MV@Top-9	*Geo_Query*	1.639	**0.471**	84.00%	46.87%
L2Geo+MV@Top-3	*Geo_Doc*	4.932	0.857	52.83%	64.05%
L2Geo+MV@Top-5	*Geo_Doc*	4.366	0.576	58.66%	54.75%
L2Geo+MV@Top-7	*Geo_Doc*	3.968	0.526	62.25%	49.96%
L2Geo+MV@Top-9	*Geo_Doc*	3.815	0.498	64.05%	47.52%
L2Geo+MV@Top-3	*Content_Query*	2.484	**0.471**	76.09%	63.15%
L2Geo+MV@Top-5	*Content_Query*	1.907	**0.471**	81.47%	54.53%
L2Geo+MV@Top-7	*Content_Query*	1.702	**0.471**	83.51%	49.99%
L2Geo+MV@Top-9	*Content_Query*	1.639	**0.471**	84.00%	46.87%
L2Geo+MV@Top-3	*Content_Doc*	4.089	0.529	60.56%	50.88%
L2Geo+MV@Top-5	*Content_Doc*	3.354	**0.471**	68.26%	42.81%
L2Geo+MV@Top-7	*Content_Doc*	3.064	**0.471**	72.09%	38.91%
L2Geo+MV@Top-9	*Content_Doc*	2.840	**0.471**	73.17%	36.57%

Table 5.7: Fine-grained geolocalisation results for the New York dataset using the majority voting algorithm. We compare our approach (*L2Geo+MV*) against the baseline (*MV*) considering the Top-3, -5, -7 and -9 most similar geotagged tweets. Table reports average error distance (*AED*), median error distance (*MED*), accuracy at 1 km (*Acc@1km*) and coverage (*Coverage*).

		New York			
Model	***Config***	***AED(km)***↓	***MED(km)***↓	***Acc@1km***↑	***Coverage***↑
MV@Top-3	*N/A*	2.522	0.461	72.76%	55.31%
MV@Top-5	*N/A*	1.878	0.428	79.85%	46.20%
MV@Top-7	*N/A*	1.610	0.412	82.52%	41.67%
MV@Top-9	*N/A*	1.448	0.405	84.00%	38.70%
L2Geo+MV@Top-3	*All*	2.402	0.457	73.06%	**64.17%**
L2Geo+MV@Top-5	*All*	1.925	0.434	78.05%	54.79%
L2Geo+MV@Top-7	*All*	1.740	0.418	80.12%	49.74%
L2Geo+MV@Top-9	*All*	1.571	0.412	81.55%	46.30%
L2Geo+MV@Top-3	*Query*	2.522	0.461	72.76%	55.31%
L2Geo+MV@Top-5	*Query*	1.878	0.428	79.85%	46.20%
L2Geo+MV@Top-7	*Query*	1.610	0.412	82.52%	41.67%
L2Geo+MV@Top-9	*Query*	1.448	0.405	84.00%	38.70%
L2Geo+MV@Top-3	*Doc*	4.019	0.668	55.54%	52.30%
L2Geo+MV@Top-5	*Doc*	3.250	0.496	63.89%	43.05%
L2Geo+MV@Top-7	*Doc*	2.881	0.463	68.04%	38.94%
L2Geo+MV@Top-9	*Doc*	2.610	0.441	70.86%	36.56%
L2Geo+MV@Top-3	*Common*	2.305	0.446	75.22%	59.21%
L2Geo+MV@Top-5	*Common*	1.656	0.416	81.64%	49.55%
L2Geo+MV@Top-7	*Common*	1.495	0.412	83.50%	45.02%
L2Geo+MV@Top-9	*Common*	1.336	**0.404**	**84.98%**	41.60%
L2Geo+MV@Top-3	*Query_Doc*	3.222	0.506	63.91%	52.61%
L2Geo+MV@Top-5	*Query_Doc*	2.595	0.457	70.90%	43.64%
L2Geo+MV@Top-7	*Query_Doc*	2.256	0.434	74.60%	39.49%
L2Geo+MV@Top-9	*Query_Doc*	2.146	0.420	76.25%	37.20%
L2Geo+MV@Top-3	*Query_Common*	2.294	0.450	75.09%	59.48%
L2Geo+MV@Top-5	*Query_Common*	1.688	0.420	81.27%	50.02%
L2Geo+MV@Top-7	*Query_Common*	1.488	0.412	83.35%	45.25%
L2Geo+MV@Top-9	*Query_Common*	**1.319**	0.406	84.93%	41.85%
L2Geo+MV@Top-3	*Doc_Common*	2.358	0.456	73.53%	63.35%
L2Geo+MV@Top-5	*Doc_Common*	1.858	0.429	78.84%	53.86%
L2Geo+MV@Top-7	*Doc_Common*	1.674	0.414	80.87%	48.92%
L2Geo+MV@Top-9	*Doc_Common*	1.510	0.412	82.35%	45.27%
L2Geo+MV@Top-3	*Geo_Query*	2.522	0.461	72.76%	55.31%
L2Geo+MV@Top-5	*Geo_Query*	1.878	0.428	79.85%	46.20%
L2Geo+MV@Top-7	*Geo_Query*	1.610	0.412	82.52%	41.67%
L2Geo+MV@Top-9	*Geo_Query*	1.448	0.405	84.00%	38.70%
L2Geo+MV@Top-3	*Geo_Doc*	4.257	0.675	53.75%	52.87%
L2Geo+MV@Top-5	*Geo_Doc*	3.517	0.524	61.29%	43.33%
L2Geo+MV@Top-7	*Geo_Doc*	3.105	0.479	65.12%	38.96%
L2Geo+MV@Top-9	*Geo_Doc*	2.847	0.465	67.65%	36.78%
L2Geo+MV@Top-3	*Content_Query*	2.522	0.461	72.76%	55.31%
L2Geo+MV@Top-5	*Content_Query*	1.878	0.428	79.85%	46.20%
L2Geo+MV@Top-7	*Content_Query*	1.610	0.412	82.52%	41.67%
L2Geo+MV@Top-9	*Content_Query*	1.448	0.405	84.00%	38.70%
L2Geo+MV@Top-3	*Content_Doc*	3.785	0.568	58.76%	43.96%
L2Geo+MV@Top-5	*Content_Doc*	2.783	0.457	69.25%	34.24%
L2Geo+MV@Top-7	*Content_Doc*	2.369	0.430	73.79%	31.22%
L2Geo+MV@Top-9	*Content_Doc*	2.117	0.412	76.54%	29.31%

dataset, when considering the Top-9 geotagged tweets, the average error distance is reduced from 1.639 km (*MV@Top-9*) to 1.484 km (*L2Geo+MV@Top-9* using *Common*), and coverage is improved from 46.87% to 50.21%. Furthermore, adding the query-tweet features to the query-dependent features (*Query_ Common*) improve performance further. The average error distance (*AED*) is reduced from 1.484 km to 1.452 km, and coverage is increased from 50.21% to 50.47%. Interestingly, when considering all the features (*L2Geo* using *All*) we achieve better performance in terms of average error distance (*AED*) when using the Top-1 (see Tables 5.4 and 5.5), but the average error distance (*AED*) is not reduced when applying the majority voting on any of the Top-N geotagged tweets.

Table 5.8: Best performing models in terms of average error distance (*AED*) over values of $N \in \{3, 5, 7, 9, 11, ..., 49\}$ for the Top-N in the Chicago dataset.

Best Performing Models (Chicago)					
Model	***Features***	***AED(km)*↓**	***MED(km)*↓**	***Acc@1km*↑**	***Coverage*↑**
MV@Top-35	*N/A*	1.490	**0.471**	86.32%	31.88%
L2Geo+MV@Top-13	*All*	1.758	**0.471**	83.03%	50.39%
L2Geo+MV@Top-35	*Query*	1.490	**0.471**	86.32%	31.88%
L2Geo+MV@Top-47	*Doc*	2.285	**0.471**	80.48%	30.37%
L2Geo+MV@Top-21	*Common*	1.465	**0.471**	86.48%	40.16%
L2Geo+MV@Top-25	*Query_ Doc*	1.955	**0.471**	82.02%	36.76%
L2Geo+MV@Top-13	*Query_ Common*	**1.441**	**0.471**	**86.53%**	46.01%
L2Geo+MV@Top-13	*Doc_ Common*	1.826	**0.471**	82.64%	**50.40%**
L2Geo+MV@Top-35	*Geo_ Query*	1.490	**0.471**	86.32%	31.88%
L2Geo+MV@Top-49	*Geo_ Doc*	2.175	**0.471**	80.30%	30.43%
L2Geo+MV@Top-35	*Content_ Query*	1.490	**0.471**	86.32%	31.88%
L2Geo+MV@Top-49	*Content_ Doc*	1.907	**0.471**	82.47%	26.53%

Regarding the best performing models, we present in Tables 5.8 (Chicago) and 5.9 (New York) the best performing configurations for our *L2Geo+MV* models against the baseline (*MV*), across values of $N \in \{3, 5, 7, 9, 11, ..., 49\}$, for the Top-N for the Chicago and New York datasets respectively. Consistently with previous observations, *Query_ Common* features exhibit the best performance in terms of average error distance (*AED*) when using the Top-13 geotagged tweets in Chicago and the Top-41 in New York. In both datasets, using *Query_ Common* features, our learning to rank approach is capable of reducing the average error distance (*AED*) while increasing coverage, compared to the baseline (*MV*). This suggests that by improving the ranking using our approach, we can increase the

Table 5.9: Best performing models in terms of average error distance (*AED*) over values of $N \in \{3, 5, 7, 9, 11, ..., 49\}$ for the Top-N in the New York dataset.

Best Performing Models (New York)					
Model	***Features***	***AED(km)***↓	***MED(km)***↓	***Acc@1km***↑	***Coverage***↑
MV@Top-47	*N/A*	1.121	0.386	87.57%	23.97%
L2Geo+MV@Top-29	*All*	1.269	0.393	84.75%	**33.41%**
L2Geo+MV@Top-47	*Query*	1.121	0.386	87.57%	23.97%
L2Geo+MV@Top-49	*Doc*	1.613	0.391	81.85%	25.59%
L2Geo+MV@Top-47	*Common*	1.081	**0.381**	88.40%	25.33%
L2Geo+MV@Top-49	*Query_ Doc*	1.515	0.392	82.41%	25.97%
L2Geo+MV@Top-41	*Query_ Common*	**1.080**	**0.381**	**88.42%**	26.63%
L2Geo+MV@Top-29	*Doc_ Common*	1.276	0.393	85.23%	33.02%
L2Geo+MV@Top-47	*Geo_ Query*	1.121	0.386	87.57%	23.97%
L2Geo+MV@Top-49	*Geo_ Doc*	1.696	0.392	81.08%	25.63%
L2Geo+MV@Top-47	*Content_ Query*	1.121	0.386	87.57%	23.97%
L2Geo+MV@Top-49	*Content_ Doc*	1.497	0.388	83.26%	22.78%

number of fine-grained predictions, which supports the main hypothesis of this chapter introduced in Section 5.1.

Furthermore, we observe in Tables 5.8 (Chicago) and 5.9 (New York) that, in terms of median error distance (*MED*), the best performance is given by all the models with 0.471 km in the Chicago dataset. Likewise, in the New York dataset almost all the models achieve similar performance. Specifically, the median error distance (*MED*) ranges between 0.381 km to 0.393 km. This behaviour can be explained by the shape of the distribution of the error distance, which is skewed towards the lowest error distances (i.e., less than 1 km). For example, this can be noticed in Figure 5.1 which shows the error distance distributions of the best performing model (*L2Geo+MV@Top-13* model using *Query_ Common*) and the baseline (*MV@Top-35*).

In both distributions, most of the predictions fall within 0 km and 1 km error distance. However our learning to rank approach (*L2Geo+MV@Top-13* using *Query_ Common*) is capable of predicting a higher number of fine-grained predictions (1 km or less) than the baseline (*MV@Top-35*) for the Chicago dataset. This is reflected in the trade-off in performance between accuracy at 1 km (*Acc@1km*) and coverage. For example, in the Chicago dataset, the *L2Geo+MV@Top-13* using *Query_ Common* model obtains 86.53% of accuracy and coverage of 46.01%, whereas the baseline (*MV@Top-35*) achieves 86.32% but with lower coverage of

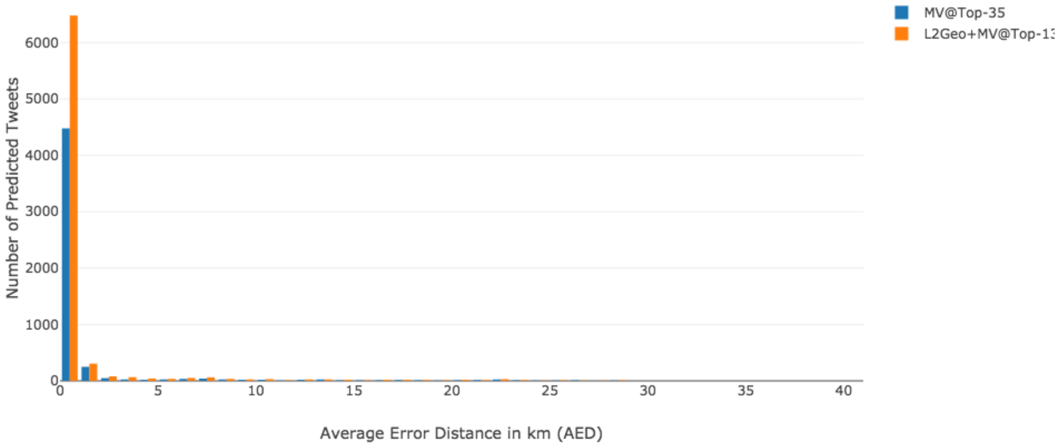

Figure 5.1: Distribution of average error distance (*AED*) for the Chicago dataset. The figure presents the distribution of our best performing learning to rank approach (*L2Geo+MV@Top-13* using *Query_ Common*) and the baseline (*MV@Top-47*).

31.88%.

Additionally, in order to see the best performing configuration we present the performance of the models when considering the Top-N most similar geotagged tweets with values of $N \in \{3, 5, 7, 9, 11, ..., 49\}$. We show the performance in terms of average error distance (*AED*) and coverage (*Coverage*) respectively for the Chicago dataset in Figures 5.2 and 5.3, and the New York dataset in Figures 5.4 and 5.5. Analysing such figures, we identify that *L2Geo* using *Common* and *Query_ Common* outperforms every model (including the baseline, IDF weighting) in average error distance while maintaining a better trade-off with respect to coverage than other models.

The set of results discussed above address research question **RQ-5.2** and supports the central hypothesis of this chapter, which states that by improving the ranking we can also improve the performance of fine-grained geolocalisation. Now, we investigate the most useful features for fine-grained geolocalisation using our learning to rank approach.

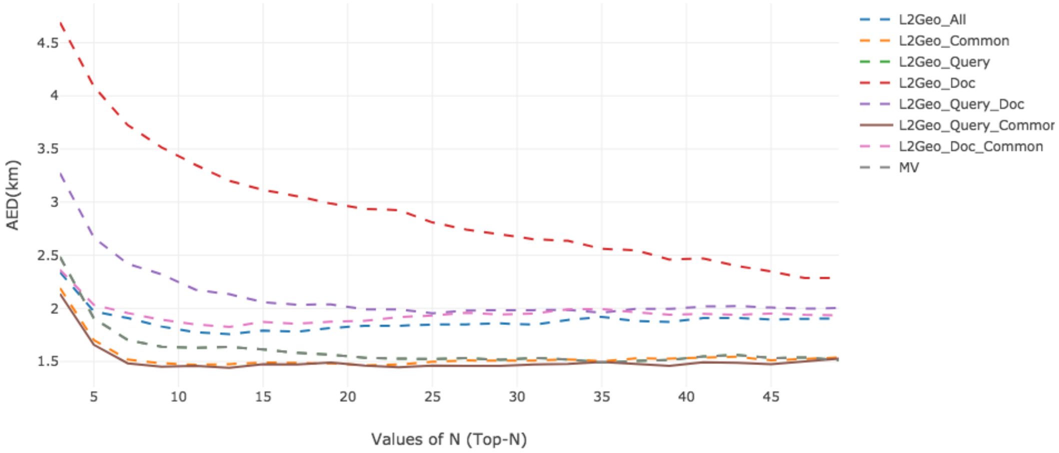

Figure 5.2: (Chicago Dataset) Distribution of average error distance (*AED*) over the values of $N \in \{3, 5, 7, 9, 11, ..., 49\}$ for the Top-N geotagged tweets considered by the majority voting algorithm. Bold line represents the best performing model (*L2Geo* using *Query_ Common*)

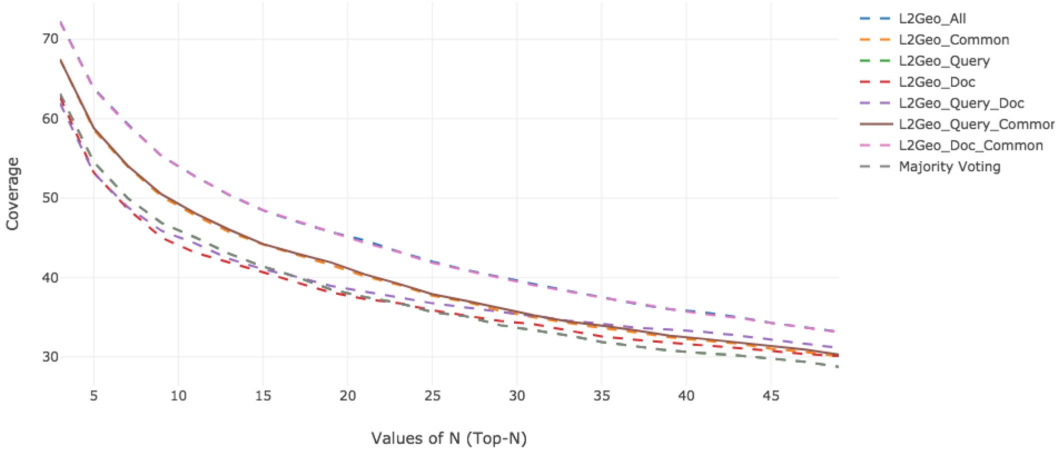

Figure 5.3: (Chicago Dataset) Distribution of coverage (*Coverage*) over the values of $N \in \{3, 5, 7, 9, 11, ..., 49\}$ for the Top-N geotagged tweets considered by the majority voting algorithm. Bold line represents the best performing model (*L2Geo* using *Query_ Common*)

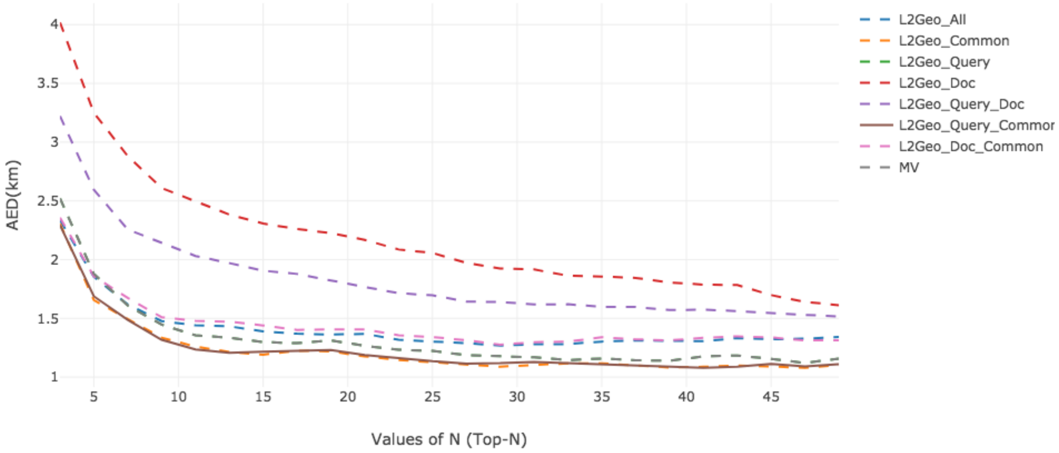

Figure 5.4: (New York Dataset) Distribution of average error distance (AED) over the values of $N \in \{3, 5, 7, 9, 11, ..., 49\}$ for the Top-N geotagged tweets considered by the majority voting algorithm. Bold line represents the best performing model (*L2Geo* using *Query_ Common*)

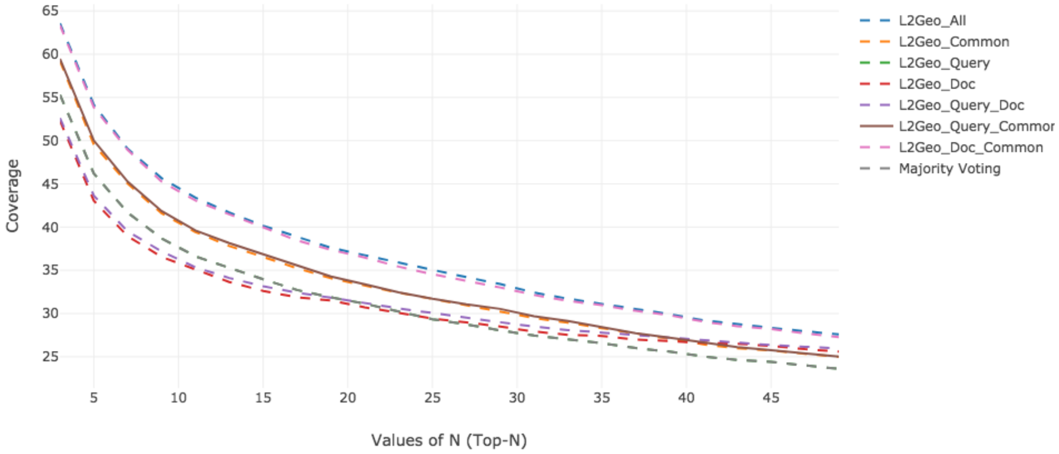

Figure 5.5: (New York Dataset) Distribution of coverage ($Coverage$) over the values of $N \in \{3, 5, 7, 9, 11, ..., 49\}$ for the Top-N geotagged tweets considered by the majority voting algorithm. Bold line represents the best performing model (*L2Geo* using *Query_ Common*)

5.4.2.2 Best Features for Fine-Grained Geolocalisation

In this section, we address research question **RQ-5.3**. First, we identify the best performing subset of features. Next, we study the individual impact of each of the features belonging to that subset.

Analysing Tables 5.8 and 5.9 for the Chicago and New York datasets respectively, we conclude that the query-dependent features (*Common*), described in Section 5.2.1.2, are the most impactful. We observe that all the models that incorporate *Common* features show improvements over the baseline (*IDF*), and outperforms other models that use any other subset of features. Besides, we observe this behaviour in all the models that combines *Common* features; *Query_ Common*, *Doc_ Common* and *All*, being *Query_ Common* features the best performing one. On the other hand, features extracted from the query-tweet (*Query*) shows better performance than features extracted from the the doc-tweet *Doc*, which exhibit the worst performance overall.

Note that *Query* and *Doc* features are extracted at document and query level alone. However, *Common* features models the relationship between the query and the document. This suggests that the information shared between tweets is the best indicator for fine-grained geolocalisation, which means that high related tweets are posted within the same fine-grained area.

Table 5.10: Best performing models in terms of average error distance (*AED*) over values of $N \in \{3, 5, 7, 9, 11, ..., 49\}$ for the Top-N in the Chicago dataset. We train single-feature model for each of the features belonging to the *Common* set, described in Section 5.3.2, to study their predictive power.

Best Performing Models (Chicago)					
Model	***Features***	***AED(km)***↓	***MED(km)***↓	***Acc@1km***↑	***Coverage***↑
MV@Top-35	*N/A*	1.490	**0.471**	86.32%	31.88%
L2Geo+MV@Top-35	*Common_ Hashtags*	1.535	**0.471**	85.61%	31.59%
L2Geo+MV@Top-23	*Common_ Hour*	1.399	**0.471**	**87.23%**	34.69%
L2Geo+MV@Top-35	*Common_ Mentions*	1.486	**0.471**	86.40%	32.02%
L2Geo+MV@Top-35	*Common_ Score*	1.517	**0.471**	85.96%	31.99%
L2Geo+MV@Top-21	*Common_ User*	1.475	**0.471**	86.67%	**39.17%**
L2Geo+MV@Top-35	*Common_ Weekday*	**1.350**	**0.471**	**87.23%**	30.55%

To study the individual predictive power of the *Common* features, for each of the features we train a single-feature model and evaluate its performance on fine-grained geolocalisation. Tables 5.10 and 5.11 present results for the best

Table 5.11: Best performing models in terms of average error distance (AED) over values of $N \in \{3, 5, 7, 9, 11, ..., 49\}$ for the Top-N in the New York dataset. We train single-feature model for each of the features belonging to the *Common* set, described in Section 5.3.2, to study their predictive power.

Best Performing Models (New York)					
Model	***Features***	***AED(km)***↓	***MED(km)***↓	***Acc@1km***↑	***Coverage***↑
MV@Top-47	*N/A*	1.121	0.386	87.57%	23.97%
L2Geo+MV@Top-47	*Common_Hashtags*	1.167	0.386	87.04%	24.26%
L2Geo+MV@Top-21	*Common_Hour*	1.046	0.385	88.55%	**29.13%**
L2Geo+MV@Top-47	*Common_Mentions*	1.121	0.386	87.56%	24.00%
L2Geo+MV@Top-47	*Common_Score*	1.120	0.386	87.76%	23.96%
L2Geo+MV@Top-47	*Common_User*	1.127	0.386	87.57%	24.66%
L2Geo+MV@Top-29	*Common_Weekday*	**0.973**	**0.380**	**89.70%**	25.15%

performing configuration of the models for the Chicago and New York dataset, respectively. First, we observe that all the features contribute similarly. However, we identify that *Common_Hour* and *Common_ Weekday* provides slightly better performance than other features in terms of average error distance (AED) and accuracy (*Acc@1km*). This means that the posting time of the tweets is a good indicator of its geolocation, which suggests that seasonal events are occurring in certain locations. On the other hand, we observe that *Common_Hashtags*, *Common_Mentions*, *Common_ User* and *Common_Score* also contribute positively to fine-grained geolocalisation, which indicates that tweets posted in a certain geolocation share the same content.

The findings described above address **RQ-5.2** and support our thesis statement introduced in Chapter 1 (Section 1.2), and discussed in Chapter 4 (Section 4.1), that postulate an existing correlation between similarity and geographical distance at a fine-grained level.

5.4.3 Comparing Behaviour Across Datasets

In this section, we explore the similarities in the behaviour of our approach across the two datasets, Chicago and New York. Particularly, we observe the same patterns in behaviour observed in the Chapter 4 approach, and described in Section 4.5.3.

In summary, we observe in Tables 5.6 and 5.7 that, as we increase the values of N for the Top-N geotagged tweets, our approach is capable of decrease average error distance (*AED*) and accuracy at 1 km (*Acc@1km*), along with a decrease in coverage (*Coverage*). Identically, in Tables 5.8 and 5.9, we note that our approach can reduce *AED* and increase *Acc@1km* using the best configuration (*L2Geo+MV@Top-9* using *Query_ Common* features), while increasing Coverage with respect to the baseline (*MV@Top-9*). Moreover, similarly to Section 4.5.3, we also notice a variation in performance in terms of *AED* and *Coverage* in both datasets, which can be explained by the difference in size between the Chicago and New York datasets - where New York contains a bigger number of test tweets than Chicago. Due to this, our approach is capable of predicting a location for a lower ratio of tweets in the New York dataset, which can explain the lower *Coverage* and *AED* achieved by our models.

In addition to previous observations, we observe in Tables 5.2 and 5.2, for Chicago and New York respectively, that LambdaMART is the best learning to rank algorithm for the ranking, outperforming other algorithms and the baseline (*IDF*). Moreover, *Query_ Common* shows to be the most useful feature set for fine-grained geolocalisation in both datasets. This consistency in behaviour across datasets suggests that our approach can be adapted to other cities, and supports the generalisation of our approach.

5.5 Conclusions

In Chapter 4 of this thesis, we explored the correlation between similarity and geographical distance between tweets in order to provide fine-grained geolocalisation. First, we based our approaches on a ranking of the Top-N most similar individual geotagged tweets to then exploit their geographical characteristics using a majority voting algorithm. However, this ranking is performed taking into account only the document frequency of terms (IDF weighting), which may limit the quality of the tweets in the ranking. Specifically, the quality of the Top-N geotagged tweets that are fed into the majority voting. In this chapter, we hypothesised that by improving the ranking of geotagged tweets, we could improve performance and obtain a higher number of fine-grained predictions. In order to improve the ranking, we adopted a learning to rank approach (see Section 5.2) and proposed a set of features for fine-grained geolocalisation. Our learning to rank models are trained based on the probability that a pair of tweets are posted within the same fine-grained area (i.e., squared areas of size length 1 km). We focused the work in this chapter on three research questions **RQ-5.1**, **RQ-5.2** and **RQ-5.3**, introduced in Section 5.1.

In order to asses the effectiveness of our learning to rank approach, we created training and testing sets for learning to rank by labelling as positive instances pairs of tweets that are located at 1km distance from each other. On the other hand, pairs of tweets that are posted at more than 1km distance from each other are labelled as negative instances. Then, we extracted features, described in Section 5.2.1, at document level (doc-tweet), query level (query-tweet) and query-dependent features that model the relation between the query-tweet and the doc-tweet. The full list of features is presented in Table 5.1.

Firstly, we experiment with different algorithms for our learning to rank approach for improving the ranking of the Top-N geotagged tweets, which aimed to answer research question **RQ-5.1**. We experimented with a set of algorithms that covers the three main state-of-the-art categories: point-wise, pair-wise and list-wise approaches. In total, we compared six different algorithms, including: MART (Friedman, 2001), Random Forests (Breiman, 2001), RankNet (Burges et al., 2005), LambdaMART (Wu et al., 2010), AdaRank (Xu and Li, 2007) and ListNet (Cao et al., 2007). We trained each algorithm optimising NDCG@3, @5, @10, @20, @30, @40 and @50, and evaluate performance at the Top-N geotagged tweets in the rank with N values of 1 (NDCG@1), 3 (NDCG@3), 5 (NDCG@5) and 10 (NDCG@10). Tables 5.2 and 5.3 shows the performance of the algorithms. As a result of this experiment, we identified LambdMART, trained at NDCG@10 in the Chicago dataset and NDCG@30 in the New York dataset, as the best performing learning to rank algorithm for ranking.

Secondly, we compared the effectiveness on fine-grained geolocalisation of our learning to rank approach when always returning the Top-1 most similar geotagged tweet (*L2Geo*) as the predicted location, against the baseline explored in Chapter 3 (*Individual*). Results are presented in Tables 5.4 and 5.5. We observed that the learning to rank approaches that incorporate the query-dependent features (*Common*) significantly outperformed the baseline. Overall, the best performing model in terms of average error distance (*AED*) combines all the features (*L2Geo* using *All*).

Furthermore, in Section 5.4.2.1 we assessed the effectiveness of our learning to rank approach in terms of the quality of the geographical evidence within the Top-N re-ranked geotagged tweets by applying a majority voting algorithm (*L2Geo+MV*) with values of $N \in \{3, 5, 7, 9, ..., 49\}$. We compared against the baseline proposed in Chapter 4 (*MV*). These results are presented in Tables 5.6 and 5.7. We observed that, as the values of N increased, we achieved a lower average error distance (*AED*) with improved coverage (*Coverage*) compared to the baseline. Additionally, in Figures 5.2 and 5.3 for Chicago, and Figures 5.4 and 5.5 for New York, we identified that our learning to rank approach using query-dependent features combined with features extracted from the query-tweet (*L2Geo+MV*) using *Query_ Common*) outperforms every model (included the baseline) in average error distance. Moreover, combining all the features led to a lower performance regarding average error distance (*AED*) when considering the Top-N geotagged tweets, in contrast to results when considering only the Top-1 geotagged tweet. These results address research question **RQ-5.2** and support the hypothesis that by improving the ranking of the Top-N most similar geotagged tweets (*L2Geo*), the performance of fine-grained geolocalisation is also improved.

Finally, observing the behaviour of our learning to rank models over previous results, we concluded that query-dependent features along with features extracted from the query-tweet (*Query_ Common*) are the most informative for fine-grained geolocalisation, which address research question **RQ-5.3**.

In summary, in this chapter, we have demonstrated that by improving the ranking of the Top-N geotagged tweets leads to a better performance of fine-grained geolocalisation, and we can obtain a higher number of fine-grained predictions. We achieved an average error distance (*AED*) of 1.441 km in Chicago (Table 5.8), and 1.080 km in New York (Table 5.9), which improves previous approaches explored in this thesis. Also, we aimed to reduce the average error distance along with an increase of coverage. These results support the main hypothesis of this chapter (introduced in Section 5.1). Also, we have contributed with an approach that is capable of predicting a high number of tweet at a fine-grained level, with an 86.56% of accuracy (*Acc@1km*) in Chicago and 88.42% of accuracy in New York (See Tables 5.8 and 5.9).

In the remainder of this thesis, we demonstrate the applicability of our fine-grained geolocalisation approach, developed throughout this thesis, in a practical application – traffic incident detection. Lastly, we provide concluding remarks of our work and present, future works and discusses the new research lines that this work opens to the community in Chapter 7.

Part III
Applicability

Chapter 6

Effectiveness of Fine-Grained Geolocalised Tweets

6.1 Introduction

In the previous chapters, we tackled the problem of inferring the geolocalisation of tweets at a fine-grained level of granularity. As a result of our research, we developed a fine-grained geolocalisation method that is based on a learning to rank approach for ranking geotagged tweets, and a majority voting algorithm to exploit the geographical evidence of the geotagged tweets. Our experimental results showed that our approach is capable of predicting the location of tweets at almost 1 km distance (See Tables 5.6 and 5.7), which represents an approximate area of 3.14 km^2. On the other hand, several real-world applications use geotagged Twitter data for their analysis. In this chapter, we use the traffic incident detection task as a case study, which aims to use Twitter as a data source for the detection of traffic incidents occurring in a city.

There are currently several examples of the use of Twitter data for traffic incident detection (Cui et al., 2014; D'Andrea et al., 2015; Gu et al., 2016; Kosala et al., 2012; Mai and Hranac, 2013; Schulz et al., 2013b; Steiger et al., 2014). These works focused on scrutinising the Twitter stream to obtain tweets with content containing information about traffic conditions and disruptions. However, traffic incidents occur in very fine-grained areas: roads or highways. Thus, it is not only essential to identify traffic incident-related content in a tweet, but also it is crucial to know the precise location of the tweets in order to acknowledge an incident reliably. However, as mentioned in Chapter 1 only 1% of the Twitter data is finely-grained geotagged (Graham et al., 2014), so that the sample sizes are quite limited for real-time incident detection.

In this chapter, we hypothesise that by geolocalising non-geotagged tweets we can obtain a representative sample of geotagged data and, therefore, improve the effectiveness on the traffic incident detection task (see **Hypothesis 4** in Section 1.2). In this chapter, we aim to explore the usefulness of our fine-grained geolocalisation method when applied in the traffic incident detection pipeline and provide evidence of whether the task can be improved by enhancing the geographic details of non-geotagged data, compared to what would be supported by geotagged data alone. In this chapter, we aim to answer the following research questions:

- **RQ-6.1** What is the effectiveness of the geolocalised traffic incident-related tweets on the traffic incident detection task?
- **RQ-6.2** Does expanding the sample of geotagged tweets with new geolocalised data improve the performance of the traffic incident detection task?

The chapter is organised as follows: in Section 6.2, we discuss the main issues motivating our research approach. In Section 6.3, we describe the datasets used for testing the traffic incident detection task. Section 6.4 we build and evaluate a text classifier to identify traffic incident-related content in tweets. Next, in Section 6.5 we evaluate our fine-grained geolocalisation method and select the best configurations to apply in the traffic incident detection pipeline. In Section 6.6, we describe the traffic incident detection pipeline that integrates our fine-grained geolocalisation method and discuss the evaluation metrics of the task. Finally, in Section 6.7 we present our experimental results and describe the performance of the traffic incident detection task with the enhanced sample of geolocalised tweets. Finally, we provide concluding remarks in Section 6.9.

6.2 Background

Popular social media such as Twitter and other sources can reveal not only historical travel patterns but also real-time traffic incidents and events. The unstructured nature of the data and the level of noise involved in inferring knowledge can pose significant challenges to their routine use in transportation operations. One major area of interest in the transportation community is automated incident detection on roadways. This task depends on a wide variety of fixed sensors (inductive loop detection systems, CCTV) and moving-object sensors (probe vehicles, transit vehicles, cellphone users) and primarily covers the detection of events that disrupt efficient traffic operations. Typically in urban areas, roadways tend to be instrumented by fixed sensors, while lower level arterial and side streets which are not as well equipped with infrastructure-based sensors are monitored by moving objects and other ad-hoc sources. This detection infrastructure is expensive and does not cover the road network completely; thus there are areas where real-time detection is not possible. Additionally, the sensors provide information about an anomaly in a road, but can not provide any context information about the event.

Because Twitter data is ubiquitous and provide first-hand real-time reports of the events provided by the users, it has attracted the attention of transportation managers to be used as an alternative data source for transportation operations (traffic incident detection) (Gu et al., 2016; Mai and Hranac, 2013). Detecting small-scale road incidents using Twitter data has now been studied by many researchers, but the problems of detection rates are pertinent research issues. In the early days of using georeferenced tweets in the detection of traffic events, only geotagged tweets are used due to high spatial granularity. Nevertheless, only about 1% of tweets are geotagged, and geotagged tweets are much more heterogeneously distributed than the overall population (Graham et al., 2014). This means that an extremely limited number of georeferenced tweets are potentially useful in the detection of traffic events with fine-grained occurrence locations.

To detect traffic events by exploiting social media, some studies used both geotagged tweets and geolocalised tweets and found more tweets than using geotagged tweets alone (Cui et al., 2014; D'Andrea et al., 2015; Gu et al., 2016; Kosala et al., 2012; Mai and Hranac, 2013; Schulz et al., 2013b; Steiger et al., 2014). Most of earlier studies on geolocalisation of tweets had limitations in either the precision of the spatial resolution recovered or the number of non-geotagged tweets for which location is estimated. Some studies geolocalised tweets at the nation or city level (Eisenstein et al., 2010a; Han and Cook, 2013; Kinsella et al., 2011; Schulz et al., 2013a). Thus, a more precise geolocalisation method is needed to provide new fine-grained geolocalised data. In this chapter, we aim to integrate our fine-grained geolocalisation approach and understand its effectiveness into the traffic incident detection pipeline using Twitter data.

One approach to addressing the above problems is to increase the sample size of tweets with precisely known geographical location. Having a larger sample of geographically located tweets would help in exploring the overall representativeness and event coverage associated with geotagged data. In this chapter, we try to retrieve new finely-grained geolocalised incident-related tweets by using the fine-grained geolocalisation approaches proposed in this thesis. Then we use both these geolocalised tweets as well as the geotagged tweets to assess their comparative performance in the detection of traffic incidents in a metropolitan area.

6.3 Data

In this chapter, we study the Chicago metropolitan region. The area is defined by a bounding box with the following longitude/latitude coordinates: -86.8112, 42.4625, -88.4359, 41.2845. We show this area later in Figure 6.1. To conduct our experiments, we collected Twitter data and traffic incident data for a period of study of a month (July 2016).

6.3.1 Twitter Data

We collect Twitter data from the Twitter Public Streaming API [1]. Spatial filtering can be applied to obtain tweets from a specific area. Geographical information is attached to tweets in two ways: (i) exact longitude and latitude if the GPS location reporting of the user device is activated (*geotagged*); and (ii) as a suggested area (bounding box) from a list that can be extrapolated to a polygon, when sending a tweet (*geobounded*). In this work, we use *geotagged* and *geobounded* tweets for our experiments (see Table 6.1). The geobounded data provides a coarse-grained location but not the spatial precision (fine-grained level) needed for the types of applications considered in this chapter. We perform the geolocalisation on geobounded tweets. Since this work explores a practical way to go beyond

[1] https://dev.twitter.com/streaming/overview

geotagged data, we use geobounded tweets for exemplification of the limitations of using geotagged tweets alone.

Table 6.1: Number of geotagged tweets and geobounded tweets (July 2016).

	Total Tweets
Geotagged Tweets	160,634
Geobounded Tweets	1,888,683

Next, in order to build our fine-grained geolocalisation method and evaluate the traffic incident detection task, we divide the dataset into three subsets. First, we obtain the tweets posted during the first four weeks of July 2016 for training (1st July to 25th July), denoted by *Training Period*, and randomly divide the last week (25th July to 1st August), denoted by *Testing Period*, into validation and testing. Also, we obtain the geobounded tweets posted during the testing period as an example of non-geotagged tweets for applying our fine-grained geolocalisation approaches. Table 6.2 shows the number of tweets for each subset.

Table 6.2: Number of geotagged and geobounded tweets distributed between training, validation and testing.

	Training Period	**Testing Period**	
	Training	*Validation*	*Testing*
Geotagged Tweets	120,503	20,169	19,962
Geobounded Tweets	-	-	459,233

For convenience and reference, we now introduce basic terminology that will be used in the rest of this chapter:

- **Geotagged:** we refer to the set of tweets with longitude/latitude GPS coordinates attached. This set is already available in the Twitter stream and represents approximately 1% of the whole stream (Graham et al., 2014).
- **Geobounded:** we refer to the set of tweets available in the Twitter stream that are not finely-grained located, but instead a bounding box that represents at a coarse-grained area is attached to them.

- **Geolocalised:** we refer to the set of tweets with attached geolocation that has been inferred using one of our geolocalisation approaches (see Section 6.5).

- **Georeferenced:** we refer to the set of tweets that have geographical information available. This set represents the union of geotagged tweets and geolocalised tweets.

6.3.2 Traffic Incident Data

Both geotagged and geolocalised tweets were compared against a ground truth dataset containing traffic crashes within the City of Chicago limits reported by the Chicago Police Department (CPD). The dataset is publicly available at the City of Chicago open data portal[1]. Specifically, we extract traffic crashes which occurred in Chicago during the *Testing Period* of study (25th July to 1st August 2016), presented before in Table 6.2.

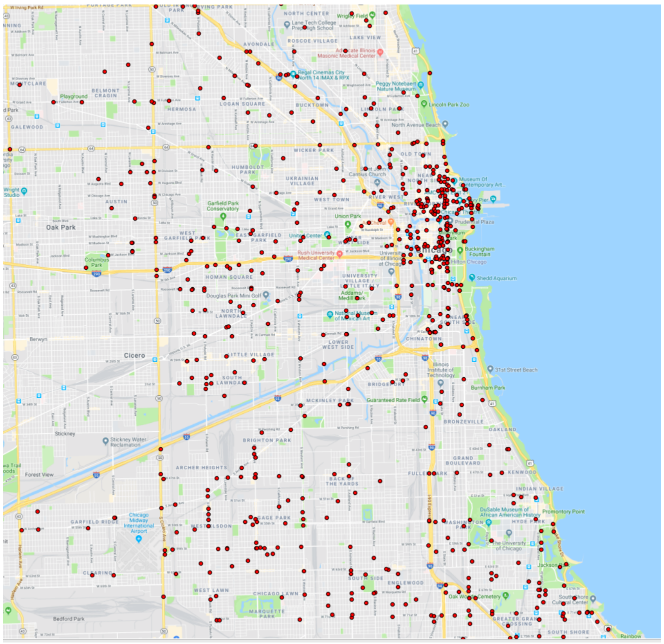

Figure 6.1: Geographical distribution of traffic crashes (N=886) in the city Chicago during the testing period (25th July to 1st August 2016).

[1] https://data.cityofchicago.org/Transportation/Traffic-Crashes-Crashes/85ca-t3if

In total, we obtain 886 traffic crashes that occurred within the city of Chicago during the testing period defined in the previous section (25th July to 1st August). Each traffic crash provides information about the incident and the vehicle, but for the purpose of this chapter, we only consider the location of the incident (longitude and latitude point) and the time of the event. Figure 6.1 shows a map with the locations of the 886 traffic crashes that occurred during the testing period (25th July to 1st August 2016).

6.4 Classification of Incident-Related Tweets

In this section, we introduce our approach for identifying traffic crash related tweets from the Twitter dataset. Inspired by previous work (D'Andrea et al., 2015; Gu et al., 2016; Schulz et al., 2013b), we build a text classifier to determine whether the content of the tweets is related to a traffic crash or not. Firstly, we describe our ground truth of human labelled traffic crash tweets on which we build our classifier. Then, we present the performance of different classifiers using different algorithms and select the best performing one for application on the traffic incident detection task.

6.4.1 Incident-Related Twitter Dataset

We use a gold standard dataset[1] generated by Schulz et al. (2013b, 2017), that contains human labelled tweets from a wide range of cities. In particular, we use the available tweets from Chicago for building our incident-related tweet classifier. Originally, the dataset is composed of 1,483 tweets posted from January 2014 to March 2014 in a 15 km radius around the city centre of Chicago. The tweets are annotated by humans and labelled as "crash" for tweets about traffic crashes, "fire" for tweets about fires in buildings, "shooting" for crime incidents involving guns shooting, and "NO" for tweets that are not about any incident.

As we aim to identify traffic crashes, we only extract "crash" tweets as positive instances and "NO" tweets as negative instances. In total, we obtain 129 tweets labelled as "crash", and 1,269 tweets annotated as "NO". Finally, we balance the distribution of positive and negative instances by randomly reducing the number

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

of negative instances to 129 tweets. In order to evaluate our incident-related tweet classifier, we randomly divided our ground truth dataset into training and test sets, containing 80% and 20% of the data respectively. Table 6.3 presents the total number of positive ("Crash") and negative instances ("NO") in the training and test sets.

Table 6.3: Number of positive instances (*Crash*) and negative instances (*NO*) in the training and testing datasets for our tweet incident classifier.

	Crash	**NO**
Training	104	107
Testing	25	22
Total	129	129

6.4.2 Evaluation

Next, we experiment with three different algorithms for classification: Multinomial Naive Bayes (Zhang, 2004), Random Forest (Breiman, 2001) and Decision Trees (Breiman, 1984; Friedman et al., 2001) as implemented in the Sckit-Learn python package[1]. As a baseline, we use a Random classifier that generates predictions uniformly at random. We run a McNemar's test (Demšar, 2006) to assess statistic significance between our classifiers and the baseline

We preprocess each tweet following the procedure described before in Section 3.3.2: remove punctuations, hyperlinks, stopwords, tokenise (1-gram) and apply Porter Stemmer. As features, we use a TF-IDF representation of the words in the document. Lastly, we train the models in the training set and evaluate their performance in the test set. For measuring the performance of classification, we report precision, recall, F1-score and accuracy formalised as follows:

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{6.1}$$

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{6.2}$$

[1] http://scikit-learn.org/

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{6.3}$$

$$F1 - Score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{6.4}$$

where TP = True Positives, TN = True Negatives, FP = False Positives, and FN = False Negatives.

Table 6.4 present results for the incident-related tweet classification task. The best performing result for each metric is highlighted in bold. We observe that the Random Forest classifier outperforms the rest of the models in terms of precision, recall and F1-score. The Random Forest achieves a precision of 0.85, recall of 0.82, an F1-Score of 0.82 and an accuracy of 0.83. This performance is consistent with previous work on the classification of traffic tweets (D'Andrea et al., 2015; Gu et al., 2016; Schulz et al., 2013b).

Table 6.4: Results for the traffic incident-related tweet classification task. We report Precision (*Prec.*), Recall (*Rec.*), F1-Score (*F1*) and Accuracy (*Acc.*) for each of the classifiers evaluated. We run a McNemar's test to assess statistical significance with respect to the baseline (*Random*).

Model	**Prec.**	**Rec.**	**F1**	**Acc.**
Ramdom	0.51	0.49	0.49	0.49
MultinomialNB	0.73**	0.73**	0.73**	0.72**
Decision Trees	0.80**	0.80**	0.80**	0.80**
Random Forest	**0.85****	**0.82****	**0.82****	**0.83****

As a result of this experiment, we select the trained Random Forest classifier to identify incident-related tweets in the traffic incident detection pipeline described next in Section 6.6.

6.5 Fine-Grained Geolocalisation

In this section, we build our fine-grained geolocalisation models following the approaches developed in Chapters 3, 4 and 5 of this thesis. We incorporate these models into the traffic incident detection pipeline described next in Section 6.6.

According to previous research, the majority of traffic incidents occur in road intersections (Hakkert and Mahalel, 1978; Thomas, 1996). Additionally, traffic congestions and secondary incidents are caused by traffic incidents and have an effect on up to 1-2 miles (1.6-3.2 km) from the incident location (Khattak et al., 2009). Moreover, the data about the transportation network in Chicago[1] shows that the majority of the majority of the roads segments have a length of 0.5-0.6 miles (approximately 0.8-1 km). Therefore, we configure our fine-grained geolocalisation models to use pre-defined squared areas of size length 1 km.

For evaluating the models, we use the Chicago Twitter dataset described in Section 6.3.1 and select, for each of them, two configurations that will be applied later in Section 6.7. In this section, we only report the performance of the selected configurations, however, for completeness, we present more detailed result in Appendix A. Now, we describe the evaluation of the models for each of the chapters:

Chapter 3 models. For models based on Chapter 3 approach, we use the *Aggregated* and *Individual* approaches introduced in Section 3.2. Then, we evaluate the models following the experimental setting described in Section 3.3. Table A.3 presents detailed results of the experiments. Finally, we use the following models for the experiments in this chapter:

- **C3-Agg:** In this model we use the best performing configuration for the *Aggregated* approach, which uses *BM25* as retrieval model.
- **C3-Indv:** In this model we use the best performing configuration fo the *Individual* approach, which uses *TF-IDF* as retrieval model.

Chapter 4 models. For models based on Chapter 4 approach, we follow the experimental settings described in Section 4.4. We evaluated the models that uses the weighted majority voting approach introduced in Section 4.3. We present complete evaluation results in Table A.2. Finally, we select the following models for rest of the experiments in this chapter:

[1] The data is available in https://support.office.com/en-us/article/create-a-histogram-in-excel-85680173-064b-4024-b39d-80f17ff2f4e8

- **C4-HA:** In this model we use the *WMV@Top-39* with $\alpha = 0.0$ configuration, which provides the highest accuracy but also low coverage.

- **C4-HC:** In this model we use the *WMV@Top-1* with $\alpha = 1.0$ configuration, which provides the highest coverage but also low accuracy.

Chapter 5 models. For building the models based on Chapter 5 approach, we select the learning to rank approach proposed in Chapter 5 that uses LambdaMART as the ranking function, trained to optimise NDCG@10. Moreover, we compute the set of *Query_ Common* features described in 5.2.1, as they showed to be the best performing one for geolocalisation according to experiments in Chapter 5. We present the complete set of results in Table A.3. Finally, we select the following models for the rest of the experiments in this chapter:

- **C5-HA:** In this model we use the *L2Geo+MV@Top-17* configuration, which provides the highest accuracy but also low coverage.

- **C5-HC:** In this model we use the *L2Geo* configuration, which provides the highest coverage but also low accuracy.

Lastly, we present the performance of each of the models selected above on fine-grained geolocalisation. Table 6.5 presents the metrics described in Section 3.3.6 for each of the models, namely average error distance (*AED*), median error distance (*MED*), accuracy at 1 km (*Acc@1km*) and coverage.

6.6 Traffic Incident Detection

In this section, we introduce the traffic incident detection task, which aims to identify traffic incident-related tweets from the Twitter stream. In order to evaluate the traffic incident detection task, we link the incident-related tweets to a ground truth of traffic crashes events reported by the Chicago Police Department (See Section 6.3.2). Previous work (D'Andrea et al., 2015; Gu et al., 2016; Schulz et al., 2013b) have used the already available set of geotagged tweets to perform the task. However, in this chapter, we aim to expand the sample of fine-grained geotagged incident-related tweets by applying our fine-grained geolocalisation approaches, presented in Section 6.5, for inferring the geolocation of new

Table 6.5: Results on fine-grained geolocalisation of the models selected for the experiments in this chapter, which follow the approaches introduced in Chapters 3, 4 and 5. We report the average error distance (*AED*), median error distance (*MED*), accuracy at 1 km (*Acc@1km*) and coverage.

	Chicago (25th July to 1st August)			
Model	**AED(km)**	**MED(km)**	**Acc@1km**	**Coverage**
C3-Agg	4.496	1.074	48.96%	99.96%
C3-Indv	1.717	**0.000**	82.07%	**100.00%**
C4-HA	1.404	0.471	87.86%	35.54%
C4-HC	3.993	0.563	58.75%	85.50%
C5-HA	**1.108**	0.435	**90.55%**	49.00%
C5-HC	1.578	0.435	86.03%	**100.00%**

non-geotagged incident-related tweets. To this end, we integrate a fine-grained geolocalisation process within the traffic incident detection pipeline, illustrated in Figure 6.2.

The remainder of the section is as follows. First, we present the output of the traffic incident classification process of the pipeline. Second, we discuss the performance of the fine-grained geolocalisation process using the geolocalisation models evaluated in Section 6.5. Finally, we link the resulting geolocalised and geotagged traffic incident-related tweets to our ground truth of traffic crashes events, and assess the performance of the traffic incident detection pipeline.

6.6.1 Identifying Traffic Incident-Related Tweets

The first process in the traffic detection pipeline, illustrated in Figure 6.2, aims to identify tweets whose content is related to a traffic crash incident. To this end, we integrate the incident-related tweet classifier that we built previously in Section 6.4. Our classifier processes the set of geotagged tweets as well as the set of geobounded tweets described in Section 6.3.1. The classifier processes each tweet and filters out those messages predicted as “NO”. Then, messages predicted as “Crash” are retained as candidate traffic incident-related tweets for traffic incident detection.

Table 6.6 presents the final number of tweets that our classifier predicts as incident-related for each of the sets considered: geotagged and geobounded. In total, we obtain 705 traffic incident-related geotagged tweets, and 6,524 traffic

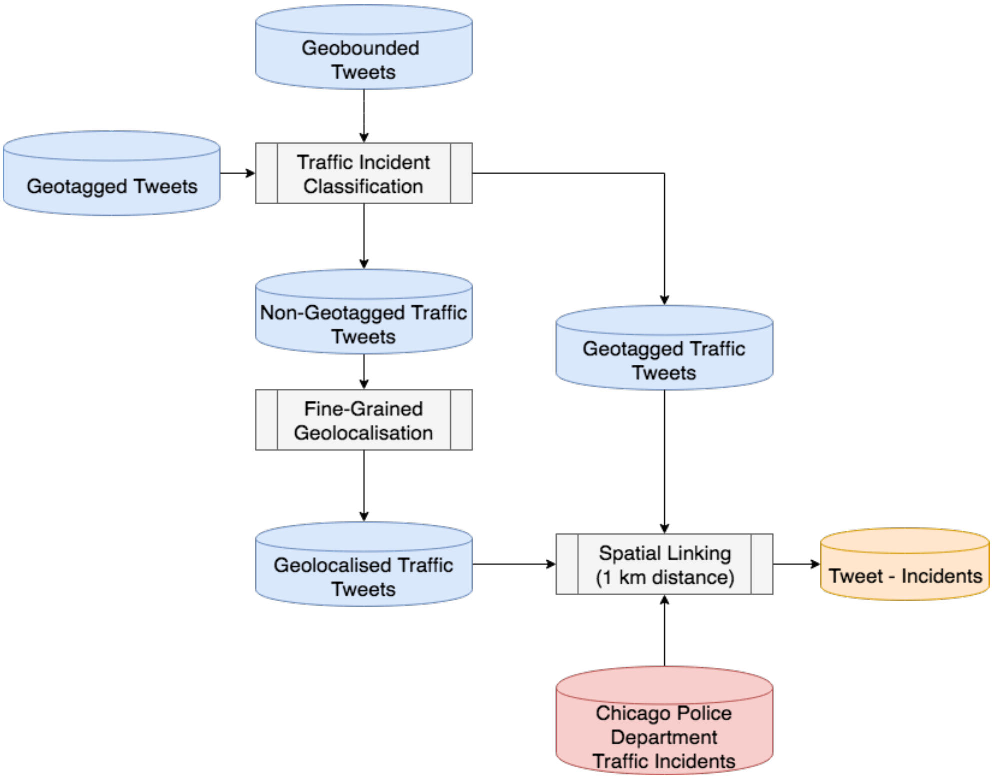
Geobounded Tweets
Geotagged Tweets
Traffic Incident Classification
Non-Geotagged Traffic Tweets
Geotagged Traffic Tweets
Fine-Grained Geolocalisation
Geolocalised Traffic Tweets
Spatial Linking (1 km distance)
Tweet - Incidents
Chicago Police Department Traffic Incidents

Figure 6.2: Traffic Incident Detection Pipeline. We integrate our fine-grained geolocalisation apporach to infer the geolocation of non-geotagged traffic tweets.

Table 6.6: Number of traffic incident-related tweet identified out of the sets of geotagged tweets and the set of geobounded tweets. We also report the total number of traffic incident-related tweets available.

	Total Tweets	**Incident-Related**
Geotagged Tweets	19,962	705
Geobounded Tweets	459,233	6,524
Total Tweets	479,196	7,229

incident-related geobounded tweets. In the next step, we feed the geobounded (non-geotagged) incident-related tweets into our geolocalisation models to predict fine-grained geolocation for them.

6.6.2 Applying Fine-Grained Geolocalisation

We integrate the fine-grained geolocalisation process into the traffic incident detection task, as illustrated in Figure 6.2, in order to increase the sample of finely-grained geolocated traffic incident-related tweets. As geolocalisation models, we use the fine-grained geolocalisation models that we built before in Section 6.5. We apply our geolocalisation models to infer a fine-grained geolocation for the 6,671 traffic incident-related tweets identified by our classifier in Section 6.6.1.

Table 6.7 present the final number of fine-grained geolocalised traffic incident-related tweets obtained by geolocalisation models.

Table 6.7: Number of tweets geolocalised by our geolocalisation models out of the total geobounded traffic incident-related (**I-R**) tweets (N=6,671).

	I-R Tweets
C3-Agg	6,497
C3-Indv	6,494
C4-HA	407
C4-HC	4,804
C5-HA	465
C5-HC	6,494

As a result of the geolocalisation process, we observe that models that provide higher coverage (i.e., *C3-Agg, C3-Indv, C4-HC and C5-HC*) are capable of finding a geolocation for a higher number of incident-related tweets. On the other hand, models that provide lower coverage but higher accuracy (i.e., *C4-HA and C5-HA*) are capable of geolocalise a smaller number of incident-related tweets. These results are consistent with the behaviour observed in our experiments in Section 6.5.

Additionally, we show in Table 6.8 the number of incident-related tweets that are already available in the set of geotagged tweets (*Geotagged*), as well as the final number of incident-related tweets we obtain as a result of expanding the sample of geotagged incident-related tweets. For instance, when adding the 465 incident-related tweets geolocalised using the *C5-HA* to the initial 705 geotagged incident-related tweets, we obtain a final set of 1,170 incident-related finely-grained georeferenced tweets.

Table 6.8: Number of incident-related geotagged tweets (*Geotagged*), and final number of georeferenced tweets, after adding new incident-related (**I-R**) tweets geolocalised using the models described in Section 6.5.

	I-R Tweets
Geotagged	705
Geotagged + C3-Agg	7,202
Geotagged + C3-Indv	7,199
Geotagged + C4-HA	1,112
Geotagged + C4-HC	5,509
Geotagged + C5-HA	1,170
Geotagged + C5-HC	7,199

After the fine-grained geolocalisation process, we next link in Section 6.6.3 the resulting geolocalised traffic incident-related tweets to the incidents from the Chicago Police Department traffic crashes dataset (See 6.3.2). Finally, we obtained the traffic incident-related tweets that are located at 1 km distance or less.

6.6.3 Spatial Linking to Traffic Incidents

To evaluate the effectiveness of the traffic incident detection task, we perform a linking process that associates traffic incident-related tweet with traffic crashes reported by the Chicago Police Department in the same period of time (testing period), described in Section 6.3.2. Our linkage strategy is based on spatial matching criteria between tweets and incidents and returns pairs of tweet-incidents that are placed between each other at 1 km distance or less.

6.6.4 Evaluation Metrics

After the linking process, we compute the following metrics to evaluate the performance of the traffic incident detection pipeline.

- **Accuracy:** We define accuracy as the percentage of traffic incident-related tweets that are linked to an incident. In this chapter, we consider that a tweet is linked to an incident is it is placed within a 1 km distance of it. Higher values represent better performance.

- **Detection Rate:** We define detection rate as the percentage of incidents that are covered by the traffic incident-related tweets. An incident is covered if it contains at least one traffic-related tweet within 1 km distance. Higher values represent better performance.

6.7 Experimental Results

In this section, we present and discuss our experimental results on the traffic incident detection task. First, in Section 6.7.1 we evaluate the traffic incident detection pipeline, described in Section 6.6, using the new incident-related tweets that are geolocalised using our fine-grained geolocalisation models, described in Section 6.5. This experiment aims to address the research question **RQ-6.1**, that investigates the effectiveness of our geolocalisation approaches on the traffic incident detection task.

Additionally, in Section 6.7.2 we evaluate how the detection of traffic incidents is improved when using an expanded sample of georeferenced traffic incident-related tweets. This georeferenced set consists of the geotagged incident-related tweets expanded with the new geolocalised incident-related tweets obtaining during the fine-grained geolocalisation process of the pipeline, as described in Section 6.7. This experiment aims to assess whether our fine-grained geolocalisation approaches can effectively expand the sample of finely-grained geolocated incident-related data and, therefore, benefits the overall performance of the traffic incident detection.

For each of the experiment mentioned above, we report the metrics described in Section 6.6.4, which aim to measure the number real incident we can cover with traffic incident-related tweets (detection rate), and what percentage of these traffic incident-related tweets are located at 1 km or less to the real locations of the incidents (accuracy). We compute these metrics when considering tweets generated at 1 to 30 minutes after the incident.

We present results in Figures 6.3 and 6.4 for experiments in Section 6.7.1, and Figures 6.5 and 6.6 for experiments in Section 6.7.2. For completeness, we report detailed experimental results in Tables B.1 and B.2 in Appendix B.

The remainder of this section is as follows. In Section 6.7.1 we address **RQ-6.1** and evaluate the effectiveness of the geolocalised traffic incident-related tweets for traffic incident detection. Next, in Section 6.7.2 we address **RQ-6.2** and evaluate whether the traffic incident detection is improved by enhancing the set of traffic geotagged tweets with geolocalised traffic tweets. Finally, we provide concluding remarks in Section 6.9.

6.7.1 Effectiveness of Geolocalised Tweets

We first evaluate the performance of traffic incident detection when considering only the traffic incident-related tweets geolocalised by our fine-grained geolocalisation approaches. We compute accuracy and detection rate (See Section 6.6.4) for the geolocalisation models described in Section 6.7.

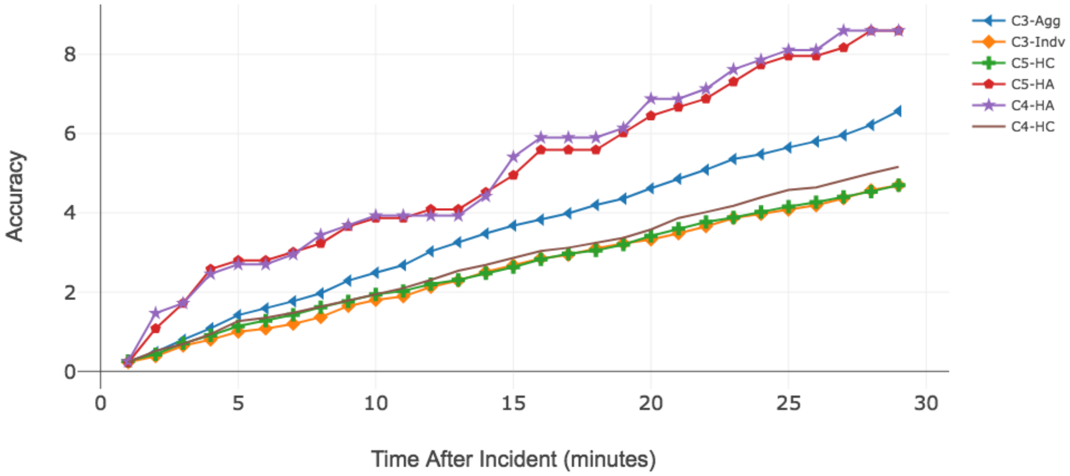

Figure 6.3: Accuracy (y-axis) for 1 minute to 30 minutes after the incident (x-axis) for the traffic incident-related geolocalised tweets using our fine-grained geolocalisation approaches.

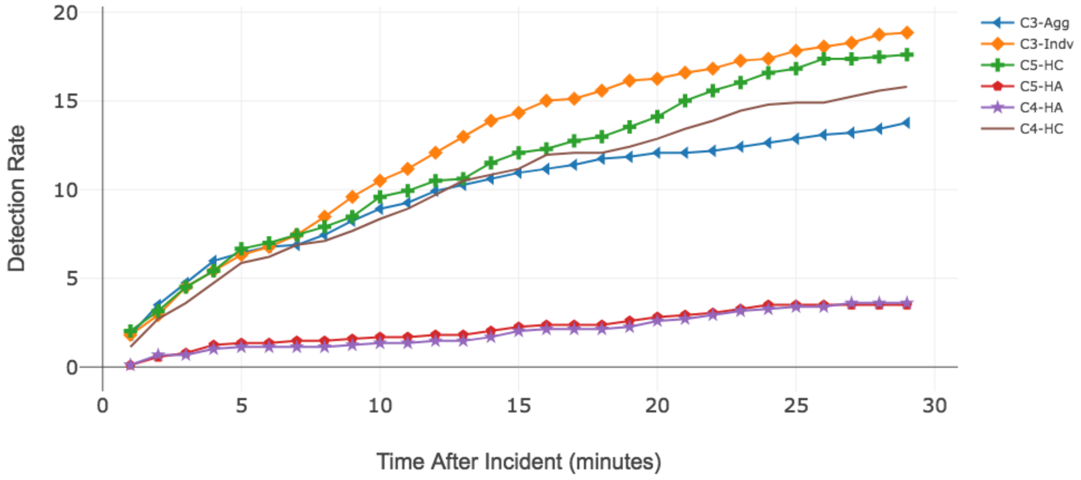

Figure 6.4: Incident Detection Rate (y-axis) for 1 minute to 30 minutes after the incident (x-axis) for the traffic incident-related geolocalised tweets using our fine-grained geolocalisation approaches.

We observe in Table B.1 that models that provide high accuracy of geolocalisation (i.e., *C4-HA and C5-HA*) also achieve higher accuracy of detection over the geolocalisation models that provide high coverage (i.e., *C3-Agg, C3-Indv, C4-HC and C5-HC*). On the other hand, we observe in Figure 6.4 that models that provide high coverage models achieve a higher detection rate compared to models that provide high accuracy. This means that high accurate geolocalisation models cover a lower percentage of the incidents, but they are capable of accurately detecting their geographical location (1 km distance). In contrast, models that provide high coverage can identify a larger number of incidents, but the geographical location of them are not accurately predicted (1 km distance).

This is the expected behaviour considering the geolocalisation performance the models, observed in Section 6.5, and this behaviour is consistent with the behaviour observed thought Chapters 3, 4 and 5 in this thesis. These results address the research question **RQ-6.1** and demonstrate the effectiveness of our fine-grained geolocalisation approach on the traffic incident detection task.

Besides, the observed behaviour of our fine-grained geolocalisation approaches evaluated in a new dataset (Chicago July 2016), presented in Section 6.5, and the consistency shows the generalisation of out fine-grained geolocalisation approaches.

6.7.2 Expanding The Sample of Geotagged Tweets

In this section, we aim to address research question **RQ.6.2**, which aims to assess whether the traffic incident detection is improved by enhancing the sample of already available traffic incident-related geotagged tweets with a new set of geolocalised traffic incident-related tweets. To this end, we compare the performance of the detection when considering the geotagged tweets alone (*Geotagged*), and when considering the set the geotagged tweets expanded with the incident-related tweets geolocalised using our fine-grained geolocalisation models described in Section 6.5.

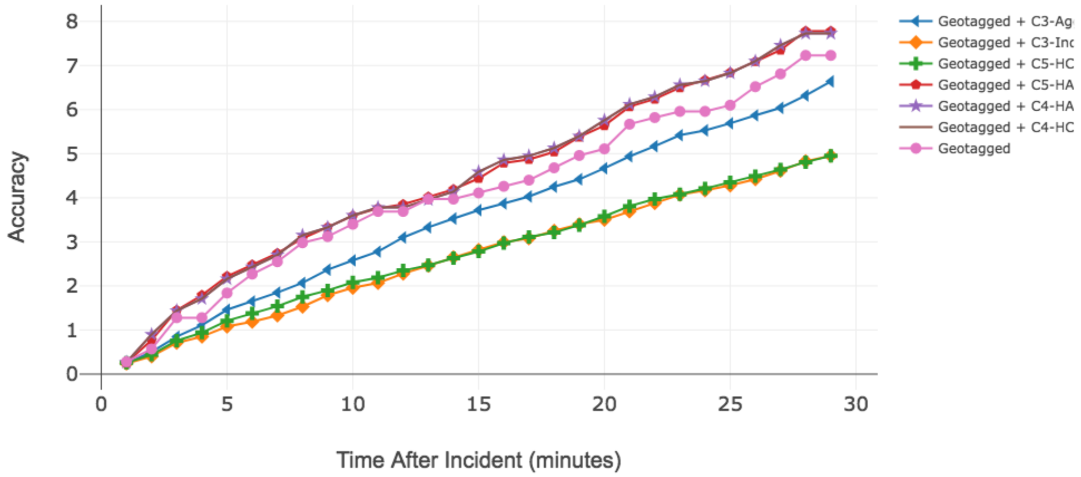

Figure 6.5: Accuracy (y-axis) for 1 minute to 30 minutes after the incident (x-axis) for the traffic incident-related geotagged tweets (*Geotagged*), and the expanded samples using tweets geolocalised using our fine-grained geolocalisation approaches.

We observe in Figure 6.5 that the overall performance of the traffic incident detection task is improved when the sample of geotagged traffic tweets is expanded over using geotagged incident-related tweets alone (*Geotagged*). In particular,

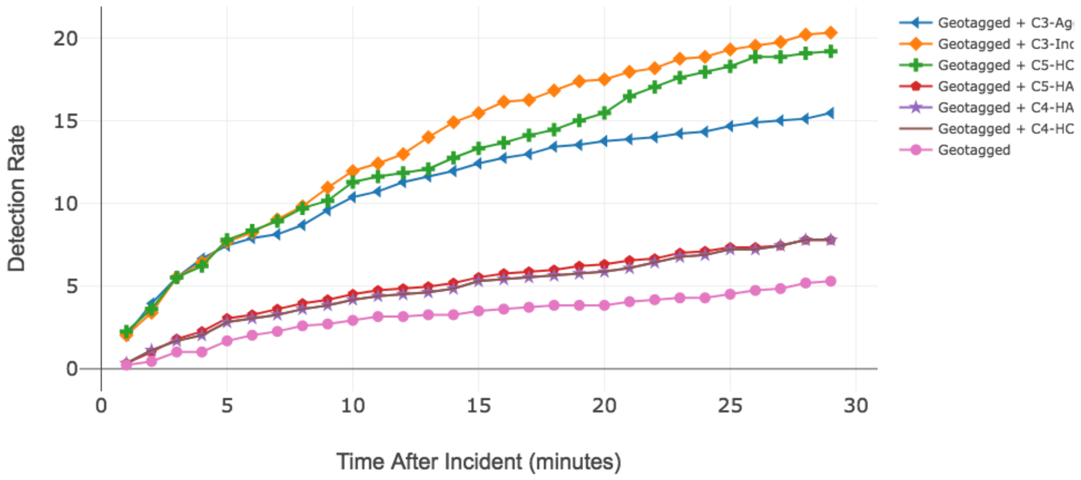

Figure 6.6: Incident Detection Rate (y-axis) for 1 minute to 30 minutes after the incident (x-axis) for the traffic incident-related geotagged tweets (*Geotagged*), and the expanded samples using tweets geolocalised using our fine-grained geolocalisation approaches.

accuracy is improved when adding the tweets geolocalised using the models that provides high accuracy of geolocalisation (i.e., *Geotagged + C4-HA and Geotagged + C5-HA*), whereas detection rate is improved when expanding using the tweets geolocalised using the models that provide high coverage (i.e., *Geotagged + C3-Agg, Geotagged + C3-Indv, Geotagged + C4-HC and Geotagged + C5-HC*).

These results address the research question **RQ-6.2** and show that the traffic incident detection is improved when considering new incident-related tweets geolocalised using our fine-grained geolocalisation approaches. This supports the hypothesis, introduced in Section 6.1, that states that by enhancing the set of already available geotagged tweets with geolocalised tweets we can improve the performance of the traffic incident detection task.

6.8 Recommendations to Transportation Managers

Previous research defines incident duration as the period between an incident occurs, and the time the incident is cleared. This period is divided into three main phases, illustrated in Figure 6.7: detection/reporting time, response time and clearance time (Nam and Mannering, 2000). According to Nam and Mannering (2000), the average times of each of the phases are 12.2 minutes for the detection time phase, 26.2 for the response time phase and 136.8 minutes for the clearance time phase.

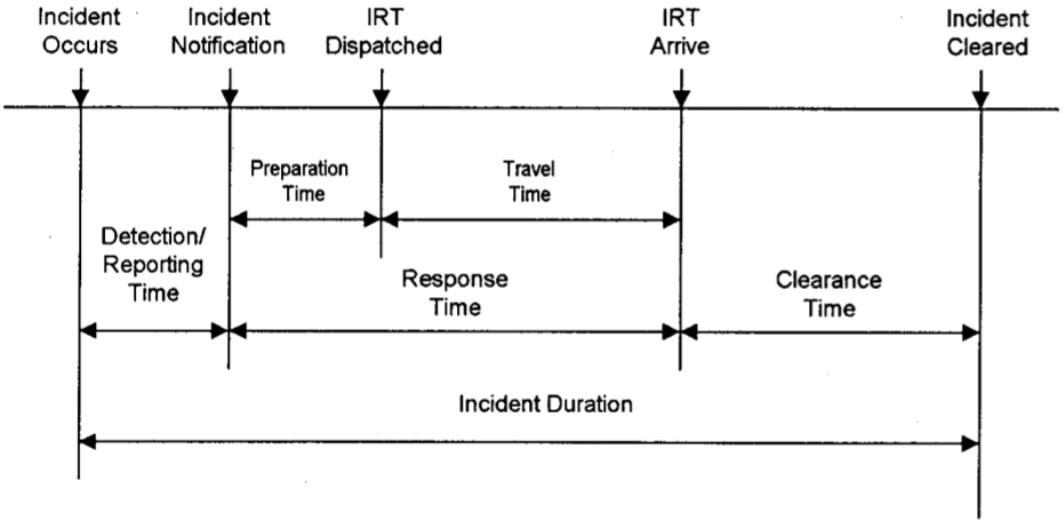

Figure 6.7: The phases of incident duration (Nam and Mannering, 2000).

On the other hand, information about the context of the incidents (i.e., number of vehicles involved, type of vehicles, injuries or fire) showed to be helpful for predicting the duration of an incident (Pereira et al., 2013). For this reason, the content of the tweets associated with the occurrences can provide crucial information to transportation managers.

The traffic incident information from the tweets becomes more decisive as it is extracted during the detection/reporting time, and closer to the time the incident occurs. Therefore, managers should aim for more accurate models as the location of the event is still unknown, and the emergency services have not verified the occurrence yet. In contrast, after detection time, the location of the incident, and emergency services are in place. Thus, accuracy is not crucial for this phase.

Besides, during the response and clearance time phases, information reported by users in real-time about the evolution of the incident, or other events occurring in its surrounding (i.e., congestions, subsequent events), can be helpful to manage the incident until the road is completely cleared efficiently. Therefore, managers should consider a model that provides high coverage in these phases, and thus maximise the amount of information extracted from the Twitter messages.

Finally, according to our results in Figure 6.5 and Table B.2 in Appendix B, and considering the average time of the detection phase reported in the literature (12.2 minutes), we suggest the following models, described in Section 6.5, for the different phases of the incident: *C5-HA* or *C4-HA* for detection/reporting time, and *C3-Indv* for response and clearance time.

6.9 Conclusions

The use of Twitter data as a complementary data source for the detection of traffic incidents have attracted the attention of transportation researchers. The reason is that Twitter data is ubiquitous and provides first-hand reports of the incidents. To perform the task, researchers have developed several machine learning approaches that aim to identify content related to traffic incidents in the tweets. However, once the content is identified, it is crucial for predicting the location of the incident from the evidence in the tweets. So far, researchers have used the set of already available geotagged tweets, which represents only 1% of the Twitter stream. This means that transportation managers rely on a very small sample of geotagged tweets to do their analysis. Thus, in this chapter, we explored whether by expanding the sample of geotagged tweets by inferring the geolocation of new non-geotagged tweets we can improve the performance of the traffic incident detection task. We used our fine-grained geolocalisation approaches developed previously in Chapters 3, 4 and 5 in this thesis, and explored the usefulness of integrating the fine-grained geolocalisation process into the traffic incident detection task.

We collected Twitter data from the city of Chicago during July 2016 to build our fine-grained geolocalisation approaches (tweets from 1st July to 25th July) and determined a period of study of a week (25th July to 1st August) as a testing period. Also, to evaluate the detection of traffic incidents we collected a dataset of traffic crashes reported by the Chicago Police Department occurred during the same period of study.

We present the traffic incident detection pipeline, illustrated in Figure 6.2, and consists of the following stages. First, we used a text classifier to identify whether the content of a tweet is about a traffic crash or not. To build our classifier, we use a gold standard dataset generated by Schulz et al. (2013b, 2017) (See Section 6.4). Then, we used the built classifier to obtain traffic incident-related tweets from the set of geotagged tweets as well as from the set of geobounded tweets (not fine-grained geotagged). Second, we passed the non-geotagged traffic incident-related tweet through the fine-grained geolocalisation process, which incorporates our fine-grained geolocalisation approaches developed in Chapters 3, 4 and 5 (See Section 6.5). Finally, we compared the obtained fine-grained geolocalised incident-related tweets to the real traffic crashes reports from the Chicago Police Department. As metrics, we reported accuracy and detection rate, described in Section 6.6.4.

Our experimental results in Table B.1 showed that the incident-related tweets geolocalised using geolocalisation models that provide high coverage (i.e., *C3-Agg, C3-Indv, C4-HC and C5-HC*) detected a large number of incidents (higher detection rate), but in contrast they were not capable of accurately predicts their geolocation (accuracy). In comparison, incident-related geolocalised tweets using geolocalisation models that provide high accuracy (i.e., *C4-HA and C5-HA*) detected a lower number of incidents, but their geolocations are predicted accurately at a distance of 1 km or less (accuracy). These results address the research question **RQ-6.1** which aims to assess the usefulness of geolocalised data on traffic incident detection. Besides, the consistency these results with the behaviour of geolocalisation observed thought this thesis shows the generalisation of out fine-grained geolocalisation approaches.

Finally, when expanding the sample of traffic incident-related geotagged tweets with the new fine-grained geolocalised traffic incident-related tweets the overall performance of the traffic incident detection is improved (regarding accuracy and detection rate). These results support the central hypothesis of this chapter (see **Hypothesis 4** in Section 1.2) and address research question **RQ-6.2**, which states that by expanding the sample of geotagged tweets we can improve the performance of the traffic incident detection task.

In the next chapter, we provide concluding remarks of the work in this thesis as well as discussing new open research question and future research directions.

Part IV

Conclusions

Chapter 7

Conclusions and Future Work

This thesis investigated the problem of inferring the geolocation of individual tweets at a fine-grained level of granularity (i.e., 1 km error distance). We argued that, by exploiting the characteristics of individual finely-grained geotagged tweets that are already available in the Twitter stream, we could achieve the geolocalisation of non-geotagged tweets at a fine-grained level. We postulated a correlation between the content similarity and geographical distance between tweets that are posted within a fine-grained area. Therefore, if two tweets contain similar content, then it is very likely that they were generated in the same location.

Across all the chapters in this thesis, we have addressed the problem of whether geolocalisation can be achieved using the content of tweets, and proposed novel approaches that advance the existing literature further by providing highly accurate geolocalisation of Twitter posts. The experiments undertaken in this thesis showed the effectiveness of our fine-grained geolocalisation approaches, so it is possible to infer the geolocation of tweets at a fine-grained level. Additionally, we investigated the effectiveness of our proposed approach in a practical application by incorporating our fine-grained geolocalisation approach into the traffic incident detection pipeline. In this chapter, we summarise the main contributions of this thesis and discuss the findings and conclusions of our research.

The remainder of this chapter is as follows. We first summarise the main contributions of this thesis in Section 7.1. Next, we discuss the main conclusions and achievements of this thesis in Section 7.2. Finally, we discuss future research directions in Section 7.3.

7.1 Contributions

The main contributions of this thesis are as follows:

- In Chapter 3, we investigated the limitations of the state-of-the-art tweet geolocalisation approaches when they are adapted to work at a fine-grained level. We provided insights to understand the drawbacks of existing research and proposed a ranking approach that alleviates such limitations, thus enabling the geolocalisation of tweets at a fine-grained level.

- In Chapter 4, we discuss the predictability of the geolocation of tweets at a fine-grained level. We postulated a correlation between content similarity and geographical distance in fine-grained predictions. Based on this, we proposed a novel approach that incorporates a weighted majority voting algorithm, which exploits the geographical evidence encoded within the Top-N most similar geotagged tweets.

- In Chapter 5, we investigated whether improving the ranking of the geotagged tweets can lead to a better performance in fine-grained geolocalisation. We proposed a learning to rank-based approach that re-ranks geotagged tweets based on their geographical proximity to a given non-geotagged tweet. Additionally, we proposed a set of features for geolocalisation and investigated the best performing combination of them.

- In Chapter 6, we demonstrated the usefulness of our proposed fine-grained geolocalisation approach in the traffic incident detection task. We incorporated our geolocalisation method into the traffic incident detection pipeline to infer the geolocalisation of non-geotagged tweets, and expand the sample of finely-grained geolocated traffic related tweets. We then showed the improvements in traffic incident detection by evaluating the pipeline over a ground truth of official incidents reports.

7.2 Findings and Conclusions

The main findings and conclusions of this thesis are that *fine-grained geolocalisation of tweets can be achieved by exploiting the characteristics of already available geotagged tweets and, in doing so, it is important to consider the following:*

- (Chapter 3) When performing fine-grained geolocalisation using a ranking approach, representing an area as a document containing the text of an individual tweet performs significantly better than aggregating the texts of the geotagged tweets from a pre-defined area into a virtual document. We increased accuracy at 1 km (*Acc@1km*) from 50.67% to 55.20% in Chicago (Table 3.4) and from 45.40% to 48.46% in New York (Table 3.5).

- (Chapter 3) Document frequency information is lost when aggregating the tweets into a single document, and this evidence is transformed into term frequency information. Moreover, document frequency information has been shown to be the most discriminative feature for fine-grained geolocalisation, and it is effectively exploited when using individual tweets to represent locations (see Figure 3.2), increasing the performance over models using the aggregation of tweets.

- (Chapter 4) The predictability of tweets at a fine-grained level is derived by the correlation between their content similarity and the geographical distance to other geotagged tweets. By ranking geotagged tweets based on content similarity and exploiting the geographical evidence encoded within the Top-N tweets in the ranking, we can find reduce the average error distance of the predicted tweets from 4.694 km to1.602 km in Chicago and 4.972 km to 1.448 in New York (see Tables 4.1 and 4.2).

- (Chapter 5). The quality of the ranking of geotagged tweets is crucial for fine-grained geolocalisation. By improving the ranking using a tailored learning to rank approach, we can decrease the average error distance of our predictions from 1.602 km to 1.451 km in Chicago and from 1.448 km to 1.319 km in New York (see Tables 5.6 and 5.7). Moreover, we are capable of increase the number of tweets for which we can find a fine-grained geolocation.

Additionally, we demonstrated the applicability of the fine-grained geolocalisation approach developed in this thesis in a practical scenario. To this end, we incorporated our approach into the pipeline of the traffic incident detection task. Our findings are the following.

- (Chapter 6) We demonstrated the effectiveness of traffic incident-related tweets that are geolocalised using our fine-grained geolocalisation approach. The geolocation inferred for such tweets has been shown to be closer to the location of real incidents occurring in the same period.
- (Chapter 6) The consistency of the behaviour of our fine-grained geolocalisation approaches observed through the chapters of this thesis and their applicability on the detection of traffic incidents, supports the generalisation of our approaches.
- (Chapter 6) Expanding the already available geotagged tweets with new geolocalised tweets increases the overall performance of the traffic incident detection task.

In the rest of this section, we elaborate each of the findings in detail.

7.2.1 Limitations of State-of-The-Art Geolocalisation Approaches

First, in Chapter 3 we investigated the limitations of existing work when performing geolocalisation at a fine-grained level of granularity. To existing approaches divide the geographical area into areas of a pre-defined size. Then, each area is represented as a document that contains the aggregated texts of the geotagged tweets belonging to the area. However, when performing such an aggregation process, important information about discriminative words that are representative of fine-grained locations is lost. Therefore, we hypothesised that by considering geotagged tweets individually we could preserve the evidence loss when adapting previous approaches at a fine-grained level, and thus we can improve the performance of fine-grained geolocalisation (see **Hypothesis 1** in Section 1.2). To test our hypothesis, we answered the following research questions:

- **RQ-3.1:** Does consider geotagged tweets individually improve the performance of fine-grained geolocalisation?
- **RQ-3.2:** What is the effect of aggregating tweets within a predefined area on accuracy when geolocalising tweets at a fine-grained level?

In order to answer these research questions, we analysed the behaviour of the existing state-of-the-art approaches in the context of fine-grained geolocalisation and compared them with our proposed solution of considering geotagged tweets individually (See Section 3.3).

The first outcome of our experiments, reported in Section 3.4.1, addressed research question **RQ-3.1** and showed that our proposed solution of using individual geotagged tweets is capable of predicting the highest number of tweets at a fine-grained level (i.e., 1 km distance), and the average error distance is significantly (statistically) reduced (See Tables 3.4 and 3.5).

Second, we observed an interesting behaviour when using the *BM25* retrieval model in both approaches; aggregating the tweets and using individual tweets. We noted that there is not a high difference in performance. In Section 3.4.1.1, we concluded that due to the nature of tweets (short documents) and the inherent characteristics of the *BM25* model, formalised in Equation (3.3), information in terms of term frequency and document length is low. This suggested that document frequency provides the strongest evidence for fine-grained geolocalisation.

Following the previous finding, we then answered **RQ-3.2** and derived a theoretical explanation of the effects that aggregating tweets have on the evidence in the form of document frequency, which is affecting geolocalisation at a fine-grained level. In Section 3.4.2, we computed the distribution of error distance committed by both approaches. Results were presented in Tables 3.4 and 3.5 for our two datasets, Chicago and New York respectively.

We found that retrieval models that rely on document frequency (IDF and TF-IDF) performed significantly (statistically) the best when using individual tweets, but in contrast performed significantly (statistically) worst when aggregating the tweets. We concluded that document frequency is the less discriminative information when aggregating the tweets, but becomes the most important evidence when using individual tweets.

On the other hand, retrieval models that rely on term frequency and document length performed the best within all the models that use the aggregation of tweets. This finding suggests that the evidence encoded in the form of document frequency information is transformed into term frequency information when is aggregating the text of the tweets into a virtual document, and such models still capture it.

7.2.2 Predictability of the Geolocation of Tweets at a Fine-Grained Level

In Chapter 4, we explored the predictability of tweets at a fine-grained level. The ranking approach proposed previously in Chapter 3 achieved an average error distance of approximately 4.693 km, which represents a confident area of 69.19 km^2 (see Table 3.4). This is not sufficient for tasks that require data geolocated at a fine-grained level as defined in this thesis; 1 km error distance, which represents 3.14 km^2. In this chapter, we hypothesised that the predictability of the geolocation of tweets at a fine-grained level is given by the correlation between their content similarity and geographical distance to finely-grained geotagged tweets (see **Hypothesis 2** in Section 1.2). We postulate some cases the content similarity of the tweets does not always correlate with their geographical distance. These cases are being considered by the ranking approach proposed in Chapter 3 which leads to an increase in the average error distance of the predictions. By identifying such cases, we can increase the quality of the predictions at a fine-grained level. To this end, we proposed to exploit the geographical evidence encoded within the Top-N geotagged tweets in the ranking using a majority voting algorithm, described in Section 4.3.

We discussed the predictability of the geolocation of tweets and postulated a correlation between similarity and geographical distance, illustrated in Figure 4.1. In the context of such postulate, we validated our hypothesis by answering the following research questions.

- **RQ-4.1:** Can we obtain fine-grained predictions based on the geographical evidence between the Top-N most similar geotagged tweets?

- **RQ-4.2:** What is the percentage of tweets that we can predict at a fine-grained level?

In Section 4.5.1, we demonstrated that by exploiting the geographical evidence within the Top-N most similar geotagged tweets in the ranking, we were capable of identifying the cases of low correlation between similarity and geographical distance (see Tables 4.1 and 4.2). Therefore, we reduced the average error distance of the predictions, which answered the research question **RQ-4.1**. However, we observed a trade-off between error distance and coverage (number of tweets for which we can find a prediction). We found that as we considered higher values of N of the Top-N geotagged tweets, we achieved lower average error distance but also lower coverage, which answered **RQ-4.2**. However, we observed that this effect could be alleviated by weighting the votes in our majority voting algorithm (see Section 4.3.2).

7.2.3 Improving The Quality of The Ranking for Fine-Grained Geolocalisation

In Chapter 5, we explored whether the quality of the ranking of the Top-N geotagged tweets is affecting the performance of the geolocalisation at a fine-grained level. We hypothesised that by improving the ranking of geotagged tweets (denoted as **doc-tweets**) with respect to a given non-geotagged tweet (denoted as **query-tweet**), we can obtain more similar and geographically closer geotagged tweets, and thus we can obtain a higher number of other fine-grained predictions (see **Hypothesis 3** in Section 1.2). To improve the ranking, we proposed a learning to rank approach that re-ranks geotagged tweets based on their geographical proximity and introduced multiple features for the task.

To validate the hypothesis presented in this chapter, we addressed the following research questions.

- **RQ-5.1:** What is the best performing learning to rank algorithm to improve the ranking?

- **RQ-5.2:** Does improving the ranking of the geotagged tweets lead to better fine-grained geolocalisation?
- **RQ-5.3:** What set of features contributes the most to improve the accuracy of fine-grained geolocalisation?

Firstly, in Section 5.4.1 we addressed **RQ-5.1** and evaluated different learning to rank algorithms to determine the best suitable for the fine-grained geolocalisation. We compared six algorithms representing the three main categories in the literature: point-wise, pair-wise and list-wise algorithms. We observed in Tables 5.2 and 5.3 that LambdaMart (Wu et al., 2010) was the best performing algorithm compared to the others.

Next, we evaluated our learning to rank approach in fine-grained geolocalisation compared to our previous approaches developed in Chapter 3 and Chapter 4, which used the Vector Space Models using IDF weighting to perform the ranking. In Tables 5.4 and 5.5, we observed that our learning to rank approach outperformed previous models, answering **RQ-5.2** and supporting the central hypothesis that by improving the ranking we can also improve the performance of fine-grained geolocalisation. Moreover, we observed in Tables 5.6 and 5.7 that we can decrease the average error distance with a small decrease in coverage, compared to the approach in Chapter 4.

Finally, we observed that features extracted from the query-tweet, combined with the features that model the relation between the query-tweet and the doc-tweet (see Section 5.2.1), provided the best performance for fine-grained geolocalisation, which addressed research question **RQ-5.3**.

7.2.4 Effectiveness of The Fine-Grained Geolocalisation for Traffic Incident Detection

In Chapter 6, we investigated the applicability of the fine-grained geolocalisation approaches developed in this thesis in a practical scenario. We used the traffic incident detection task as a case study, which aims to use Twitter as a data source for detecting traffic incidents occurring in a city. We hypothesised that by geolocalising non-geotagged tweets we could obtain a more representative sample of geotagged data and, therefore, improve the effectiveness of the traffic incident detection task (see **Hypothesis 4** in Section 1.2). To this end, we integrated our fine-grained geolocalisation approach into the traffic incident detection pipeline, as illustrated in Figure 6.2. To validate our hypothesis, we answered the following research questions:

- **RQ-6.1:** What is the effectiveness of geolocalised traffic incident-related tweets on the traffic incident detection task?
- **RQ-6.2:** Does expand the sample of geotagged tweets with new geolocalised data improves the performance of the traffic incident detection task?

In order to evaluate the traffic incident detection task, we compare the location reported by the traffic incident-related tweets (identified using a state-of-the-art text classifier) to the real locations of official reports of incidents occurring in the same period of time. We geolocalised new traffic incident-related tweets using different configurations of the geolocalisation approaches developed in Chapter 2, Chapter 4 and Chapter 5. Based on previous evaluations of the geolocalisation approaches (see Section 6.5), we selected two configurations of each approach that provides different trade-offs between accuracy and coverage. Approaches that provided high error distance and high coverage (denoted as *HC*), and approaches that provided low error distance and low coverage (denoted as *HA*).

In Section 6.7.1, we evaluated the effectiveness in traffic incident detection (see Section 6.6.4) of new traffic incident-related tweets geolocalised using the above mentioned geolocalisation approaches. We observed that the new geolocalised tweets were geographically closer to the real incidents. Moreover, in line with the evaluation of the geolocalisation approaches, *HC* models were capable of detecting a higher number of incidents, but most of their locations were not accurately predicted (at 1 km distance or less). In contrast, most of the geolocalised tweets using *HA* models were capable of predicting the location of the incidents accurately, but in contrast, this model detected a lower number of incidents. These results addressed the research question **RQ-6.1** and demonstrated the effectiveness of our fine-grained geolocalisation approach. Additionally, such consistency in behaviour between the geolocalisation evaluation and the traffic incident detection of tweets supports the generalisation of our fine-grained geolocalisation approaches.

Finally, in order to address research question **RQ-6.2**, in Section 6.7.2 we enhanced the already available traffic incident geotagged tweets with the new tweets geolocalised using our approaches and evaluated the overall performance of the traffic incident detection task. We observed an increase in performance when using the enhanced sample against using the geotagged sample alone.

The previous results support the central hypothesis that by geolocalising non-geotagged tweets and expanding the sample of already available geotagged tweets, we can improve the performance of the incident detection task. Moreover, we demonstrated the effectiveness of our fine-grained geolocalisation approach in a practical application.

7.3 Future Research Directions

This thesis has opened several interesting research directions to be investigated in the future. The first research direction is to investigate the effect of the temporal aspect of tweets in our model. It is known that time is an important feature to take into account to improve geolocalisation (Dredze et al., 2016). Currently, our model does not take temporal characteristics into account. Also, in this thesis, we have evaluated our approaches using a period of a month (three weeks for training and one week for testing). It would be interesting to investigate how the stability of our model is affected by varying the size of the time windows for the training and testing periods.

The second research direction could investigate the drawbacks of using grids in our approach. The strategy of dividing the geographical space into fixed-size cells suffers from the data sparsity problem since some cells may not have sufficient data points, and thus might be under-represented. It could be interesting to test the performance of our geolocalisation approach when using different strategies of dividing the geographical space. There are several alternatives to discretise the geographical space that can tackle the data-sparsity problem. For example, an adaptive grid can be created by using a k-d tree data structure (Bentley, 1975), which provides high granularity (smaller areas) in dense regions and coarse granularity (bigger areas) elsewhere. Another option could be to use a density-based clustering algorithm, such as DBSCAN (Ester et al., 1996; Sander et al., 1998), to find dense, fine-grained regions and use them as candidate locations.

The third research direction could be 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 ambiguity issue can affect the accuracy of our model. Especially, the ambiguity problem can be a significant issue when dealing with non-geotagged tweets from the Twitter stream, which can originate anywhere in the world. So far, we have evaluated our approach in the context of a limited geographical area, which means that the fine-grained geolocalisation can be applied at the end of a pipeline that has previously used a coarse-grained geolocalisation method to infer the location at the city level. Therefore, it could be interesting to investigate the effect that the ambiguity issue has on the effectiveness of fine-grained geolocalisation, so we can incorporate the best techniques to alleviate this problem in our approaches. This can lead to the creation of a more generalised model that can be applied directly to the Twitter stream.

Finally, it could be interesting to evaluate the effectiveness of our fine-grained geolocalisation approach in other practical applications, and how it can improve their performance. Examples of alternative applications that require precise geolocated Twitter data are real-time event detection (Atefeh and Khreich, 2015; Crooks et al., 2013; Sakaki et al., 2010; Walther and Kaisser, 2013; Watanabe et al., 2011; Xia et al., 2014; Zhang et al., 2016a), sentiment analysis (Agarwal et al., 2011; Baucom et al., 2013; Kouloumpis et al., 2011; Pak and Paroubek, 2010), urban planning Frias-Martinez et al. (2012), topic detection (Hong et al., 2012b), and disaster and emergency analysis (Ao et al., 2014; Imran et al., 2015; McCreadie et al., 2016).

Appendix A

Fined-Grained Geolocalisation Models for Traffic Incident Detection

This appendix contains the evaluation results for the fine-grained geolocalisation approaches specified in Chapter 6, Section 6.5. Fine-grained geolocalisation approaches are evaluated over a dataset of geotagged tweet collected in Chicago during July 20126, as described in Section 6.3.1.

We report the geolocalisation evaluation metrics presented in Section 3.3.6, namely Average Error Distance in kilometres (AED), Median Error Distance in kilometres (MED), Accuracy at 1 kilometre (Acc@1km) and Coverage. The following tables are presented:

- Table A.1 presents resutls for the approaches discussed in Chapter 3.
- Table A.2 presents resutls for the approaches discussed in Chapter 4.
- Table A.3 presents resutls for the approaches discussed in Chapter 5.

Table A.1: Evaluation results for the Chapter 3 models. The table present the Average Error Distance in kilometres (AED), Median Error Distance in kilometres (MED), Accuracy at 1 kilometre (Acc@1km) and Coverage. Significant (statistically) differences with respect to the best Baseline (Aggregated using LMD) are denoted by $*$ (p$<$0.01).

Chicago (25th July to 1st August)					
Model	**Function**	**AED(km)↓**	**MED(km)↓**	**Acc@1km↑**	**Coverage↑**
Aggregated	*BM25*	4.496	1.074	48.96%	99.96%
Aggregated	*IDF*	14.044	14.201	10.40%	99.96%
Aggregated	*TF_IDF*	8.132	4.063	41.54%	99.96%
Aggregated	*DFR*	6.325	1.966	46.00%	99.96%
Aggregated	*LMD*	6.588	2.501	44.49%	99.96%
Individual	*BM25*	1.762	**0.000**	81.70%	**100.00%**
Individual	*IDF*	1.735	**0.000**	81.87%	**100.00%**
Individual	*TF_IDF*	**1.717**	**0.000**	**82.07%**	**100.00%**
Individual	*DFR*	1.767	**0.000**	81.64%	**100.00%**
Individual	*LMD*	1.765	**0.000**	81.77%	**100.00%**

Table A.2: Evaluation results for the Chapter 4 models. The table presents the Average Error Distance in kilometres (AED), Median of Error distance (MDE), Accuracy at Grid (A@Grid), Accuracy at 1 kilometre (A@1km) and Coverage for our proposed approach (WMV) using the Top-N (@TopN) elements in the rank and values of α.

Chicago (25th July to 1st August)					
Model	***Config***	***AED(km)*↓**	***MED(km)*↓**	***Acc@1km*↑**	***Coverage*↑**
WMV@Top-3	$alpha = 0.0$	3.379	0.474	67.01%	70.37%
WMV@Top-3	$alpha = 1.0$	3.993	0.563	58.75%	**85.50%**
WMV@Top-5	$alpha = 0.0$	2.240	0.471	78.34%	56.56%
WMV@Top-5	$alpha = 1.0$	3.681	0.520	62.00%	78.44%
WMV@Top-7	$alpha = 0.0$	1.850	0.471	82.49%	50.95%
WMV@Top-7	$alpha = 1.0$	3.110	0.471	67.92%	68.20%
WMV@Top-9	$alpha = 0.0$	1.719	0.471	84.29%	48.03%
WMV@Top-9	$alpha = 1.0$	2.651	0.471	72.49%	61.73%
WMV@Top-15	$alpha = 0.0$	1.566	**0.465**	86.60%	43.86%
WMV@Top-15	$alpha = 1.0$	2.216	0.471	77.47%	51.96%
WMV@Top-19	$alpha = 0.0$	1.533	0.470	86.45%	42.25%
WMV@Top-19	$alpha = 1.0$	2.080	0.471	78.86%	48.58%
WMV@Top-25	$alpha = 0.0$	1.505	0.471	87.03%	39.69%
WMV@Top-25	$alpha = 1.0$	2.031	0.471	79.68%	45.69%
WMV@Top-29	$alpha = 0.0$	1.444	0.471	87.42%	38.52%
WMV@Top-29	$alpha = 1.0$	2.010	0.471	79.58%	44.26%
WMV@Top-35	$alpha = 0.0$	1.424	0.471	87.75%	36.52%
WMV@Top-35	$alpha = 1.0$	2.024	0.471	79.28%	42.66%
WMV@Top-39	$alpha = 0.0$	**1.404**	0.471	87.86%	35.54%
WMV@Top-39	$alpha = 1.0$	2.007	0.471	79.40%	41.73%
WMV@Top-45	$alpha = 0.0$	1.408	0.471	**87.93%**	34.35%
WMV@Top-45	$alpha = 1.0$	1.970	0.471	79.29%	40.74%
WMV@Top-49	$alpha = 0.0$	1.423	0.471	87.74%	33.90%
WMV@Top-49	$alpha = 1.0$	1.940	0.471	79.47%	39.90%

Table A.3: Evaluation results for the Chapter 5 models. We present results for our learning to rank approaches (*L2Geo*) and (*L2Geo+MV*) considering the Top-3, to Top-49 most similar geotagged tweets. Table reports average error distance (*AED*), median error distance (*MED*), accuracy at 1 km (*Acc@1km*) and coverage (*Coverage*).

Chicago (25th July to 1st August)					
Model	***Features***	***AED(km)***↓	***MED(km)***↓	***Acc@1km***↑	***Coverage***↑
L2Geo	*Query_ Common*	1.578	0.435	86.03%	**100.00%**
L2Geo+MV@Top-3	*Query_ Common*	1.221	**0.434**	89.36%	73.47%
L2Geo+MV@Top-5	*Query_ Common*	1.127	0.435	90.34%	65.15%
L2Geo+MV@Top-7	*Query_ Common*	1.111	0.435	90.61%	60.29%
L2Geo+MV@Top-9	*Query_ Common*	1.123	0.435	**90.63%**	56.97%
L2Geo+MV@Top-11	*Query_ Common*	1.127	0.435	**90.63%**	54.26%
L2Geo+MV@Top-13	*Query_ Common*	1.122	0.435	90.61%	52.32%
L2Geo+MV@Top-15	*Query_ Common*	1.113	0.435	90.62%	50.67%
L2Geo+MV@Top-17	*Query_ Common*	**1.108**	0.435	90.55%	49.00%
L2Geo+MV@Top-19	*Query_ Common*	1.126	0.436	90.29%	47.69%
L2Geo+MV@Top-21	*Query_ Common*	1.152	0.439	90.12%	46.53%
L2Geo+MV@Top-23	*Query_ Common*	1.194	0.440	89.84%	45.47%
L2Geo+MV@Top-25	*Query_ Common*	1.191	0.442	89.74%	44.79%
L2Geo+MV@Top-27	*Query_ Common*	1.233	0.447	89.43%	43.99%
L2Geo+MV@Top-29	*Query_ Common*	1.262	0.449	89.08%	43.14%
L2Geo+MV@Top-31	*Query_ Common*	1.252	0.450	89.12%	42.37%
L2Geo+MV@Top-33	*Query_ Common*	1.244	0.451	89.14%	41.71%
L2Geo+MV@Top-35	*Query_ Common*	1.246	0.458	88.97%	41.33%
L2Geo+MV@Top-37	*Query_ Common*	1.251	0.465	88.94%	40.73%
L2Geo+MV@Top-39	*Query_ Common*	1.257	0.468	88.86%	40.26%
L2Geo+MV@Top-41	*Query_ Common*	1.255	0.468	88.87%	39.69%
L2Geo+MV@Top-43	*Query_ Common*	1.253	0.470	88.77%	39.06%
L2Geo+MV@Top-45	*Query_ Common*	1.266	0.471	88.55%	38.62%
L2Geo+MV@Top-47	*Query_ Common*	1.246	0.471	88.61%	38.14%
L2Geo+MV@Top-49	*Query_ Common*	1.253	0.471	88.46%	37.80%

Appendix B

Detailed Results for Traffic Incident Detection

B.1 Effectiveness of Geolocalised Tweets

Table B.1: Accuracy and Detection Rate at 5, 10, 15, 20, 25 and 30 minutes after the incident (**TAI**) of geolocalised traffic incident-related tweets. We present results for tweets geolocalised by the geolocalisation approaches described in Section 6.6.1.

	Accuracy↑					
TAI	*5 min*	*10 min*	*15 min*	*20 min*	*25 min*	*30 min*
C3-Agg	1.42%	2.49%	3.68%	4.62%	5.65%	6.71%
C3-Indv	1.00%	1.80%	2.68%	3.33%	4.08%	4.84%
C4-HA	2.70%	**3.93%**	**5.41%**	**6.88%**	**8.11%**	**8.85%**
C4-HC	1.27%	1.94%	2.87%	3.58%	4.58%	5.29%
C5-HA	**2.80%**	3.87%	4.95%	6.45%	7.96%	8.82%
C5-HC	1.14%	1.94%	2.63%	3.42%	4.16%	4.87%

	Detection Rate↑					
TAI	*5 min*	*10 min*	*15 min*	*20 min*	*25 min*	*30 min*
C3-Agg	6.43%	8.92%	10.95%	12.08%	12.87%	14.11%
C3-Indv	6.32%	**10.50%**	**14.33%**	**16.25%**	**17.83%**	**19.41%**
C4-HA	1.13%	1.35%	2.03%	2.60%	3.39%	3.61%
C4-HC	5.87%	8.35%	11.17%	12.87%	14.90%	16.03%
C5-HA	1.35%	1.69%	2.26%	2.82%	3.50%	3.50%
C5-HC	**6.66%**	9.59%	12.08%	14.11%	16.82%	17.83%

B.2 Expanding The Sample Of Geotagged Tweets

Table B.2: Accuracy and Detection Rate at 5, 10, 15, 20, 25 and 30 minutes after the incident (**TAI**) of traffic crash geotagged tweets expanded with traffic incident-related geolocalised tweets. We present results for the geotagged tweets alone (*Geotagged* compared to the expanded sample using crash-realated tweets geolocalised using the geolocalisation approaches described in Section 6.6.1.

	Accuracy↑					
TAI	*5 min*	*10 min*	*15 min*	*20 min*	*25 min*	*30 min*
Geotagged	1.84%	3.40%	4.11%	5.11%	6.10%	7.52%
Geotagged+C3-Agg	1.46%	2.58%	3.72%	4.67%	5.69%	6.79%
Geotagged+C3-Indv	1.08%	1.96%	2.82%	3.50%	4.28%	5.10%
Geotagged+C4-HA	2.16%	**3.60%**	**4.59%**	**5.76%**	6.83%	8.00%
Geotagged+C4-HC	1.34%	2.12%	3.03%	3.78%	4.77%	5.57%
Geotagged+C5-HA	**2.22%**	3.59%	4.44%	5.64%	**6.84%**	**8.03%**
Geotagged+C5-HC	1.21%	2.08%	2.78%	3.58%	4.35%	5.13%

	Detection Rate↑					
TAI	*5 min*	*10 min*	*15 min*	*20 min*	*25 min*	*30 min*
Geotagged	1.69%	2.93%	3.50%	3.84%	4.51%	5.53%
Geotagged+C3-Agg	7.45%	10.38%	12.42%	13.77%	14.67%	15.91%
Geotagged+C3-Indv	7.67%	**11.96%**	**15.46%**	**17.49%**	**19.30%**	**20.88%**
Geotagged+C4-HA	2.82%	4.18%	5.30%	5.87%	7.22%	8.01%
Geotagged+C4-HC	7.11%	9.93%	12.42%	13.77%	16.03%	17.04%
Geotagged+C5-HA	3.05%	4.51%	5.53%	6.32%	7.34%	8.01%
Geotagged+C5-HC	**7.79%**	11.29%	13.32%	15.46%	18.28%	19.41%

Bibliography

Mohammad-Ali Abbasi, Sun-Ki Chai, Huan Liu, and Kiran Sagoo. Real-world behavior analysis through a social media lens. In *Proceedings of the 5th International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction*, SBP'12, pages 18–26. Springer-Verlag, 2012. 2

Younos Aboulnaga, Charles L. A. Clarke, and David R. Cheriton. Frequent itemset mining for query expansion in microblog ad-hoc search. *TREC Microblog*, 2012. 27

Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. Sentiment analysis of twitter data. In *Proceedings of the workshop on languages in social media*, pages 30–38. Association for Computational Linguistics, 2011. 27, 142

Giambattista Amati. *Probability models for information retrieval based on divergence from randomness.* PhD thesis, University of Glasgow, 2003. 20, 21, 22, 23

Gianni Amati and Cornelis Joost Van Rijsbergen. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM TOIS, pages 357–389*, pages 357–389, 2002. 46

Ji Ao, Peng Zhang, and Yanan Cao. Estimating the locations of emergency events from twitter streams. *Procedia Computer Science*, 31:731–739, 2014. 2, 29, 142

Arvind Arasu, Junghoo Cho, Hector Garcia-Molina, Andreas Paepcke, and Sriram Raghavan. Searching the web. *ACM Transactions on Internet Technology (TOIT)*, 1(1):2–43, 2001. 25

Farzindar Atefeh and Wael Khreich. A survey of techniques for event detection in twitter. *Computational Intelligence*, 31(1):132–164, 2015. 2, 28, 142

Ricardo Baeza-Yates, Berthier Ribeiro-Neto, et al. *Modern information retrieval*, volume 463. ACM press New York, 1999. 11, 13

Jordan Bakerman, Karl Pazdernik, Alyson Wilson, Geoffrey Fairchild, and Rian Bahran. Twitter geolocation: A hybrid approach. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 12(3):34, 2018. 33, 34

Eric Baucom, Azade Sanjari, Xiaozhong Liu, and Miao Chen. Mirroring the real world in social media: twitter, geolocation, and sentiment analysis. In *Proceedings of the 2013 international workshop on Mining unstructured big data using natural language processing*, pages 61–68. ACM, 2013. 27, 142

Jon Louis Bentley. Multidimensional binary search trees used for associative searching. *Communications of the ACM*, 18(9):509–517, 1975. 141

Adam Berger and John Lafferty. Information retrieval as statistical translation. In *ACM SIGIR Forum*, volume 51, pages 219–226. ACM, 2017. 18, 19

Avrim Blum. On-line algorithms in machine learning. In *In Proceedings of the Workshop on On-Line Algorithms, Dagstuhl*, pages 306–325. Springer, 1996. 63

Abraham Bookstein and Don R Swanson. Probabilistic models for automatic indexing. *Journal of the Association for Information Science and Technology*, 25(5):312–316, 1974. 21

Robert S. Boyer and J. Strother Moore. *MJRTY - A Fast Majority Vote Algorithm*, pages 105–117. Springer Netherlands, 1991. 63

Leo Breiman. *Classification and regression trees*. Routledge, 1984. 114

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001. 25, 87, 103, 114

Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In *Proc. 22nd ACM ICML, pages 89–96*, 2005. 87, 103

Christopher J Burges, Robert Ragno, and Quoc V Le. Learning to rank with non-smooth cost functions. In *Advances in neural information processing systems*, pages 193–200, 2007. 25

Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. Learning to rank: from pairwise approach to listwise approach. In *Proceedings of the 24th international conference on Machine learning*, pages 129–136. ACM, 2007. 88, 103

Lewis Carroll. *Alice's adventures in wonderland*. Broadview Press, 2011. 12

Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. Information credibility on twitter. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pages 675–684. ACM, 2011. 62, 65

Hau-wen Chang, Dongwon Lee, Mohammed Eltaher, and Jeongkyu Lee. @phillies tweeting from philly? predicting twitter user locations with spatial word usage. In *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*, ASONAM '12, pages 111–118. IEEE Computer Society, 2012. 28

Fuxing Cheng, Xin Zhang, Ben He, Tiejian Luo, and Wenjie Wang. A survey of learning to rank for real-time twitter search. In *Joint International Conference ICPCA/SWS, pages 150–164*, 2012. 79, 81

Zhiyuan Cheng, James Caverlee, and Kyumin Lee. You are where you tweet: a content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 759–768. ACM, 2010. 28, 42

T-H Chiang, H-Y Lo, and S-D Lin. A ranking-based knn approach for multi-label classification. In *Asian Conference on Machine Learning*, pages 81–96, 2012. 62, 63

William S Cooper. A definition of relevance for information retrieval. *Information storage and retrieval*, 7(1):19–37, 1971. 17

W Bruce Croft, Donald Metzler, and Trevor Strohman. *Search engines: Information retrieval in practice*, volume 283. Addison-Wesley Reading, 2010. 25

Andrew Crooks, Arie Croitoru, Anthony Stefanidis, and Jacek Radzikowski. #earthquake: Twitter as a distributed sensor system. *Transactions in GIS*, 17(1):124–147, 2013. 28, 142

Jian Cui, Rui Fu, Chenghao Dong, and Zuo Zhang. Extraction of traffic information from social media interactions: Methods and experiments. In *Intelligent Transportation Systems (ITSC), 2014 IEEE 17th International Conference on*, pages 1549–1554. IEEE, 2014. 3, 107, 109

Firas Damak, Karen Pinel-Sauvagnat, Mohand Boughanem, and Guillaume Cabanac. Effectiveness of state-of-the-art features for microblog search. In *Proceedings of the 28th Annual ACM Symposium on Applied Computing*, SAC '13, pages 914–919. ACM, 2013. 81

Fred J Damerau. An experiment in automatic indexing. *Journal of the Association for Information Science and Technology*, 16(4):283–289, 1965. 21

Eleonora D'Andrea, Pietro Ducange, Beatrice Lazzerini, and Francesco Marcelloni. Real-time detection of traffic from twitter stream analysis. *Intelligent Transportation Systems, IEEE Transactions on*, 16(4):2269–2283, 2015. 3, 107, 109, 113, 115, 117

Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research*, 7(Jan):1–30, 2006. 114

Martin Dillon. Introduction to modern information retrieval: G. salton and m. mcgill. mcgraw-hill, new york (1983)., 1983. 16, 45

Mark Dredze, Miles Osborne, and Prabhanjan Kambadur. Geolocation for twitter: Timing matters. In *HLT-NAACL*, pages 1064–1069, 2016. 141

Yajuan Duan, Long Jiang, Tao Qin, Ming Zhou, and Heung-Yeung Shum. An empirical study on learning to rank of tweets. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 295–303. Association for Computational Linguistics, 2010. 81

Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 1277–1287. Association for Computational Linguistics, 2010a. 3, 33, 109

Jacob Eisenstein, Brendan O'Connor, Noah A Smith, and Eric P Xing. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1277–1287. Association for Computational Linguistics, 2010b. 28, 30, 34

Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231, 1996. 141

Paul Ferguson, Neil O'Hare, James Lanagan, Owen Phelan, and Kevin McCarthy. An investigation of term weighting approaches for microblog retrieval. In *European Conference on Information Retrieval*, pages 552–555. Springer, 2012. 27, 52, 54

David Flatow, Mor Naaman, Ke Eddie Xie, Yana Volkovich, and Yaron Kanza. On the accuracy of hyper-local geotagging of social media content. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pages 127–136. ACM, 2015. 30, 34, 47, 61

Vanessa Frias-Martinez, Victor Soto, Heath Hohwald, and Enrique Frias-Martinez. Characterizing urban landscapes using geolocated tweets. In *Proceedings of the 2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust*, SOCIALCOM-PASSAT '12, pages 239–248. IEEE Computer Society, 2012. 142

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, NY, USA:, 2001. 114

Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001. 25, 87, 103

Jinhua Gao, Guoxin Cui, Shenghua Liu, Yue Liu, and Xueqi Cheng. Ictnet at microblog track in trec 2013. *TREC Microblog*, 2013. 27

Jorge David Gonzalez Paule, Yashar Moshfeghi, Joemon M Jose, and Piyushimita Vonu Thakuriah. On fine-grained geolocalisation of tweets. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '17, pages 313–316. ACM, 2017. 9

Jorge David Gonzalez Paule, Yashar Moshfeghi, Craig Macdonald, and Iadh Ounis. Learning to geolocalise tweets at a fine-grained level. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 1675–1678. ACM, 2018a. 9

Jorge David Gonzalez Paule, Yeran Sun, and Yashar Moshfeghi. On fine-grained geolocalisation of tweets and real-time traffic incident detection. *Information Processing & Management*, 2018b. 9

Jorge David Gonzalez Paule, Yeran Sun, and Piyushimita Vonu Thakuriah. Beyond geotagged tweets: Exploring the geolocalisation of tweets for transportation applications. pages 1–21, 2019. 9

Irena Grabovitch-Zuyev, Yaron Kanza, Elad Kravi, and Barak Pat. On the correlation between textual content and geospatial locations in microblogs. In *Proceedings of Workshop on Managing and Mining Enriched Geo-Spatial Data*, GeoRich'14, pages 3:1–3:6. ACM, 2007. 66

Mark Graham, Scott A Hale, and Devin Gaffney. Where in the world are you? geolocation and language identification in twitter. *The Professional Geographer*, 66(4):568–578, 2014. 3, 28, 30, 108, 109, 111

Yiming Gu, Zhen Sean Qian, and Feng Chen. From twitter to detector: Real-time traffic incident detection using social media data. *Transportation research part C: emerging technologies*, 67:321–342, 2016. 3, 107, 109, 113, 115, 117

A Shalom Hakkert and David Mahalel. Estimating the number of accidents at intersections from a knowledge of the traffic flows on the approaches. *Accident Analysis & Prevention*, 10(1):69–79, 1978. 116

Bo Han and Paul Cook. A stacking-based approach to twitter user geolocation prediction. In *In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013): System Demonstrations*, pages 7–12, 2013. 3, 28, 109

Bo Han, Paul Cook, and Timothy Baldwin. Geolocation prediction in social media data by finding location indicative words. *Proceedings of COLING 2012*, pages 1045–1062, 2012a. 34

Bo Han, Paul Cook, and Timothy Baldwin. Text-based twitter user geolocation prediction. *Journal of Artificial Intelligence Research*, 49:451–500, 2014. 41

Bo Han, Afshin Rahimi, Leon Derczynski, and Timothy Baldwin. Twitter geolocation prediction shared task of the 2016 workshop on noisy user-generated text. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pages 213–217, 2016. 32, 34

Zhongyuan Han, Xuwei Li, Muyun Yang, Haoliang Qi, Sheng Li, and Tiejun Zhao. Hit at trec 2012 microblog track. In *TREC*, volume 12, page 19, 2012b. 81

Stephen P Harter. A probabilistic approach to automatic keyword indexing. *Journal of the American Society for Information Science (pre-1986)*, 26(5): 280, 1975a. 21

Stephen P Harter. A probabilistic approach to automatic keyword indexing. part ii. an algorithm for probabilistic indexing. *Journal of the Association for Information Science and Technology*, 26(5):280–289, 1975b. 21

Brent Hecht, Lichan Hong, Bongwon Suh, and Ed H Chi. Tweets from justin bieber's heart: The dynamics of the location field in user profiles. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 237–246. ACM, 2011. 30

Djoerd Hiemstra. A linguistically motivated probabilistic model of information retrieval. In *International Conference on Theory and Practice of Digital Libraries*, pages 569–584. Springer, 1998. 18, 19

Liangjie Hong, Amr Ahmed, Siva Gurumurthy, Alexander J. Smola, and Kostas Tsioutsiouliklis. Discovering geographical topics in the twitter stream. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, pages 769–778. ACM, 2012a. 66

Liangjie Hong, Amr Ahmed, Siva Gurumurthy, Alexander J Smola, and Kostas Tsioutsiouliklis. Discovering geographical topics in the twitter stream. In *Proceedings of the 21st international conference on World Wide Web*, pages 769–778. ACM, 2012b. 2, 142

Binxuan Huang and Kathleen M Carley. On predicting geolocation of tweets using convolutional neural networks. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*, pages 281–291. Springer, 2017. 31

Mans Hulden, Miikka Silfverberg, and Jerid Francom. Kernel density estimation for text-based geolocation. In *AAAI*, pages 145–150, 2015. 31, 34, 38, 40, 41, 44, 45

Muhammad Imran, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. Processing social media messages in mass emergency: A survey. *ACM Comput. Surv.*, 47 (4):67:1–67:38, 2015. 2, 29, 142

Lamjed Ben Jabeur, Firas Damak, Lynda Tamine, Guillaume Cabanac, Karen Pinel-Sauvagnat, and Mohand Boughanem. Irit at trec microblog track 2013. *TREC Microblog*, 2013. 27

Zongcheng Ji, Aixin Sun, Gao Cong, and Jialong Han. Joint recognition and linking of fine-grained locations from tweets. In *Proceedings of the 25th International Conference on World Wide Web*, pages 1271–1281. International World Wide Web Conferences Steering Committee, 2016. 30

Hideo Joho, Leif A Azzopardi, and Wim Vanderbauwhede. A survey of patent users: an analysis of tasks, behavior, search functionality and system requirements. In *Proceedings of the third symposium on Information interaction in context*, pages 13–24. ACM, 2010. 16

Asad Khattak, Xin Wang, and Hongbing Zhang. Are incident durations and secondary incidents interdependent? *Transportation Research Record: Journal of the Transportation Research Board*, 2099:39–49, 2009. 116

Sheila Kinsella, Vanessa Murdock, and Neil O'Hare. I'm eating a sandwich in glasgow: modeling locations with tweets. In *Proceedings of the 3rd international workshop on Search and mining user-generated contents*, pages 61–68. ACM, 2011. 3, 31, 32, 34, 38, 40, 41, 42, 44, 45, 47, 109

Raymondus Kosala, Erwin Adi, et al. Harvesting real time traffic information from twitter. *Procedia Engineering*, 50:1–11, 2012. 3, 107, 109

Efthymios Kouloumpis, Theresa Wilson, and Johanna D Moore. Twitter sentiment analysis: The good the bad and the omg! *Icwsm*, 11(538-541):164, 2011. 27, 142

Abhinav Kumar and Jyoti Prakash Singh. Location reference identification from tweets during emergencies: A deep learning approach. *International Journal of Disaster Risk Reduction*, 33:365–375, 2019. 32

Chenliang Li and Aixin Sun. Fine-grained location extraction from tweets with temporal awareness. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 43–52. ACM, 2014. 30

Chenliang Li and Aixin Sun. Extracting fine-grained location with temporal awareness in tweets: A two-stage approach. *Journal of the Association for Information Science and Technology*, 68(7):1652–1670, 2017. 30

Yaliang Li, Jing Gao, Chuishi Meng, Qi Li, Lu Su, Bo Zhao, Wei Fan, and Jiawei Han. A survey on truth discovery. *SIGKDD Explor. Newsl.*, 17(2):1–16, 2016. 62

Jimmy Lin and Miles Efron. Overview of the trec-2013 microblog track. Technical report, NIST, 2013. 20, 27

Jimmy Lin, Miles Efron, Yulu Wang, and Garrick Sherman. Overview of the trec-2014 microblog track. Technical report, NIST, 2014. 20, 27

Jimmy Lin, Miles Efron, Yulu Wang, Garrick Sherman, and Ellen Voorhees. Overview of the trec-2015 microblog track. Technical report, NIST, 2015. 27

Nick Littlestone and Manfred K. Warmuth. The weighted majority algorithm, 1992. 63

Tie-Yan Liu et al. Learning to rank for information retrieval. *Foundations and Trends® in Information Retrieval*, 3(3):225–331, 2009. xviii, 24, 25, 79

Rachel Tsz-Wai Lo, Ben He, and Iadh Ounis. Automatically building a stopword list for an information retrieval system. In *Journal on Digital Information Management: Special Issue on the 5th Dutch-Belgian Information Retrieval Workshop (DIR)*, volume 5, pages 17–24, 2005. 13

Julie Beth Lovins. Development of a stemming algorithm. *Mech. Translat. & Comp. Linguistics*, 11(1-2):22–31, 1968. 13

Hans Peter Luhn. A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of research and development*, 1(4):309–317, 1957. 13

Craig Macdonald, Rodrygo LT Santos, and Iadh Ounis. The whens and hows of learning to rank for web search. *Information Retrieval*, 16(5):584–628, 2013. 24

Eric Mai and Rob Hranac. Twitter interactions as a data source for transportation incidents. In *Proc. Transportation Research Board 92nd Ann. Meeting*, volume 13-1636, 2013. 3, 107, 109

Christopher D Manning, Christopher D Manning, and Hinrich Schütze. *Foundations of statistical natural language processing*. MIT press, 1999. 18

Jermaine Marshall, Munira Syed, and Dong Wang. Hardness-aware truth discovery in social sensing applications. In *Distributed Computing in Sensor Systems (DCOSS), 2016 International Conference on*, pages 143–152. IEEE, 2016. 62

Richard McCreadie, Craig Macdonald, and Iadh Ounis. Eaims: Emergency analysis identification and management system. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 1101–1104. ACM, 2016. 2, 29, 62, 63, 65, 142

Donald Metzler and W Bruce Croft. Linear feature-based models for information retrieval. *Information Retrieval*, 10(3):257–274, 2007. 25

David R. H. Miller, Tim Leek, and Richard M. Schwartz. A hidden markov model information retrieval system. In *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, pages 214–221. ACM, 1999a. 19

David RH Miller, Tim Leek, and Richard M Schwartz. A hidden markov model information retrieval system. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 214–221. ACM, 1999b. 18

Yasuhide Miura, Motoki Taniguchi, Tomoki Taniguchi, and Tomoko Ohkuma. A simple scalable neural networks based model for geolocation prediction in twitter. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pages 235–239, 2016. 32

Yasuhide Miura, Motoki Taniguchi, Tomoki Taniguchi, and Tomoko Ohkuma. Unifying text, metadata, and user network representations with a neural network for geolocation prediction. In *Proceedings of the 55th Annual Meeting* *of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1260–1272, 2017. 32

Mawloud Mosbah and Bachir Boucheham. Majority voting re-ranking algorithm for content based-image retrieval. In *Research Conference on Metadata and Semantics Research*, pages 121–131. Springer, 2015. 62, 63

Doohee Nam and Fred Mannering. An exploratory hazard-based analysis of highway incident duration. *Transportation Research Part A: Policy and Practice*, 34(2):85–102, 2000. xx, 127

Nasir Naveed, Thomas Gottron, Jérôme Kunegis, and Arifah Che Alhadi. Searching microblogs: coping with sparsity and document quality. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 183–188. ACM, 2011. 27, 54

I. Ounis, C. Macdonald, J. Lin, and I. Soboroff. Overview of the trec-2011 microblog track. In *TREC Microblog*, 2011. 27

Iadh Ounis, Gianni Amati, Vassilis Plachouras, Ben He, Craig Macdonald, and Christina Lioma. Terrier: A high performance and scalable information retrieval platform. In *Proceedings of the OSIR Workshop*, pages 18–25, 2006. 22

Ozer Ozdikis, Heri Ramampiaro, and Kjetil Nørvåg. Locality-adapted kernel densities for tweet localization. In *The 41st International ACM SIGIR Conference on Research &#38; Development in Information Retrieval*, SIGIR '18, pages 1149–1152. ACM, 2018a. 33

Ozer Ozdikis, Heri Ramampiaro, and Kjetil Nørvåg. Spatial statistics of term co-occurrences for location prediction of tweets. In *European Conference on Information Retrieval*, pages 494–506. Springer, 2018b. 33, 34

Alexander Pak and Patrick Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *LREc*, volume 10, pages 1320–1326, 2010. 27, 142

Pavlos Paraskevopoulos and Themis Palpanas. Fine-grained geolocalisation of non-geotagged tweets. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, ASONAM '15, pages 105–112. ACM, 2015. 31, 32, 34, 38, 40, 41, 44, 45, 47, 62

Francisco C Pereira, Filipe Rodrigues, and Moshe Ben-Akiva. Text analysis in incident duration prediction. *Transportation Research Part C: Emerging Technologies*, 37:177–192, 2013. 127

Jay M Ponte and W Bruce Croft. A language modeling approach to information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 275–281. ACM, 1998. 18, 19

Martin F Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980. 13, 44

Reid Priedhorsky, Aron Culotta, and Sara Y Del Valle. Inferring the origin locations of tweets with quantitative confidence. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 1523–1536. ACM, 2014. 30, 33, 34

Stephen E Robertson. The probability ranking principle in ir. *Journal of documentation*, 33(4):294–304, 1977. 17

Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. Okapi at trec-3. *Nist Special Publication Sp*, 109:109, 1995. 18, 46, 51

C Carl Robusto. The cosine-haversine formula. *The American Mathematical Monthly*, 64(1):38–40, 1957. 47

Jesus A Rodriguez Perez, Andrew J McMinn, and Joemon M Jose. University of glasgow (uog_twteam) at trec microblog 2013. *TREC Microblog*, 2013. 27

Jesus Alberto Rodriguez Perez. *Microblog retrieval challenges and opportunities*. PhD thesis, University of Glasgow, 2018. 28, 54

Jesus Alberto Rodriguez Perez and Joemon M. Jose. On microblog dimensionality and informativeness: Exploiting microblogs' structure and dimensions for adhoc retrieval. In *Proceedings of the 2015 International Conference on The Theory of Information Retrieval*, ICTIR '15, pages 211–220. ACM, 2015. 28, 54

Lior Rokach. *Pattern classification using ensemble methods*, volume 75. World Scientific, 2010. 62, 63

Stephen Roller, Michael Speriosu, Sarat Rallapalli, Benjamin Wing, and Jason Baldridge. Supervised text-based geolocation using language models on an adaptive grid. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1500–1510. Association for Computational Linguistics, 2012. 31, 34, 38

Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: Real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 851–860. ACM, 2010. 28, 142

Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523, 1988. 45

Gerard Salton, Anita Wong, and Chung-Shu Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975. 16, 45

Jörg Sander, Martin Ester, Hans-Peter Kriegel, and Xiaowei Xu. Density-based clustering in spatial databases: The algorithm gdbscan and its applications. *Data mining and knowledge discovery*, 2(2):169–194, 1998. 141

Axel Schulz, Aristotelis Hadjakos, Heiko Paulheim, Johannes Nachtwey, and Max Mühlhäuser. A multi-indicator approach for geolocalization of tweets. In *ICWSM*, 2013a. 3, 29, 30, 34, 42, 109

Axel Schulz, Petar Ristoski, and Heiko Paulheim. I see a car crash: Real-time detection of small scale incidents in microblogs. In *The Semantic Web: ESWC 2013 Satellite Events*, pages 22–33. Springer, 2013b. 3, 107, 109, 113, 115, 117, 129

Axel Schulz, Christian Guckelsberger, and Frederik Janssen. Semantic abstraction for generalization of tweet classification: An evaluation of incident-related tweets. *Semantic Web*, 8(3):353–372, 2017. 113, 129

Luke Sloan and Jeffrey Morgan. 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. 3

Ian Soboroff, Iadh Ounis, Craig Macdonald, and Jimmy J Lin. Overview of the trec-2012 microblog track. In *TREC Microblog*, 2012. 27

Karen Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21, 1972. 16

Enrico Steiger, Timothy Ellersiek, and Alexander Zipf. Explorative public transport flow analysis from uncertain social media data. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information*, GeoCrowd '14, pages 1–7. ACM, 2014. 3, 107, 109

Yeran Sun and Jorge David Gonzalez Paule. Spatial analysis of users-generated ratings of yelp venues. *Open Geospatial Data, Software and Standards*, 2(1):5, 2017. 9

Jaime Teevan, Daniel Ramage, and Merredith Ringel Morris. # twittersearch: a comparison of microblog search and web search. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 35–44. ACM, 2011. 20, 25, 39

Piyushimita Thakuriah, Katarzyna Sila-Nowicka, and Jorge Gonzalez Paule. Sensing spatiotemporal patterns in urban areas: analytics and visualizations using the integrated multimedia city data platform. *Built Environment*, 42(3): 415–429, 2016. 9

Isabelle Thomas. Spatial data aggregation: exploratory analysis of road accidents. *Accident Analysis & Prevention*, 28(2):251–264, 1996. 116

Sarvnaz Karimi Jie Yin Paul Thomas. Searching and filtering tweets: Csiro at the trec 2012 microblog track. *TREC Microblog*, 2012. 27

C. Van Rijsbergen. Information retrieval, 2nd edition. *Butterworths, London*, 1979. 13, 14, 15, 16, 17

Maximilian Walther and Michael Kaisser. Geo-spatial event detection in the twitter stream. In *Proceedings of the 35th European Conference on Advances in Information Retrieval*, ECIR'13, pages 356–367. Springer-Verlag, 2013. 28, 142

Dong Wang, Jermaine Marshall, and Chao Huang. Theme-relevant truth discovery on twitter: An estimation theoretic approach. In *ICWSM*, pages 408–416, 2016. 62

Yazhe Wang, Jamie Callan, and Baihua Zheng. Should we use the sample? analyzing datasets sampled from twitter's stream api. *ACM Transactions on the Web (TWEB)*, 9(3):13, 2015. 34

Kazufumi Watanabe, Masanao Ochi, Makoto Okabe, and Rikio Onai. Jasmine: a real-time local-event detection system based on geolocation information propagated to microblogs. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 2541–2544. ACM, 2011. 28, 142

Benjamin P Wing and Jason Baldridge. Simple supervised document geolocation with geodesic grids. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 955–964. Association for Computational Linguistics, 2011. 31, 34, 38

Qiang Wu, Christopher JC Burges, Krysta M Svore, and Jianfeng Gao. Adapting boosting for information retrieval measures. *Information Retrieval*, 13(3):254–270, 2010. 25, 88, 103, 139

Chaolun Xia, Raz Schwartz, Ke Xie, Adam Krebs, Andrew Langdon, Jeremy Ting, and Mor Naaman. Citybeat: Real-time social media visualization of hyper-local city data. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14 Companion, pages 167–170. ACM, 2014. 28, 142

Jun Xu and Hang Li. Adarank: a boosting algorithm for information retrieval. In *Proc. 30th ACM SIGIR, pages 391–398*, pages 391–398, 2007. 25, 88, 103

Zhen Yang, Guangyuan Zhang, Shuyong SI, Yingxu LAI, and Kefeng FAN. Bjut at trec 2013 microblog track. *TREC Microblog*, 2013. 27

Xiaoxin Yin, Jiawei Han, and Philip S. Yu. Truth discovery with multiple conflicting information providers on the web. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '07, pages 1048–1052. ACM, 2007. 62

ChengXiang Zhai and Hui Fang. Axiomatic analysis and optimization of information retrieval models. In *Proceedings of the 2013 Conference on the Theory of Information Retrieval*, page 3. ACM, 2013. 24

Chengxiang Zhai and John Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *ACM SIGIR Forum*, volume 51, pages 268–276. ACM, 2017. 19, 20, 45

Chao Zhang, Guangyu Zhou, Quan Yuan, Honglei Zhuang, Yu Zheng, Lance Kaplan, Shaowen Wang, and Jiawei Han. Geoburst: Real-time local event detection in geo-tagged tweet streams. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, pages 513–522. ACM, 2016a. 28, 142

Daniel Yue Zhang, Rungang Han, Dong Wang, and Chao Huang. On robust truth discovery in sparse social media sensing. In *Big Data (Big Data), 2016 IEEE International Conference on*, pages 1076–1081. IEEE, 2016b. 62

Harry Zhang. The optimality of naive bayes. *AA*, 1(2):3, 2004. 114

Siming Zhu, Zhe Gao, Yajing Yuan, Hui Wang, and Guang Chen. PRIS at TREC 2013 microblog track. Technical report, NIST, 2013. 27