Real-Time Event Analysis and Spatial Information Extraction From Text Using Social Media Data

Doctoral thesis

for the fulfillment of the requirements
for the academic degree

Doctor of Engineering (Dr.-Ing.)

Accepted by
the Department of Civil Engineering,
Geo and Environmental Sciences of the
Karlsruhe Institute of Technology (KIT)

Submitted by

M.Sc. André Dittrich

from Augsburg, Germany

Day of examination
July 7, 2016

Main referee	Prof. Dr.-Ing. habil. Stefan Hinz
Co-referees	Prof. Dr. rer. nat. Martin Breunig
	Prof. Dr.-Ing. habil. Monika Sester

Karlsruhe 2016

André Dittrich

Real-Time Event Analysis and Spatial Information Extraction From Text Using Social Media Data

PhD Thesis, May 6, 2016
Referees:
Prof. Dr.-Ing. habil. Stefan Hinz
Prof. Dr. rer. nat. Martin Breunig
Prof. Dr.-Ing. habil. Monika Sester

Karlsruhe Institute of Technology
Department of Civil Engineering, Geo and Environmental Sciences
Institute of Photogrammetry and Remote Sensing
Englerstr. 7
76131 Karlsruhe

Some ideas presented in this thesis have been partly published in the following peer-reviewed publications.

Dittrich, André and Christian Lucas (2013). "A step towards real-time analysis of major disaster events based on tweets". In: *Proceedings of the 10th International ISCRAM Conference*. Ed. by T. Comes, F. Fiedrich, S. Fortier, J. Geldermann, and T. Müller. Baden-Baden, Germany, pp. 868–874.

Dittrich, André and Christian Lucas (2014). "Is this Twitter Event a Disaster?" In: *Connecting a Digital Europe through Location and Place, Proceedings of the AGILE'2014 International Conference on Geographic Information Science*. Ed. by Joaquin Huerta, Sven Schade, and Carlos Granell. Castellón, Spain.

Dittrich, André, Daniela Richter, and Christian Lucas (2015). "Analysing the Usage of Spatial Prepositions in Short Messages". In: *Progress in Location-Based Services 2014*. Ed. by Georg Gartner and Haosheng Huang. Lecture Notes in Geoinformation and Cartography. Vienna, Austria: Springer International Publishing, pp. 153–169.

Dittrich, André, Maria Vasardani, Stephan Winter, Timothy Baldwin, and Fei Liu (2015). "A Classification Schema for Fast Detection of Locative Expressions in Social Media Data". In: *Proceedings of the 6th ACM SIGSPATIAL International Workshop on GeoStreaming (IWGS)*. Seattle, USA.

Abstract

Since the advent of websites and platforms that enable users to participate and interact with each other by sharing content in different forms, a plethora of possibly relevant information is at scientists' fingertips. Data on any type of topic – from sports news to private discussions to disaster events – occurs in a wide range of different formats such as text, pictures or videos. However, this literally never-ending stream of data, which is often referred to as *Big Data*, also introduces new challenges and calls for new and practical approaches, which are in fact able to find the information needles in this data haystack.
Consequently, this thesis elaborates on two distinct approaches to extract valuable information from social media data and sketches out the potential joint use case in the domain of natural disasters.

The first part develops an operational framework for real-time event analysis on a global-scale, exploiting the real-time stream of georeferenced TWITTER messages (so-called tweets). The spatio-temporal statistical analysis of local tweet frequencies based on an equidistant grid reliably detects significantly increased volumes on a per cell and per time step basis. Then the messages in the identified cells are subject to a keyword-based thematic classification using a domain taxonomy to possibly assign one or more domain-relevant topics, which are ranked according to a document similarity score. Events with large impact areas as well as with temporally lasting effects on social media users can be detected and monitored, as a spatial-thematic and temporal clustering is applied at the end of each time step. The framework is completed by an automatic e-mail notification containing the most important facts on the event and an ad hoc visualization.
The operational prototype for the analysis of natural disasters is evaluated based on a ground truth dataset of earthquake events.

The second part introduces the idea of automatically extracting spatial information from text in the form of spatial relations between physical entities, which are encoded by a preposition. In unrestricted natural language input however, the majority of prepositional phrases does not carry physically spatial information but conveys other meanings of the involved preposition such as temporal, modal, causal, or semantically transformed cases as in metaphors.
The developed extraction and disambiguation approach is constraint to English utterances but involves a range of pre-processing steps to extend its applicability from standard natural language to the often ungrammatical and noisy nature of social media texts.
A definition of physically spatial relations for the scope of this thesis is given and a manual decision schema is derived in order to build a solid basis for the evaluation of the approach.

The automatic approach is subdivided in three components: The identification of potentially spatial prepositions (step 1) is based on approximate regular expression matching to allow for minor spelling mistakes. The subsequent extraction of the involved entities (step 2) – the subject and object of the preposition – utilizes the idea of a candidate selection and ranking, as well as a straightforward rule-based algorithm, respectively. The resulting output together with carefully engineered linguistic features, and features relying on an external knowledgebase, finally builds the necessary input for the disambiguation process of spatial versus non-spatial prepositional use cases (step 3). The last step makes use of the combined output of several current machine learning classifiers and an informed feature selection to push the capability of the automatic approach close to human performance. The evaluation is conducted based on a hand-annotated corpus, whose consistency is verified by an annotator agreement study.

Kurzfassung

Seit dem verstärkten Aufkommen von Internetseiten und online Plattformen, welche Nutzern die Möglichkeit bieten an deren verschiedenen Inhalten mitzuwirken und untereinander zu interagieren, ist eine Fülle an potentiell relevanter Information für Forscher vermeintlich zum Greifen nahe. Von Sportnachrichten über persönliche Diskussionen bis zu Katastrophenereignissen treten Daten zu verschiedensten Themen in einer Bandbreite unterschiedlicher Formen auf, wie beispielsweise als Text, als Bilder oder als Videos. Dieser tatsächlich endlose Datenstrom, der oft auch unter dem Namen *Big Data* firmiert, bringt jedoch auch neue Herausforderungen mit sich. Folglich werden vermehrt praktische Ansätze benötigt, die tatsächlich erfolgreich die vereinzelten Informations-Nadeln in diesem Daten-Heuhaufen ausfindig machen.
An dieser Herausforderung setzt die vorliegende Arbeit an und stellt zwei unterschiedliche Ansätze vor zur Extraktion von relevanter Information aus *social media* Daten. Zusätzlich wird eine potentielle Verbindung über die Anwendungsschale Naturkatastrophen aufgezeigt.

Der erste Teil der Arbeit stellt Forschung zu einem operationellen System zur Echtzeitanalyse von Ereignissen auf globaler Ebene dar, welches den Echtzeitdatenstrom georeferenzierter *Twitter* Nachrichten (sog. Tweets) nutzt. Die raumzeitliche statistische Analyse lokaler Tweet Häufigkeiten basierend auf einem äquidistanten Gitter, ermöglicht die zuverlässige Detektion von signifikant erhöhtem Tweet Aufkommen in einzelnen Zellen pro Zeitabschnitt. Im Folgenden wird für die so identifizierten Zellen eine auf Stichworten basierende thematische Klassifikation durchgeführt. Diese nutzt eine domänenspezifische Taxonomie um die Zellen einer oder mehrerer domänenrelevanter Ereignisarten zuzuordnen und diese basierend auf einem Ähnlichkeitsmaß für Textdokumente in einer Rangfolge einzuordnen. Durch räumlich-thematische und zeitliche Aggregation relevanter Zellen nach jedem Zeitabschnitt, können auch Ereignisse mit ausgeweitetem Einflussgebiet und länger andauernden Auswirkungen auf Nutzer sozialer Medien, sowohl detektiert als auch zeitlich verfolgt werden. Abgerundet wird das System durch eine automatische E-Mail Benachrichtigung mit den wichtigsten Fakten zum Ereignis und einer ad hoc Visualisierung.
Der operationelle Prototyp zur Analyse von Naturkatastrophen wird an Hand eines *ground truth* Datensatzes von Erdbeben bewertet.

Der zweite Teil beschreibt den Ansatz, Rauminformation in Form räumlicher Relationen zwischen physischen Entitäten automatisiert aus Text zu extrahieren, welche mit Hilfe von Präpositionen beschrieben sind. Die Mehrheit der Präpositionalphrasen übermittelt jedoch keine physisch-räumliche Information, sondern bildet andere Bedeutungen ab, wie beispielsweise bei temporalen, modalen und kausalen Präpositionen oder auch bei semantisch veränderter Anwendung (z.B. in Metaphern).
Der entwickelte Extraktions- und Disambiguierungsansatz beschränkt sich auf die englische Sprache, ist aber durch eine Reihe von Vorverarbeitungsschritten auf die oft grammatikalisch fehlerhaften und "verrauschten" Texte in sozialen Medien anwendbar.
Eine Definition physisch-räumlicher Relationen im Rahmen dieser Arbeit wird dargestellt und

ein Schema zur manuellen Unterscheidung abgeleitet.
Der automatisierte Ansatz gliedert sich in drei Komponenten: Die Identifikation potentiell räumlicher Präpositionen (Schritt 1) basiert auf einer approximierten Übereinstimmung mit regulären Ausdrücken um geringfügigen Rechtschreibfehlern Rechnung zu tragen. Die folgende Extraktion der beteiligten Entitäten (Schritt 2) – dem intendierten Objekt und dem Referenzobjekt der Präposition – macht sich zum einen die Idee der Kandidatenvorauswahl mit anschließdem Ranking zu nutze, als auch zum anderen einen unkomplizierten regelbasierten Algorithmus. In Kombination mit wohldurchdachten linguistischen Attributen, sowie Attributen die auf eine externe linguistische Wissensdatenbank zurückgreifen, bildet das Resultat aus Schritt 2 die notwendige Eingangsgröße für den Disambiguierungsprozess zwischen räumlichen und nicht räumlichen Anwendungsfällen von Präpositionen (Schritt 3). Dieser letzte Schritt setzt schließlich auf die kombinierten Ergebnisse mehrerer aktueller maschineller Lernverfahren und einer fundierten Attributauswahl, um die Leistungsfähigkeit des Systems so nah wie möglich an die Fähigkeiten eines menschlichen Operateurs heranzuführen. Die Bewertung des Ansatzes wird anhand eines manuell annotierten Textkorpus durchgeführt, dessen Konsistenz durch eine sogenannte *annotator agreement* Studie verifiziert wird.

Contents

> *I am willing to expose my ignorance, hoping that it will be slightly shielded by my intentions.*

— Warren Weaver
(Pioneer of machine translation)

1 The Big (Data) Picture

In this introductory chapter the work is motivated based on the potential of so-called *Big Data* in the form of **u**ser-**g**enerated **c**ontent (UGC) on social media platforms. This will be referred to as *social media data* in the following. After explaining its distinctive characteristics, the two main goals constituting this research are described, and how they can be combined in a mutual use case. Eventually the tripartite structure of this thesis is explained.

1.1 Motivation

In 1999, Sir Timothy J. Berners-Lee, the inventor of the modern web, envisioned the idea of interactivity on the internet as "the possibility of jointly creating things or solving problems together" (Berners-Lee et al., 1999). With the advent of websites that allowed internet users to participate and interact with each other by sharing content in different forms (e.g. text, pictures or videos), this idea prospered.

The development can be roughly dated back to the years around the millennium with the launch of sites such as NAPSTER (1998), WIKIPEDIA (2001), MYSPACE (2003) and FACEBOOK (2004). Back then, these websites focused on rather long lasting and detailed content. Later on, however, other social media sites (e.g. TWITTER (2006), TUMBLR (2007), SINAWEIBO (2009) or INSTAGRAM (2010)) started aiming at sharing smaller chunks and satisfying the users' desire for more immediate information. The distinction is getting fuzzier with more sites adding different options of interaction and content sharing. In the end, all these sites have contributed to the explosion of the amount of available social media data that could be observed during the last 16 years.

Today, millions of small information chunks such as news bulletins, text messages, pictures or videos are uploaded and shared on various social media sites every day. The idea of automatically extracting valuable information from this large resource of often freely available data is obviously intriguing. However, two questions arise: "Is there in fact any relevant or valuable information hidden inside?" and if so, "How can it be retrieved in an efficient way?". In the scope of this thesis I will present two different ways of exploiting social media data in order to extract valuable information, and I will provide a mutual use case as recurrent theme for examples and a prototypical implementation.

1.2 Characteristics of Social Media Data

The term *Big Data* – often denoted as buzz word – in fact fits quite well for describing the characteristics of social media data and the involved challenges when it is processed. *Big Data* does not solely refer to the absolute data volume but rather to the so-called four **V**'s (cf. Cielen et al. (2015) and Jagadish et al. (2014)).

Volume *– the scale of the data*

A look at usage statistics of some of the most popular social media sites illustrates the tremendous volume of potential data available for analyzing – 1.55B[1] monthly active users on FACEBOOK, 5M check-ins shared per day on SWARM, 300M active users on GOOGLE+, 80M pictures posted daily on INSTAGRAM, 120M daily posts on TUMBLR, 430K hours of video uploaded daily on YOUTUBE, 600M tweets sent daily on TWITTER (statistics from Smith (2015) and Internet Live Stats (2016)).

Velocity *– the arrival rate of the data*

Another characteristic of social media data is the velocity the data comes into any kind of processing system. Whereas volume is more about the amount of data in collections, velocity refers to massive and continuous data streams. Again the statistics can provide a good grasp of the issue – 7K tweets every second on TWITTER, 900 pictures every second on INSTAGRAM, 1.4K posts every second on TUMBLR (statistics from Smith (2015) and Internet Live Stats (2016)).

Variety *– the different forms of the data*

The data the users upload or post on social media sites occurs in a large variety of formats. Besides the most common types – text, pictures and videos – the data can include locations (coordinates, addresses, etc.), **U**niform **R**esource **L**ocator (URL), dates, calendar items, animated **G**raphics **I**nterchange **F**ormat (GIF), audio files and any other kind of file format.

Veracity *– the uncertainty of the data*

In this context, veracity carries several meanings, e.g. uncertainty, reliability (unknown source), bias (subjective opinion) or noisiness. In particular concerning textual social media data the challenges are manifold. Due to the usually limited content, tweets for example suffer from spelling mistakes, ungrammatical sentences, (uncommon) abbreviations and acronyms, colloquial terms and lexical variants, mixed language use, etc. This characteristic will be referred to as noisiness (cf. Baldwin, Cook, et al., 2013; Han and Baldwin, 2011; Han, Cook, et al., 2012; Petrovi et al., 2012).

Social Media Mining The challenge that arises from these characteristics is the question how to operationalize a meaningful usage of social media data.
Consequentially, Goodchild (2007) introduced the idea of interpreting *citizens as sensors* that voluntarily contribute to the creation, assembly, and dissemination of information. Although he

[1] Throughout this work, I will use the short forms K for thousand, M for million and B for billion, i.e. a thousand million

focused on the purposive sharing of (geographic[2]) information (i.e. *crowdsourcing*), the idea quickly spread to general use cases and often a more passive role of the users – this is often called *social media mining*.

In general, social media users are *not* common sensors but provide a rather complementary type of information. Usually sensors are evenly distributed in space or placed at specifically chosen positions. Moreover, they either provide measurements at fixed points in time or can even be (remotely) triggered to provide an ad hoc measurement. Thus, they normally yield very accurate information, optimized for a very narrow and specific purpose, e.g. seismographic networks for detecting ground waves generated by earthquakes. In contrast, messages on social media platforms could be disseminated from almost anywhere in the world, at every time of day and even containing any relevant or irrelevant content, but almost always without prior knowledge of the system. Social media data has, at least in urban areas, a more spatially comprehensive character than common sensors. However, albeit the implicit advantages such as high mobility, high versatility of captured information and rapid distribution, all the inherent drawbacks have to be dealt with as well, e.g. subjectivity and varying quality and quantity. Moreover, social media mining uses this data in a passive way, that means the original intention of the message sender is not to provide data for an automated system, but rather to inform friends, followers or other platform users in general – i.e. other human beings. Thus, the information is given in a very informal way which poses even more challenges for a computational analysis.

Topic and Event Detection Although the topics discussed by social media users cover a wide range of news stories, sports, politics, disasters, etc. the major part consists of irrelevant content such as daily chatter, simple non-sense or offensive language (cf. Chen et al., 2012; Java et al., 2007; Xiang et al., 2012). The velocity and volume of social media data can nonetheless be exploited for important real-time topic and event detection. In many cases the identified messages incorporate not only topical content but also *spatial* references within the text or even geographical coordinates, allowing to assign geographic context to an event.

Especially the usage of social media data in the field of natural disasters or disaster management has been proven to yield valuable general as well as spatial information such as on-site accounts of first responders, estimated impact areas, and identification of smaller scale hot-spots within large-scale disaster areas (cf. Backfried et al., 2013; Imran et al., 2013; Stollberg et al., 2012; Tapia et al., 2013; Terpstra et al., 2012; Zin et al., 2013).

1.3 Research Goals and Use Case

Based on the identified potentials and challenges of social media data in the preceding section, now a first overview of the two main research goals addressed in the scope of this thesis will

[2]This is often called **V**olunteered **G**eographic **I**nformation (VGI).

be given. Moreover, this section will sketch out how these goals could be brought together in one mutual use case.

Goal I

The first goal consists of developing a fully automatic (operational) framework for global-scale, real-time event analysis using social media data.

The analysis should incorporate the temporal and spatial detection and identification of an event, its domain specific classification, as well as its temporal monitoring.

Goal II

The second goal consists of developing methods to identify, extract and disambiguate spatial information, encoded as so-called *locative expressions* – i.e. incorporating a preposition – from English social media text.

The methods should be able to account for the noisiness of social media data. A specific aim for the approach is the capability to semantically disambiguate between spatial and non-spatial use cases of prepositions – a problem that has not been sufficiently addressed in state-of-the-art approaches.

In the following section I will outline, how the two research goals could be combined in a joint workflow targeting the mutual use case of *natural disaster events*.

1.3.1 Natural Disasters as Mutual Use Case

In case of a natural disaster event the research goals can be expressed in a conceptual workflow consisting of two main consecutive steps:

1. mass data analysis
2. single message analysis

These two steps depict the idea of first taking a large-scale view on a disaster, i.e. gathering context information, and then zooming in and searching for local hot-spot information describing spatial scenes as in [1.1][3].

[1.1] Road cracks and wall cracks along the hospital road.

In the first step, a real-time social media stream is monitored and processed to detect natural disaster events. Context information such as the estimated impact area of the event, the time the disaster started to affect people and the classified type of the disaster (e.g. earthquake or hurricane) is derived. Additionally, incoming messages related to the disaster are continuously aggregated and stored in an event database. The first detection of a disaster event triggers the

[3] Throughout this thesis, brackets are used to denote linguistic examples numbered by chapter, whereas parentheses refer to mathematical equations.

second step, the single message analysis. Consequently, the second step can use the output of the first and benefit in different ways. First, the filtered relevant messages can be used as initial input to identify and extract **L**ocative **E**xpressions (LE) that describe local spatial scenes. Second, the context information can be employed to query other social media sources and thus retrieve more small-scale spatial descriptions which may be more relevant and also of greater detail. Lastly, the estimated impact area could serve as a-priori knowledge to resolve ambiguous place references in spatial descriptions – e.g. only by knowing the affected city, the reference to "hospital road" in [1.1] could be resolved.

A system that exhibits the aforementioned capabilities can provide real-time on-site information in case of natural disasters in populated areas. This is often faster than relying on traditional information sources and hence, it can complement them in providing more up-to-date situational awareness. Getting such a rapid understanding of the situation on-site is often crucial for disaster response organizations like the fire fighters or the Red Cross to be able to coordinate their actions and to provide help.

1.4 Thesis Outline

The remainder of this thesis is structured in three parts. As described in the preceding section, Part I and Part II will present two stand-alone methodologies with respect to their scientific goals. However, throughout the work the mutual use case of natural disasters will serve as a recurrent theme for examples. Additionally, I suggest reading this thesis in the given order as basic methods for processing linguistic data are introduced in the first part that are also applied in the second part.

Part I is subdivided into three thematic chapters that will treat the issue of real-time event analysis. Chapter 2 will first define important terms with respect to the context of this work and introduce the subject of **I**nformation **R**etrieval (IR) with respect to topic detection. Eventually, selected approaches will be reviewed to reveal the current research gap. Chapter 3 then details the developed real-time event analysis framework concerning the used input data, developed methods and technical information on the operational prototype. It depicts the scientific core contributions of Part I. Finally, Chapter 4 describes the experimental event detection results based on a publicly available earthquake ground truth data set.

Part II is subdivided into three chapters that will treat the issue of spatial information extraction from text. Chapter 5 will explain the fundamentals of spatial language and detail important related work. Selected methods and algorithms of **N**atural **L**anguage **P**rocessing (NLP) and computational linguistics that are essential for this work will be introduced. Then Chapter 6 details the developed extraction process for LEs and the approach for their semantic disambiguation. It comprises the scientific core contributions of Part II. Finally, in Chapter 7 an annotator agreement study as well as the extraction and disambiguation of LEs is evaluated and discussed based on a hand-annotated corpus.

Part III finalizes the complete work in one concluding chapter. Hence, Chapter 8 summarizes the achievements, provides concluding remarks and points out possible optimizations as well as future directions.

Part I

REAL-TIME EVENT ANALYSIS

2 Introduction

In recent years, several social media platforms have gained high popularity among researchers. Fostered by their relatively simple accessibility through **A**pplication **P**rogramming **I**nterfaces (API), the large amount of data these platforms generate have become the subject of a multitude of interesting research questions. From single user behavior to networking variations (cf. Chu et al., 2010; Hecht et al., 2011; Java et al., 2007; K. Lee et al., 2011; Lotan, 2011; Stefanidis et al., 2013) and from simple statistics to complex spatial and temporal distributions (cf. Hahmann et al., 2014; Huck et al., 2015; R. Lee and Sumiya, 2010; R. Lee, Wakamiya, et al., 2011; Leetaru et al., 2013), a lot of different aspects have been examined.
Beyond these examples, the idea of retrieving the topic(s) of current data streams by means of computer algorithms, often with respect to time and location, has gained high popularity (cf. Aiello et al., 2013; Guzman et al., 2013; Jackoway et al., 2011; Kireyev et al., 2009; C.-H. Lee, C.-H. Wu, et al., 2011; Mathioudakis et al., 2010; Naaman et al., 2011; Petrovi et al., 2012; Sankaranarayanan et al., 2009). In order to obtain the topic or subject of user interactions, the approaches often focus on extracting certain informative words to apply clustering algorithms. Some of these topics are concerned with people's general interests or habits and therefore often reveal simple patterns of re-occurrence in accordance with ordinary everyday life. Other topics however, have a more irregular type of trigger – unexpected *events*.
Depending on the characteristics of the event, a subsequent increase in general user activity as well as event-related activity can be observed (cf. Dittrich et al., 2013; Krumm et al., 2015; C. Li et al., 2012; Walther et al., 2013). Thus, from understanding the current topic of user interaction, the type of event that caused it can be inferred. Ultimately, the general goal is to automate this process of detecting and classifying events based on social media data streams. An important aspect introduced by the nature of social media data (cf. Section 1.2) is the temporal efficiency of the detection and, depending on the specific source, also the localization.

This chapter will first define the terms *event* and *natural disaster* (cf. Section 2.1), as well as the term *real-time* (cf. Section 2.2), all in the context of this work's conducted research. In Section 2.3 the general idea of IR and topic detection will be described, and some typical terminology and methods that are applied in the approach will be introduced. Finally, Section 2.4 will take a closer look at the most important related work in event analysis from social media. The respective strengths and shortcomings of the different approaches will be elaborated with particular focus on a comparison to the capabilities of the approach developed in this work.

2.1 What Is an Event?

The general primary dictionary definitions (cf. Cambridge University Press, 2015; Merriam-Webster, Inc., 2015; Oxford University Press, 2015) of the term *event* can be concentrated as

> something that happens or takes place, especially something important or *unusual*

Following Merriam-Webster, Inc. (2015) and Oxford University Press (2015) the second notions can be merged to

> a *planned* public or social occasion or activity (such as a social gathering)

Due to the context of event detection and the particular use case of natural disasters, the primary definition and the aspect of an event being something *unusual* will be taken as essential. Also the notion of an event as something that possibly triggers unusual real-world reactions, i.e. deviations from usual behavioral patterns that can be identified.

2.1.1 Events on Social Media

Several definitions have been proposed in the field of event analysis using social media data. The initial idea of topic detection kept the definition of an *event* rather general.
Y. Yang et al. (1998) claim that an event should identify something (non-trivial) happening in a certain place at a certain time, while according to Cieri et al. (2002) an event is a specific thing that happens at a specific time and place along with all necessary preconditions and unavoidable consequences.
Later on, the definitions were formulated more narrowly in order to apply them to modern social media, but leaving a somehow fuzzy border between topics and events:

- The information flow between a group of social actors on a specific topic over a certain time period.
  (Q. Zhao et al., 2007)
- A set of messages that are highly concentrated on some issues in a period of time.
  (C.-H. Lee, H.-C. Yang, et al., 2011)
- A real-world occurrence with an associated time period and a time-ordered stream of (TWITTER) messages, of substantial volume, discussing the occurrence and published during the time period.
  (Becker et al., 2011)

However, in the vast majority of publications, the authors have refrained from giving any definition of the events they want to detect or analyze using social media data.

Here, parts from the aforementioned definitions are borrowed to describe the interpretation of an event that I aim to detect through social media data. The basis can be put as

> something *observable* that happens rather unexpectedly in the real world at a specific time and in a limited region.

The term *observable* is meant to be interpreted in a way that a critical mass of people has to be affected and triggered to react to the event. Of course the idea is that, eventually, enough social media users will be activated too, as they are used as a proxy to the whole affected population. The aim is to observe the triggered reaction in the quantity of social media usage as well as in the disseminated content. This proxy is, of course, strongly biased to a very specific subgroup of the population. However, one attempt of this thesis is to reveal what is still possible with all the inherent drawbacks of the input data.
In contrast to the modern definitions given above, this thesis has a clear focus on spatially grounded events rather than only confining itself on the temporal extent. This corresponds with the desired speed of the detection – i.e. real-time as defined in Section 2.2 – and the goal of pinning the event down to a certain location or direct impact area. The timeliness of the detection is often essential for a successful localization. Naturally, the information dispersion on social media sites such as TWITTER is rapid and a message can be forwarded (i.e. *retweeted* in TWITTER terms) easily without the user having to have experienced the event himself. Thus, an event can cause an increase in topically related messages far away from its original location after some time.

2.1.2 Natural Disaster Events

Natural disasters are the main focus of the operational aspects of this work (cf. Section 3.6). Thus, a solid definition what constitutes an event as a natural disaster in general and its notion in the scope of this work is needed – that means in the context of event analysis using social media data.

First of all, a distinction has to be made between the terms natural *hazard* and natural *disaster*. The former refers to a:

> natural process or phenomenon that *may* cause loss of life, injury or other health impacts, property damage, loss of livelihoods and services, social and economic disruption, or environmental damage.
> (**U**nited **N**ations **I**nternational **S**trategy for **D**isaster **R**eduction (UNISDR), 2009)

while the latter is described as a:

> natural event [that] is so intense that people suffer and material assets are affected to a substantial degree . . . [It] depends not so much on the absolute force of the

event as on the vulnerability of the affected region.
(**CE**nter for **DI**saster **M**anagement and Risk Reduction Technology (CEDIM), 2005)

In this relation, hazards are seen as latent thread (natural process or phenomenon) and disasters are actual instances (natural event) of these hazards in vulnerable regions.
In accordance to the rather general usage of the term event in the scope of this work, also a less strict view on the term *disaster* is employed. The aim is to detect natural events caused by natural hazards that *affect* a critical mass of people without any constraint on the degree of suffering or damage. Hence, as long as there are enough people that directly experience – i.e. hear, feel, see, etc. – the natural event and are *triggered* to show a reaction, it should be detected. It can be put even simpler the other way around – if an earthquake happens in an uninhabited area it is *not* considered a natural disaster in the scope of this work.

Types of Natural Disasters According to Below et al. (2009), natural disasters can be divided into six disaster groups with several main-types, sub-types and sub-sub-types. The groups partly overlap because they are based on a "triggering hazard" logic. Their comprehensive categorization is shown in Table 2.1. The special groups "Biological" and "Extra-terrestrial" are excluded. In the topic classification step (Section 3.4), a simplified and slightly adapted version will be used as hierarchical structure – i.e. as domain taxonomy.

Natural Disaster Analysis using Social Media Several authors have shown that social media data can be used in the process of detecting and analyzing a natural disaster. The following list depicts several types of natural disasters that have been investigated and the respective references. It is meant as a quick overview but does not claim to be complete.

wild fire	De Longueville et al., 2009; Starbird et al., 2010; Vieweg et al., 2010
storm	Krumm et al., 2015; Terpstra et al., 2012
snowstorm	Krumm et al., 2015
flood	Starbird et al., 2010; Vieweg et al., 2010
tsunami	Kireyev et al., 2009
earthquake	Doan et al., 2012; Kireyev et al., 2009; Krumm et al., 2015; Sakaki et al., 2010
tornado	Sakaki et al., 2010; Saleem et al., 2014
hurricane	Hughes et al., 2009; Saleem et al., 2014

Table 2.1: Detailed categorization of disaster types according to Below et al. (2009)

Group	Main-Type	Sub-Type	Sub-Sub-Type
Geophysical	Earthquake	Ground shaking	
		Tsunami	
	Volcano	Volcanic eruption	
	Mass movement (dry)	Rockfall	
		Avalanche	Snow avalanche
			Debris avalanche
		Landslide	Mudslide
		Subsidence	Sudden subsidence
			Long-lasting subsidence
Meteorological	Storm	Tropical storm	
		Winter storm	
		Local/Convective storm	Thunderstorm/Lightning
			Snowstorm/Blizzard
			Sandstorm/Duststorm
			Generic (severe) storm
			Tornado
			Orographic storms
			Long-lasting subsidence
Hydrological	Flood	General (river) flood	
		Flash flood	
		Coastal flood	
	Mass movement (wet)	Rockfall	
		Landslide	Debris flow
		Avalanche	Snow avalanche
			Debris avalanche
		Subsidence	Sudden subsidence
			Long-lasting subsidence
Climatological	Extreme temperature	Heat wave	
		Cold wave	Frost
		Extreme winter conditions	Snow pressure
			Icing
			Freezing rain
			Debris avalanche
	Drought	Drought	
	Wild fire	Forest fire	
		Land fires (grass, bush, etc.)	

2.2 Real-Time

According to the International Organization for Standardization (2015), the adjective *real time* or *real-time* in the technical sense, is

> pertaining to the processing of data by a computer in connection with another process outside the computer according to time requirements imposed by the outside process.

In the context of this work, the data to be processed is represented by the steady flow of new information from social media streams, and the process outside the computer refers to an occurring real-world event. The time requirements imposed by the outside process on the other hand, are strongly dependent on the type of event. Therefore, no general quantification of the desired minimum time delay between the event occurrence and the detection can be given.
Consequently, the emphasis is rather on the importance of the system's *responsiveness*, i.e. the permanent capability

1. to process incoming, real-time data and
2. to yield respective results within a fixed time range.

This includes that all incoming data, i.e. each single message, is processed and contributes to the result which must be available *before* the next analysis loop starts.
No individual, absolute time ranges are set for specific event types, as it would strongly limit the generic nature of the approach. Instead, I am focusing on developing a system that taps the full velocity potential of the specific input source on a global scale (see Section 3.3.2).

2.3 Information Retrieval and Topic Detection

Commonly, IR is defined as:

> finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).
> (Manning, Raghavan, et al., 2008)

In short, general IR systems, already dating back to the 1940s, aim to identify and locate information that matches – or is at least relevant to – some user query. The system typically searches in collections of unstructured or semi-structured data, such as documents, images, videos, web pages, etc. (cf. Sanderson et al., 2012).
In the mid 1990s, however, the need for more specialized methods that could handle large streams of data rather than just static collections arose. Thus, from September 1996 through October 1997 the **D**efense **A**dvanced **R**esearch **P**rojects **A**gency (DARPA), the Carnegie Mellon

University, Dragon Systems, and the University of Massachusetts at Amherst, launched the first **T**opic **D**etection and **T**racking (TDT) pilot study. It originally started out as referring to automatic methods for the discovery of topically related entities of information in streams of data, in particular broadcast news from television and radio. Several follow-up workshops took place to investigate the current state-of-the-art in finding and following new events in a stream of broadcast news stories (cf. Allan et al., 1998).

In more detail, the research objectivesof TDT can be split into three (consecutive) subtasks:

1. *Segmentation* – finding topically homogeneous chunks in a news stream
2. *Detection* – detecting the occurrence of new events
3. *Tracking* – tracking the recurrence of known events

Today the general pattern of *Segmentation*, *Detection*, and *Tracking* in essence still holds. However, the input source has changed to mainly web based services and social media platforms where the information is disseminated via large data streams. Hence, the volume and velocity of incoming data increases. Despite the overall growth in data volume, single documents are oftentimes much shorter. Nonetheless, modern approaches in TDT as well as IR still rely on well established ideas of word occurrences and frequencies. These ideas can be aggregated under the term **V**ector **S**pace **M**odel (VSM), and will be presented in the following.

2.3.1 Vector Space Models

The basic idea underlying all VSMs is to represent each document in a collection as a vector in a vector space. Thus, vectors that are close in the chosen vector space, are expected to be semantically more similar than vectors that are far apart from each other, which in turn should be semantically distant (Turney et al., 2010). The information need expressed as a user query is then mapped to the same vector space, thus enabling distance measurements, also called *semantic similarity* in the context of IR.
According to Turney et al. (2010), the term VSM only refers to models where the values of the vector elements are derived from event frequencies, e.g. the number of times that a certain word occurs in a given context. This constraint emphasizes the derivation of VSMs from the rather general *statistical semantic hypothesis*.

statistical semantic hypothesis

Statistical patterns of human word usage can be used to figure out what people mean (Furnas et al., 1983; Weaver, 1955).

The *bag of words hypothesis* puts that in a more tangible and concrete context.

bag of words hypothesis

The frequencies of words in a document tend to indicate the relevance of the document to a query (Salton et al., 1975).

An important prerequisite for VSMs is the definition of a document unit. If the units get too small, the terms defining the topic will be distributed over several documents and important passages might be missed. In contrast, if the units are too large, systems tend to retrieve undesired matches and the relevant information is hard to find (cf. Manning, Raghavan, et al., 2008). The choice of a specific unit is strongly application-dependent. I will introduce my document aggregation approach for short messages based on space-time slices in Section 3.4.

2.3.2 Term-Document Matrix

In general mathematics, a *bag* is related to a set, in the sense that it also ignores the order of elements it contains. In a bag, however, duplicates are allowed. Thus, the sentences in this small example document

[2.1] The car is in front of the house. The owner is in the house.

can be represented as a **Bag of Words** (BoW)

$$BoW = \{car, front, house, house, in, in, is, is, of, owner, the, the, the, the\}$$

and consequently as a vector $\mathbf{x} = \{1, 1, 2, 2, 2, 1, 1, 4\}$ capturing in this case the number of occurrences of each word in the document.

In a typical setting a collection D of n documents d_j is given with $j = 1, ..., n$, i.e. a set of bags, and a *vocabulary* V of m distinct terms t_i with $i = 1, .., m$ from D. The *term-document matrix* is then defined as the $m \times n$ matrix $\mathbf{T}$ whose columns correspond to the vector representations of the documents $d_j = (t_{1,j}, t_{2,j}, ..., t_{m,j})$. The elements $t_{i,j}$ denote the weighting of term t_i for document t_j. The simplest weighting scheme is the boolean model where $\mathbf{T}$ is a binary term-document (incidence) matrix, capturing for each term t_i of V if it occurs in a certain document d_j of the collection D or not, so that the matrix entries are compiled according to

$$t_{i,j} = \begin{cases} 1, & \text{if } t_i \in d_j \\ 0, & \text{otherwise.} \end{cases} \tag{2.1}$$

However, a boolean model can only handle boolean queries and therefore just retrieve exact matches, which is often not desirable. Moreover, it is not able to incorporate cumulative evidence, that means it can not account for several occurrences of a word in a document. Hence, the more valuable approach is an algebraic model that introduces term frequency as basic weighting scheme from which more advanced weighting schemes can be derived. The term frequency (*tf*) is simply the number of occurrences of term t in document d and usually denoted by $tf_{t,d}$.
One major issue is still left in the weighting scheme though – i.e. simple term frequency assumes that each term is equally important for assessing the relevance to a query. Hence, the weighting for terms that occur in many documents of the collection D need to be scaled down as they have less discriminative power to score document relevance. For this purpose the

so-called **i**nverse **d**ocument **f**requency (*idf*) is used. According to Jurafsky et al. (2009) and Turney et al. (2010) the *idf* of a term t_j is commonly defined as

$$idf_t = \log \frac{n}{n_j} \tag{2.2}$$

where n_j is the number of documents that contain the term t_j. The logarithm is usually applied to dampen the effect of very large document collections. As I will explain in detail in Section 3.4, my approach introduces the notion of dynamic document collections but of a rather small size. Accordingly, the simple ratio is used and the logarithm is omitted, which is a better fit for the framework.
For readability's sake, the combination of the *tf* and *idf* is denoted as tf-idf$_{t,d}$ or simply tf-idf in the following. Following Manning, Raghavan, et al. (2008) the tf-idf$_{t,d}$ essentially assigns a weight to each term t in document d which is

1. highest when the term occurs many times within a small number of documents,
2. lower when the term occurs fewer times in a document, or occurs in many documents,
3. lowest when the term occurs in virtually all documents.

When applying the tf-idf$_{t,d}$ weighting scheme or one of its modifications, the patterns of frequencies that arise in the columns are a kind of signature of the corresponding documents over the vocabulary space V. Now according to the above defined *bag of words hypothesis*, these patterns capture to some degree an aspect of the meaning of the corresponding document – i.e. the *topic* of the document.

Although the BoW approach ignores word order as well as any other structural elements inherent in the text, such as punctuation, phrases, sentences, paragraphs, etc., it is widely and very successfully applied in topic detection and modeling, search engines and other branches of IR. However, due to the special characteristics of language as input data, certain pre-processing steps are often essential for the effective adoption of VSMs.

2.3.3 Linguistic Pre-Processing

In this section, I will shortly describe the typical steps for pre-processing raw textual input data. Some of the methods are also an important basis for the approach of *Spatial Information Extraction from Text*, which will be the topic of Part II. Additional steps necessary for pre-processing textual social media data will be explained in Section 3.4.

An important aspect of the VSM is a plausible notion of what constitutes a *term*, and eventually, which terms are relevant enough to be represented in the BoW capturing the semantics of a document. Although, it might be desirable to differentiate between the concepts *word* and *term*[1] in some applications, they will be used interchangeably unless explicitly noted.

[1] E.g. *New York* might be considered as one (compound) term in IR and still consist of two words.

Tokenization To extract the terms from documents, a process is employed that is called *tokenization* or sometimes more general *lexical analysis*. In the field of NLP, it is the task of cutting a character sequence (i.e. a string) into identifiable linguistic units (also called lexical entities) that constitute a piece of natural language data, e.g. a document or a sentence (cf. Bird et al., 2009).
The first and main step for most so-called *segmented languages*[2] can be as simple as splitting a sequence of characters at whitespace and dismissing the punctuation. Other languages such as Japanese and Thai do not separate the words by special characters. Hence, they require more complex approaches. However, even for English and other segmented languages, there are a lot of complicated cases (cf. Trim, 2013) such as

- apostrophes for possession and contractions
- acronyms with punctuation
- words containing periods
- different types of hyphens (end-of-line hyphen, true hyphen, lexical hyphen, sententially determined hyphenation)
- numerical and special expressions (email addresses, URLs, telephone numbers, dates, time)

The definition of a token is usually given as being only a single instance of a sequence of characters in a document that is aggregated as a useful semantic unit for processing, i.e. a term (cf. Manning, Raghavan, et al., 2008; Mitkov, 2003). Terms are also not just the set of unique tokens, though, but they are usually derived from those tokens by selection and normalization processes.

Stop Word Removal The selection of terms is commonly conducted with the help of a stop word list. Stop words are terms that appear extremely often in a certain language or domain, but are of little value for topic detection or query matching, and mostly carry little semantic content. They are excluded from the vocabulary for any further processing. According to Manning, Raghavan, et al. (2008), the general strategy for determining a stop word list is to order the terms by their frequency in the collection or another large corpus of the application domain. Then, sometimes after a manual filtering for the semantic content with respect to the domain, the most frequent terms get assigned to a *stop word list*. Figure 2.1 depicts such a first step in stop word list generation. In this case, the 20 most common terms are displayed with their respective number of occurrence from a sample of 380,683 English tweets[3] – punctuation, numeric characters and special signs (e.g. @, #, =, etc.) already excluded. Prepositions such as *to*, *in*, *of*, *for*, *at* and *on*, are included as stop words in this approach, which is justified for topic detection, however as will be described in Part II they can sometimes carry quite important semantic content.

[2] Segemented languages, are all languages that use a Latin-, Cyrillic-, or Greek-based writing system.
[3] These are all georeferenced tweets of one day from the **C**oordinated **U**niversal **T**ime (UTC) $-8\,\mathrm{h}$ time zone constraint to the United States of America, i.e. the **P**acific **S**tandard **T**ime (PST).

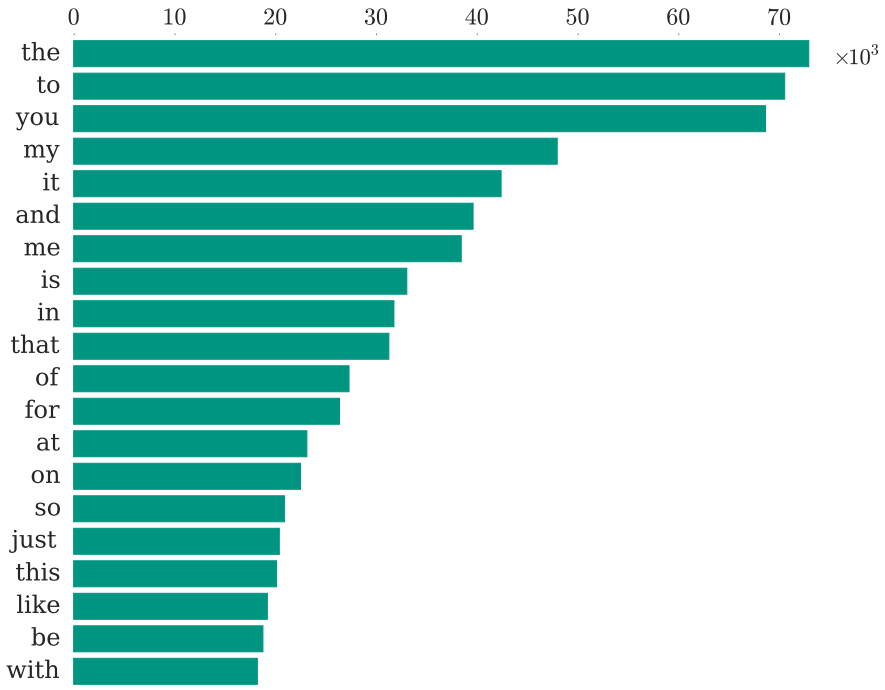

Figure 2.1: High frequency terms in a sample of 380,683 English tweets with punctuation, numeric characters and special signs (e.g. @, #, =, etc.) already excluded.

Normalization In this context normalization means canonicalizing tokens that exhibit superficial differences in their character sequences, i.e. their *surface form*, but at the same time are derived from a mutual base form – in a lexical and a semantic sense.

[2.2] Foxes often have a brownish-red color.

[2.3] The fur of a fox has several shades of a rusty red.

When a system is confronted with the input given in [2.2] and [2.3], it should be capable of grasping the topical similarity of these two documents. However, after the tokenization and stop word removal they have either none or one single token in common (that is "red", as *a* is subject to stop word removal), depending on how the applied tokenizer handles the hyphen. The steps to arrive at the desired information – that both documents share a similar topic, which can roughly be described by the words "red" and "fox" – are *case folding* and *stemming*. Case folding can be interpreted as the step that separates the lexical base form from a surface form. The common approach to achieve this is simply mapping all letters to lowercase. Thus, capitalized instances at the beginning of a sentence such as "Foxes" will match queries including the intra-sentence representation "foxes". This kind of modification might not always be advantageous. Some proper nouns, such as company names, organizations and persons, can loose their distinctiveness compared with the common words from which they are usually derived. In most scenarios, however, more complex approaches such as *truecasing*[4] did not yield the expected improvements with respect to their costs.

The second step *stemming* relies on a more complex process referred to as *morphological analysis*, the study of the inner structure of a word. Words often combine a stem and added

[4] *Truecasing* is the general term for different machine learning sequence models that make the decision of when to case-fold based on feature learning (cf. Manning, Raghavan, et al., 2008).

affixes (inflections), such as past tenses, continuous forms and plurals – e.g. “foxes” with its stem “fox” and affix “-es”. Stemming tries to reduce such inflected words to their stems (Turney et al., 2010). A more complex example is the verb “go” and its past tense form “went”. In this case a simple removal of the ending does not deliver the desired output. Stemming algorithms that can handle these advanced cases are sometimes called lemmatizers, because they reduce the inflected words to their *lemma*, i.e. their dictionary form. However, there is no strict definition, neither for stemming nor for lemmatization. So the two terms are used interchangeably throughout this work and are detailed when needed.

2.3.4 Document Similarity

Ultimately, in topic detection, the aim is to quantify the similarity of a certain document to a specific topic. This topic is either defined by a user query as in search engine applications or a predefined topic model, i.e. usually given as a set of distinctive keywords. Either way the VSM approach can be adopted and the respective representation interpreted as a BoW vector in the same vocabulary space as the document applying the same weighting scheme (here tf-idf). Thus, a score is given that describes the similarity between them.

Naturally, there exists a plethora of possible measures that could be used to derive a plausible score. In his book *Information Retrieval*, Rijsbergen (1979) showed that if proper normalization has been applied, the difference in retrieval performance using different measures is insignificant.
Nonetheless, Bullinaria et al. (2007) conducted a study with five popular distance measures as well as the *cosine similarity* measure and compared their respective effectiveness in four different tasks incorporating word (co-)occurrence statistics. The investigated distance measures were Euclidean distance and Manhattan distance (as geometric measures) as well as Hellinger distance , Bhattacharya distance, and Kullback-Leibler distance (as measures from information theory). In their settings, the cosine measure yielded the best results for all tasks.

Cosine Similarity The cosine similarity is based on the dot product (also called inner product) from linear algebra. The dot product of two vectors $\mathbf{x} = \{x_1, \ldots, x_n\}$ and $\mathbf{y} = \{y_1, \ldots, y_n\}$ is defined as

$$\mathbf{x} \cdot \mathbf{y} = \sum_{i=1}^{n} x_i \cdot y_i \tag{2.3}$$

Obviously, it can act as a similarity measure because it will be higher when $\mathbf{x}$ and $\mathbf{y}$ have large values in the same dimensions, and in contrast, it will be closer to 0 if the vectors have zeros in many different dimensions (Jurafsky et al., 2009). Still, it suffers from the shortcoming of a strong document length dependence as the relative distributions of words may be similar, but the absolute term frequencies of one document may be far larger, just because it is much longer (Manning, Raghavan, et al., 2008). The vectors are normalized to unit length to account for

the sensitivity to the absolute magnitudes of the various dimensions. Thus, the cosine similarity measure is given by

$$\cos(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{|\mathbf{x}| \cdot |\mathbf{y}|} = \frac{\sum_{i=1}^{n} x_i \cdot y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 \cdot \sum_{i=1}^{n} y_i^2}} \tag{2.4}$$

with the intuitive interpretation that it computes the cosine of the angle between two vectors. Consequently, if two documents are identical, with respect to the vector space, the cosine will be 1 and if they do not share any terms it will be 0. In VSMs the cosine is limited to positive values, as all the vector elements represent weightings based on term frequencies, i.e. positive counts.

2.4 Related Work

The body of work concerned with event analysis from social media sources comprises investigations in various research fields such as artificial intelligence, computational linguistics, computer science, electrical engineering, information science, social science, as well as research with industrial background (e.g. HP, MICROSOFT, PHILIPS, IBM, etc.). Hence, a considerable amount of approaches has been proposed to tackle the task for social media sources in general (e.g. Chae et al., 2012; Nurwidyantoro et al., 2013; Sayyadi et al., 2009; Valkanas et al., 2013; Q. Zhao et al., 2007). However, the overwhelming majority of approaches either use TWITTER exclusively or at least as their experimental dataset (Aggarwal et al., 2012; Atefeh et al., 2015; Becker et al., 2011; Benson et al., 2011; Dong et al., 2015; Krumm et al., 2015; C.-H. Lee, H.-C. Yang, et al., 2011; C. Li et al., 2012; R. Li et al., 2012; Ritter, Mausam, et al., 2012; Sakaki et al., 2010; Sugitani et al., 2013; Walther et al., 2013; Wang et al., 2013; Watanabe et al., 2011; Weng et al., 2011; W. X. Zhao et al., 2012).

Due to this huge amount of research it seems reasonable to provide a more focused review. Most of the aforementioned approaches only deal with a small aspect of event analysis. As the framework developed in this work depicts a holistic view on the matter, the three most important approaches that also cover several essential aspects are considered.
Each of the approaches is reviewed as detailed as necessary and elaborated on its respective strengths and shortcomings. The inspection is based on important features allowing an informed comparison to the approach presented in this thesis. The targeted features are

real-time capabilities	e.g. potential speed of detection, operational prototype, notification mechanism
event localization	e.g. granularity, accuracy, spatial coverage
clustering capabilities	e.g. spatial, thematic and temporal (i.e. monitoring)
event classification	e.g. distinctive class labels, language coverage, domain dependency

A Novel Approach for Event Detection by Mining Spatio-temporal Information on Microblogs, C.-H. Lee, H.-C. Yang, et al., 2011
In their proposed approach, C.-H. Lee, H.-C. Yang, et al. (2011) try to provide a comprehensive

spatio-temporal viewpoint of an event by detecting and grouping emerging topics in real-time and assign a geo-location to the identified topics. They motivate their work based on the information needs of situational awareness for event control. Moreover, they emphasize the use case of utilizing spatio-temporal information from TWITTER to support emergency planning, risk assessment and damage estimation in case of natural disasters such as earthquakes and tsunamis.

The system architecture is a two-pass model comprising the consecutive steps of *thematic topic categorization* and *spatial analysis*. The authors use an incremental **D**ensity-**B**ased **S**patial **C**lustering of **A**pplications with **N**oise (DBSCAN) algorithm to constantly group related messages into topics. The clustering is based on a dynamic feature space which keeps messages in a sliding window model. They rely on a BoW approach with dynamic term weighting based on history, originally introduced by C.-H. Lee, C.-H. Wu, et al. (2011). So-called hot topics are identified by analyzing the detected clusters – however, the authors provide no further explanation *how* they analyze the clusters. Finally, they try to exploit the spatial distribution of messages in a topic cluster, using textual location mentions to estimate the location where the event occurred. As the approach aims for detecting local events, they penalize higher amounts of different locations within one topic.

Evaluation

The approach of C.-H. Lee, H.-C. Yang, et al. (2011) lacks the proof of real-time capabilities as their analysis is carried out retrospectively. The authors use two hour chunks of the TWITTER random sample stream as their sliding window width, resulting in roughly 68K messages per chunk. It is therefore questionable if local events generate enough traffic to be detectable. The proposed method for topic detection is unbound from any domain, but it runs short of a real classification of the event type as it only yields a list of keywords. Although the localization method can be applied on a global scale, it suffers from several issues, as it is based on location mentions in the messages. Location references provided in natural language data are inherently highly ambiguous, as they are usually given in low granularity (e.g. city level) and mostly not reliable (cf. Section 3.1.3). Finally, the approach allows cluster shapes – i.e. the distribution of the distinctive terms describing the topic – to change over time. Nonetheless, it is not capable of linking temporally distant topic occurrences to one mutual event.

Beyond Trending Topics: Real-World Event Identification on Twitter, Becker et al., 2011

As described in Section 2.1, Becker et al. (2011) define an event as a real-world occurrence with (i) an associated time period and (ii) a time-ordered stream of Twitter messages of substantial volume, discussing the occurrence, and published during the time period. The goal is to identify real-world event content in an online fashion.

Incremental online clustering and filtering is conducted without initially defining the number of clusters. The threshold parameters are empirically tuned during training. The different cluster features are categorized into four groups:

- *temporal* – volume of frequent cluster terms and deviation from expected message volume

- *social* – percentages of messages containing different types of user interaction (retweets, replies, mentions)
- *topical* – describing the topical coherence of a cluster, relying on the hypothesis that event clusters have one central topic
- *twitter-centric* – usage of tags and presence of multi-word hashtags

The subsequent classification is based on a machine learning approach using the described cluster features in one hour time steps. Human annotators are employed to label clusters for both the training and testing phases. Ambiguous clusters as well as clusters where the annotators disagreed are not considered. The classifier is trained with a balanced dataset of event and non-event clusters. A variety of classifiers is considered and support vector machines yielded the best results for this setup. A following logistic regression model allows to obtain probability estimates of the class assignment. The top 20 events during one hour according to their event probability are selected for a baseline comparison. The authors show that their classifier is superior to the baseline – a Naïve Bayes classifier for text similar to the approach of Sankaranarayanan et al. (2009).

Evaluation

The approach of Becker et al. (2011) has similar shortcomings as the approach of C.-H. Lee, H.-C. Yang, et al. (2011), i.e. (i) there is no proof of real-time capabilities because their analysis is carried out retrospectively, (ii) their topic detection incorporates no real classification of the event type and (iii) the approach is not capable of linking similar topics from the distinct one hour chunks to one possibly mutual event.

A set of 2.6M TWITTER messages from February 2010 is used where the user account location information states *New York City* – i.e. they assume a predefined spatial extent[5] without incorporating any event localization in their approach.

Eyewitness: Identifying Local Events via Space-Time Signals in Twitter Feeds, Krumm et al., 2015

In contrast to C.-H. Lee, H.-C. Yang, et al. (2011) and Becker et al. (2011), Krumm et al., 2015 show that what they call *geotagged* tweets, i.e. including a pair of geographical coordinates, are sufficient for high-precision detection of local events. They conclude that it is not necessary to infer locations from tweet text or user profiles.

The key assumption is similar to the one in this work – significant local events will trigger people to suddenly write messages in a limited region and limited time. Their definition of an event is kept rather simple as "something that happens at some specific time and place".

The discretization of the earth's surface is accomplished using the **H**ierarchical **T**riangular **M**esh (HTM) according to Szalay et al. (2005). The step to a higher resolution is achieved by dividing each triangle into four smaller triangles. Four resolution levels are incorporated resulting in triangle areas of roughly $15\,\mathrm{km}^2$ to $1000\,\mathrm{km}^2$. Concerning the time discretization, the authors used uniform, disjoint intervals ranging from $20\,\mathrm{min}$ to $24\,\mathrm{h}$. They refer to the different combinations of temporal intervals and spatial extents as *space-time prisms*. The

[5] Above all they rely on the highly ambiguous location field; see Section 3.1.3 for a thorough discussion.

detection of anomalies in these space-time prisms is conducted using a regression function that predicts the amount of messages based on five features: the time of day, the day of the week and the number of tweets in the three adjacent triangles. The authors declare an event when the prediction error is larger than three times the standard deviation of the regression. The approach provides a so-called event summary, i.e. it uses the SUMBASIC algorithm by Vanderwende et al. (2007) to select the five messages that best summarize the event.
103 crowdsourced human judges were employed to produce binary ground truth for 2400 candidate events, yielding a precision of 70% for the automatic detection. They were able to increase the precision up to 93% by using the ground truth data as input for a machine learning approach[6]. The feature relevance investigation shows that the most important features for the classifier are – by far – the actual number of messages in the respective space-time prism, the normalized prediction error and the relation of the two.

Evaluation

Krumm et al. (2015) present a well-informed approach to local event detection using tweets. Their method produces a high detection precision based on machine learning and human annotated ground truth data. As their study is conducted retrospectively, they acknowledge that their current regression function has to be adapted to enable a real-time detection. The study is limited to a bounding box containing the United States of America, but it is very likely that their spatial discretization would also work on a global scale for their rather long time intervals. For real-time analysis in the range of a few minutes though, the approach might suffer from the complexity of the HTM.
On the one hand, Krumm et al. (2015) make use of the most accurate location information provided with a tweet, but then they leave their localization at the spatial granularity of the triangle. Exploiting the messages that caused the detection in a spatial clustering would largely improve the localization accuracy. Moreover, the authors do not account for events spreading over several cells, i.e. there is no spatial clustering of triangles that host the same event. Instead of a real event classification the approach only provides five representative messages. Eventually, the approach also lacks thematic clustering capabilities over distinct time intervals.

2.4.1 Research Gap

The literature review revealed several drawbacks of current state-of-the-art approaches for real-time event analysis using social media.
First of all, the generally missing operational prototype, which might be considered as a technical issue rather than a scientific one. However, I argue that the challenges in case of real-time systems go beyond simple coding tasks, but directly influence the actual feasibility of the applied methods. Throughout the design of my framework (cf. Chapter 3) the operational implementation was a key aspect. Specific implementation details will be presented in Section 3.6.
With a restriction to English messages, a global event detection is unlikely to be successful.

[6] They used a FASTRANK algorithm that learns an ensemble of decision trees.

Only C.-H. Lee, H.-C. Yang, et al. (2011) account at least for 13 different languages in their topic detection, but only in terms of removing so-called stop words (see Section 2.3.3). Their system would still interpret two equivalent keywords in different languages as distinct and thus skew the resulting topic. In order to overcome this issue, I incorporate a domain restricted translation engine for 64 common languages (see Section 3.4.2).
Furthermore, the approaches are not taking advantage of the full potential of the data source concerning the speed of detection. Only Krumm et al. (2015) use time intervals below one hour. In contrast, the framework developed in this work operates on a one-minute moving-window basis aiming for real-time event detection and classification. In terms of localization granularity, the approaches either do not use all the available information, or, in case of Krumm et al. (2015), not to the full extent.
Finally, the approaches are missing a real event classification in terms of assigning a concrete class label to an event. They either provide only the most frequent keywords or a presumably representative subsample of the messages.

Considering these shortcomings of current approaches, the need for a holistic approach of real-time event analysis using social media data becomes apparent. In the following chapter I will present my event analysis framework offering (i) real-time, global event detection, (ii) fine-grained localization capabilities, (iii) multi-lingual event classification, (iv) spatial-thematic clustering and (v) temporal monitoring as well as (vi) immediate e-mail notification and (vii) ad hoc visualization.

3 Event Analysis Framework

In this chapter, the developed framework is presented that monitors the worldwide georeferenced data stream of a social media platform to automatically analyze events in real-time. The analysis of an event includes the spatio-temporal detection and monitoring of the event's impact on the platform users as well as the classification of the event type based on the textual content of the messages sent by the users.
The main contributions of the approach are

- the applicability to a global, high-volume and real-time data stream,
- the generic implementation for domain-dependent event classification,
- the extensive multi-lingual coverage, and
- the spatial-thematic and temporal clustering capabilities.

To accomplish these tasks, the framework follows a pattern of first *filtering* the massive data stream and afterwards *aggregating* the extracted information. In a more concrete sense, the steps that the system traverses can be described as:

1. Identifying areas with unusually high message volume
   (filter step one)

2. Testing the identified areas for domain relevance and assigning class labels
   (filter step two)

3. Spatially clustering classified areas produced by the same event
   (aggregation step one)

4. Temporally monitoring spatial event clusters
   (aggregation step two)

The complete workflow will be explained in detail in the subsequent sections which are organized as follows:
Initially, the input data will be described concerning its source, format and different content aspects such as textual, temporal and locational information (see Section 3.1). Subsequently, the investigation of important space- and time-dependent, quantitative characteristics for its usage in a detection approach follows (see Section 3.2). On this basis, the developed spatio-temporal model is detailed, which uses a grid-based spatial discretization and a temporal moving window approach (see Section 3.3). A short review of alternative spatial discretization

approaches will also be provided.
The thematic event classification uses the filtered output obtained by the spatio-temporal model and conducts a multi-lingual keyword-based analysis to assign a certain class label to an event (see Section 3.4). The classification incorporates a domain taxonomy to account for the usually unspecific reporting style of social media users. Due to size and durational variabilities of events, spatial-thematic and temporal clustering is essential to capture the specific characteristics of an event (see Section 3.5).
Eventually, a description of operational aspects of the prototype such as computational resources for the real-time data processing, as well as notification and visualization capabilities will be given (see Section 3.6). Section 3.7 concludes this chapter with a short summary.

3.1 Input Data

Similar to the three approaches detailed in Section 2.4 and also most other approaches, I also use the popular microblogging platform TWITTER as data source. The reasons for this choice are mainly its large user base, the worldwide coverage[1], and its real-time nature – i.e. the platform users tend to disseminate information on something they have just experienced (cf. Java et al., 2007).
TWITTER was launched in October 2006 and has been reaching 316M monthly active users as of June 30, 2015 according to its official website (Twitter Inc., 2015b). On average, 600M tweets, i.e. 140-character messages[2], are sent via TWITTER per day (Internet Live Stats, 2016). 80% of the vast amount of users interact with the platform via mobile devices. Additionally, the message can be sent through web-based services and applications. Java et al. (2007) demonstrate the main types of user intentions to be: daily chatter, conversations, sharing information and reporting news.
TWITTER provides access to the Firehose, the real-time stream of all tweets being sent, through its Streaming API. Single tweets are received as documents in **J**ava **S**cript **O**bject **N**otation (JSON) consisting of compulsory and optional fields containing different types of alphanumerical information. The relevant fields for the current research contain the timestamp the tweet was posted (`created_at`-field), the position from where it was sent (`coordinates`-field) and the message itself, i.e. the textual information (`text`-field). Listing 3.1 shows an excerpt of a tweet that was sent after an earthquake in La Verne, California on September 19, 2013.
In the following sections the textual (Section 3.1.1) and the temporal information (Section 3.1.2) of a tweet are described. Then the various forms of embedded locational information are detailed, their respective characteristics described and the exclusive choice of the `coordinates`-field is motivated (Section 3.1.3).

[1] Except for China where TWITTER has been blocked permanently from the government since June 2009 (Guardian News and Media Ltd., 2009). In some other countries such as Iran, United Arab Emirates, Russia and Turkey, occasional blockage exists (Bender, 2015; Privax Ltd., 2015). According to the "Twitter Transperancy Report" (Twitter Inc., 2015a), other countries that have recently requested TWITTER to filter or remove tweets include Brazil, Japan, Netherlands, Germany, South Korea, Canada, India, Indonesia, Iraq, Italy, Kazakhstan, Malaysia, Mexico, Mongolia, Pakistan, Spain, United Kingdom and the United States of America. North Korea has only an intra-net, i.e. no access to TWITTER.
[2] They will be referred to as tweets or (short) messages interchangeably in the remainder of this work.

```
{
    id_str : "380664856422006784",
    text : "just felt an earthquake for the first time.",
    coordinates : {
        type : "Point",
        coordinates : [-117.88524217, 34.12811191]
    },
    created_at : "Thu Sep 19 12:09:08 +0000 2013",
    place : ...
    user : {
        location : "",
        statuses_count : NumberInt(2466),
        lang : "en",
        ...
    }
     retweeted : false,
     lang : "en",
     ...
}
```

Listing 3.1: Excerpt of a tweet in JSON-format containing locational information in the form of geographical coordinates (based on the World Geodetic System 1984 (WGS84)).

3.1.1 Textual Information

The textual information disseminated by the user is accessible through the field `text`. The content is limited to 140 characters per message. It can be observed in the examples [3.1], [3.2] and [3.3] that this limitation aggravates the typical properties that textual social media content exhibits - i.e. ungrammatical and incomplete sentences, erroneous or missing punctuation, abbreviations and acronyms (both often non-standard), spelling mistakes, slang and colloquial terms (e.g. colloquial place names) and a lot of offensive language. This characteristic needs special treatment in the form of suitable pre-processing steps using various NLP methods (for details see Section 3.4 and Section 6.5).

[3.1] Had meh too gud ofa wrkout this mornin .

[3.2] fuck school nigga imma be a drug dealer lol

[3.3] FOLLOW ME BACK pleaseeeeeeee

3.1.2 Temporal Information

The information of when the tweet was created is provided in the field `created_at` in UTC standard. The format represents the day of the month, the time and the year as numerical values. The day of the week and the month, however, are provided as text as depicted in Table 3.1. This is supposed to prevent misinterpretations in terms of zero or one-based counting

(months) and different geographical standards concerning the definition of the beginning of a week - Sunday (e.g. in the United States) or Monday (e.g. in Germany).

Table 3.1: Field names of the TWITTER timestamp format using the value from Listing 3.1

Day of week	Month	Day of month	Hour of day	Minute	Second	UTC-offset	Year
Thu	Sep	19	12	9	8	+0000	2013

Due to the processing in the TWITTER database ecosystem (e.g. message indexing for the Search API), the subsequent redistribution through the streaming endpoints is bound to be subject to a small delivery delay. Hence, the timestamp in the `created_at` – field will not match the time the message actually arrives in a monitoring system's database. To quantify the delay, the time difference is calculated for all georeferenced[3] messages obtained through the Streaming API on Monday, June 10, 2013 ($8{,}235{,}883$ messages). The latency period proves to be relatively stable with a median of $2\,\mathrm{s}$ and a 99.9th percentile of $4\,\mathrm{s}$. The measured latencies are only exceeding $10\,\mathrm{s}$ for $7.2 \cdot 10^{-2}\ \%$ of the messages. There are no significant effects on the latency neither concerning the location from where the message was sent nor the time of day it was created at.
The latency was also tested for longterm changes. Therefore, another day with a time delta of two years was investigated - Wednesday, June 10, 2015 ($8{,}757{,}570$ messages). With a median of $2\,\mathrm{s}$ and a 99.9th percentile of $4\,\mathrm{s}$ seconds, the results reveal a longterm consistency of the messages' latency. Only the amount of messages with a latency exceeding $10\,\mathrm{s}$ of $4.6 \cdot 10^{-3}\ \%$ suggests an improvement over time on the side of TWITTER dissemination mechanisms in terms of large delays. Nonetheless, the magnitude of the latency for the vast majority of messages is negligible for the approach at hand (cf. Section 3.3).

3.1.3 Locational Information

The locational information provided as part of a tweet can be represented in different ways. The formats range from unstructured to semi-structured to structured locational information.

Unstructured Format The unstructured format is given if the user provides spatial references within the textual part of the message. The process of parsing text to identify terms associated with geographic places is often referred to as *geoparsing* or toponym recognition. That means that it is a sub-problem of **N**amed **E**ntity **R**ecognition (NER), which aims at identifying relevant types of named entities in text, such as persons, organizations, places, etc. NER is also one of the main tasks in NLP and applied in other fields such as general Computer Linguistics, Text Mining, IR and **I**nformation **E**xtraction (IE) (see Section 5.3 for details).
In Part II of this thesis, several methods of NLP will be explained and employed to extract and disambiguate special cases of spatial references from text, namely spatial relations incorporating a preposition – so-called *locative expressions*.

[3] The definition of a georeferenced message for this work is given in Section 3.1.3.

Semi-structured Format The semi-structured representation refers to the field `user.location`, which the user usually populates when setting up his account on the TWITTER homepage or in official clients for mobile platforms, or in one of the innumerable third-party applications. According to Leetaru et al. (2013) the field is available in 71.4% of all tweets. The provided term often refers to the hometown of the user or another place at city level. However, as the user can set any string as input, several issues are also frequent[4].

- several toponyms e.g. "seattle // los angeles"
- vernacular place-names e.g. "the city of angels"
- unofficial spelling e.g. "losangeles"
- non-sense terms/phrases e.g. "From the Country to the beach"

Hecht et al. (2011) found that 66% of the users in their large corpus provided "any inkling of real geographic information" with a vast majority at city level followed by state level granularity. However, they included vernacular place-names as long as their human annotators could decode the information. One of their examples was "kcmo–call da po po". Their coders were able to determine that the user referred to "Kansas City, Missouri". Such decoding capabilities are still not possible with automatic systems. Moreover, the authors also included examples such as "Bieberville, California" as geographic information although the city is not real. Hence, the respective estimation is most likely too optimistic.
Even in the cases where the user provides a correct place-name such as "Frankfurt", the inherent ambiguity of toponyms is still an open issue – i.e. multiple instances of the same term in different geographic regions. The process of matching these toponyms in text to the correct physical location is called *geocoding* (cf. W. Zhang et al., 2014). A study of Leetaru et al. (2013) revealed that "[N]early one third of all locations on earth share their name with another location somewhere else on the planet [...]". The task of geocoding is an active field of research, which is, however, not in the scope of the presented work here.

In summary, the information provided in the `user.location` field is neither reliable nor fine-grained nor explicit enough to be used as georeference for the approach.
For the sake of completeness, the fields `user.time_zone` and `user.utc_offset` are also mentioned, which are often populated (78.4% and 74.9% of all tweets, respectively - cf. Leetaru et al., 2013) and represent locational information as well. However, the granularity of timezones obviously does not satisfy the needs to detect local events and, moreover, the user can easily enter false information.

Structured Format Finally, the structured format can again be subdivided into a textual (encoded in the `place`-field) and a numerical representation (encoded in the `coordinates`-field) which are, however, often closely connected with each other. Messages containing at least one form of structured locational information will be referred to as *geo-tagged* in the following. Different numbers have been reported on the percentage of geo-tagged tweets in the total

[4] Hecht et al. (2011) provide an extensive and interesting analysis of the information entered in the `user.location` field.

number of tweets, ranging from 0.7% to 3% (cf. Hecht et al., 2011; Krumm et al., 2015; Leetaru et al., 2013; Watanabe et al., 2011). Custom tests in this work showed varying percentages ranging from 1.5% to 2%.
The textual representation is an unambiguous place object as depicted in Listing 3.2 containing a unique identifier (from the TWITTER place database), the place's bounding box represented in **Geographical JSON** (GeoJSON)[5], the type of place, a short form of the name, the official country code, the URL of this place object, the country name and the full name of the place.

```
{
place : {
    id : "3b77caf94bfc81fe",
    bounding_box : {
        type : "Polygon",
        coordinates : [[
        [-118.668176, 33.704554],[-118.668176, 34.337306],
        [-117.753334, 34.337306],[-117.753334, 33.704554]]]
        },
    place_type : "city",
    name : "Los Angeles",
    country_code : "US",
    url : "https://api.twitter.com/1.1/geo/id/3b77caf94bfc81fe.json",
    country : "United States",
    full_name : "Los Angeles, CA"
    }
}
```

Listing 3.2: Place object in JSON-format as it is embedded in a tweet

The numerical representation, in contrast, is a pair of geographical coordinates in GeoJSON-format, i.e. the coordinate order is longitude first then latitude, and the reference system is the WGS84. The place object can be set manually by the user via a software menu for each tweet individually or as default setting for all following tweets. The options presented to the user are based on the last known location of the device or on its current location if the locational sensors are activated. The place object is then derived by TWITTER through reverse geocoding[6] and presented to the user in different granularities from city to country level, e.g.

- Bowie (city)
- Prince George's County (county)
- Maryland (state)
- United States of America (country)

In rare cases this is not possible and thus only a tiny fraction of approximately 0.1% of geo-tagged messages do not feature a place object. In contrast, messages just containing a place

[5] Geographical JSON Working Group (2008)
[6] the process of deriving a readable street address or toponym in text format, given a coordinate-based georeferencing (cf. Hill, 2006)

object and no data in the `coordinates`-field make up for approximately 10% of geo-tagged messages in custom tests. However, apart from the options in the menu, the user can also input any other existing place-name that is available in the TWITTER place database and thus even overrule the reverse geocoding result. Consequently, the place object does not need to represent the true location where the message was sent from, even if the tweet also contains the numerical representation, i.e. the true position. Although the place object is an explicit locational information, it still lacks the desired granularity and reliability for the approach.
Taking these considerations into account, the approach in this work exclusively uses messages featuring locational information in the form of geographical coordinates and ignores the information in the `place`-field. Hence, approximately 90% of all geo-tagged messages are used. This constraint yields a total of 8-10M messages worldwide per day. To avoid confusion, this subgroup of geo-tagged messages will be referred to as *georeferenced* throughout this work.

Accuracy of Coordinates The provided coordinates either originate from a **G**lobal **P**ositioning **S**ystem (GPS) or **G**lobal **N**avigation **S**atellite **S**ystem (GNSS) sensor of the usually involved mobile device (depending on satellite signal coverage) or they are derived through other positioning methods, e.g. Wi-Fi positioning or cellular positioning (also called cell ID method). In practice, the methods show strongly differing levels of accuracy. Modern mobile phones usually use a combination of these methods – e.g. in the form of **A**ssisted **GPS** (A-GPS) – and can reach horizontal accuracies up to $2\,\mathrm{m}$ under good multipath conditions, and accuracies of $10\,\mathrm{m}$ and above under adverse multipath conditions (see Pesyna et al., 2014). An earlier study with a more practice-related setup from Zandbergen (2009) quantified the root mean square error for an iPhone 3G to $8.3\,\mathrm{m}$ with a maximum error of $18.5\,\mathrm{m}$. In another set of experiments from Zandbergen and Barbeau (2011), the authors state that as long as a GPS position fix could be obtained, the maximum positional error never exceeded $100\,\mathrm{m}$ for indoor environments. However, the indoor positioning of mobile phones is in practice often achieved through the wifi-signal, which allows accuracies in the range from $30\,\mathrm{m}$ to $50\,\mathrm{m}$ according to Bauer (2013). Zandbergen (2009) reports larger errors with a median of $74\,\mathrm{m}$. The very rare, worst case scenario is the positioning of a mobile phone solely based on the closest cellular network towers. In theory, this could yield accuracies varying from $10\,\mathrm{m}$ to $35\,\mathrm{km}$. In practice however, the cell id is enhanced with different techniques such as the signal strength, the angle of arrival, the time of arrival, and the so-called *timing advance value*. The accuracies usually lie in the range from $100\,\mathrm{m}$ to $550\,\mathrm{m}$ (see Willaredt, 2011).
The quality of Wi-Fi and cellular positioning decreases with the distance to urban areas due to the lower density of Wi-Fi access points and cell phone towers (cf. Paek et al., 2011; Zandbergen, 2009). In contrast, the GPS/GNSS accuracy is often higher in rural areas than in larger cities with high buildings, due to less multipath effects and usually an unobstructed view of the satellites, and thus also often more visible satellites.
In conclusion, the position of a georeferenced tweet is acknowledged to range from $2\,\mathrm{m}$ to $550\,\mathrm{m}$, however, with a bias to urbanized areas the assumption that the majority of messages has a positional accuracy lower than $100\,\mathrm{m}$ seems reasonable.

3.2 Spatio-Temporal Message Flow

As described in the preceding paragraphs, TWITTER's architecture enables real-time propagation of a large amount of UGC including temporal, locational and textual information. In order to exploit the potential of the platform using statistical methods, it is essential to understand its characteristics in terms of spatial and temporal variability in message volume. Accordingly, this facilitates the spatio-temporal identification of significant increases, which are indicative to an event. Consequently, several aspects of message distribution, periodicity and recurring patterns are investigated.

3.2.1 Spatial Distribution

First, the characteristics of the spatial distribution of georeferenced tweets are analyzed.

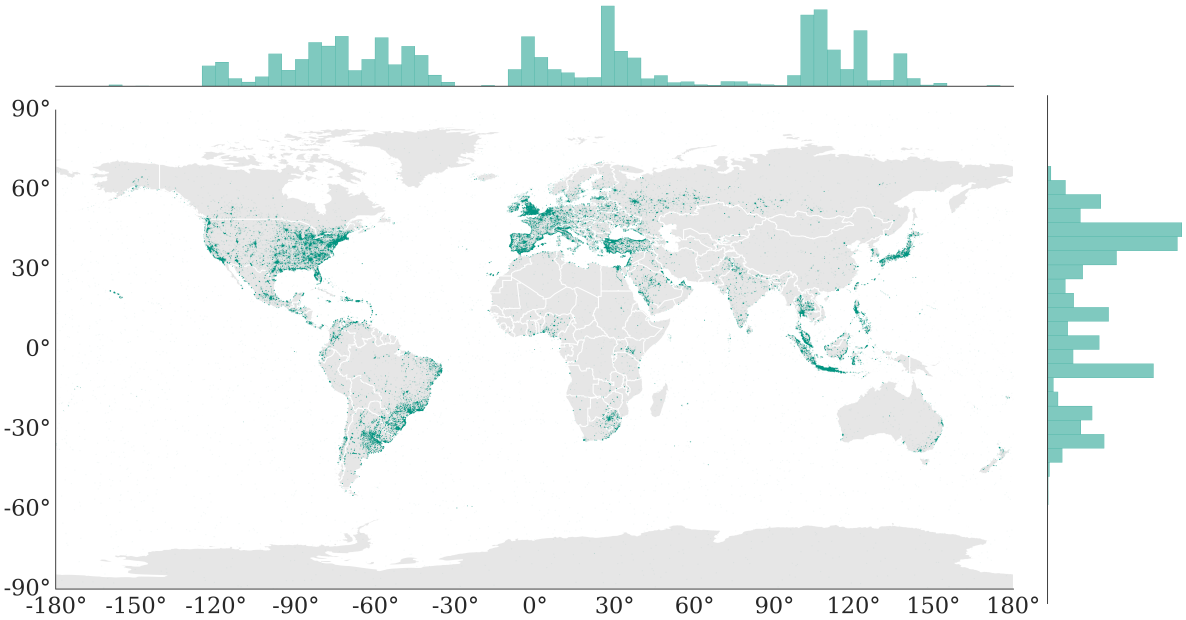

Figure 3.1: The typical global, spatial distribution of georeferenced tweets per day. On the top and on the right, the histograms illustrate the longitudinal and latitudinal distribution, respectively, in $5°$ bins

Figure 3.1[7] depicts the typical distribution of georeferenced tweets on a global scale with the distinct longitude and latitude histograms on the marginals. The histogram bins have a width of $10°$ and the respective y-axes are dropped to focus on the relative differences in the amounts. Different, active areas can be observed, such as Japan, southeast Asia, western Europe and the east coast of the United States of America. At least to some degree there is a correlation between the population density and the amount of messages sent. In smaller scales (e.g. within one country), other contributing factors such as service availability (and availability of similar other services), internet access, distribution of mobile devices, age structure, data privacy sensitization, social media affinity, etc., are rather similar and may lead to a higher

[7] All figures presenting a global map are in equidistant cylindrical projection with the equator as the standard parallel unless explicitly noted – this is also known as the *plate carée projection*.

correlation with population density (cf. Figure 3.2). A study of Malik et al. (2015) investigated the correlation of georeferenced tweets and fine-grained census data of the USA. They report biases towards younger users, users of higher income, and users in urbanized areas, as well as a surprisingly weak correlation of population density and message density. On a global scale the variability of all contributing factors is of course even higher and thus they have a very volatile influence on the amount of georeferenced messages. India, China, Pakistan and Nigeria, for example, account together for over 40% of the world's population, but only for 1.5% of all georeferenced tweets (sent per day).

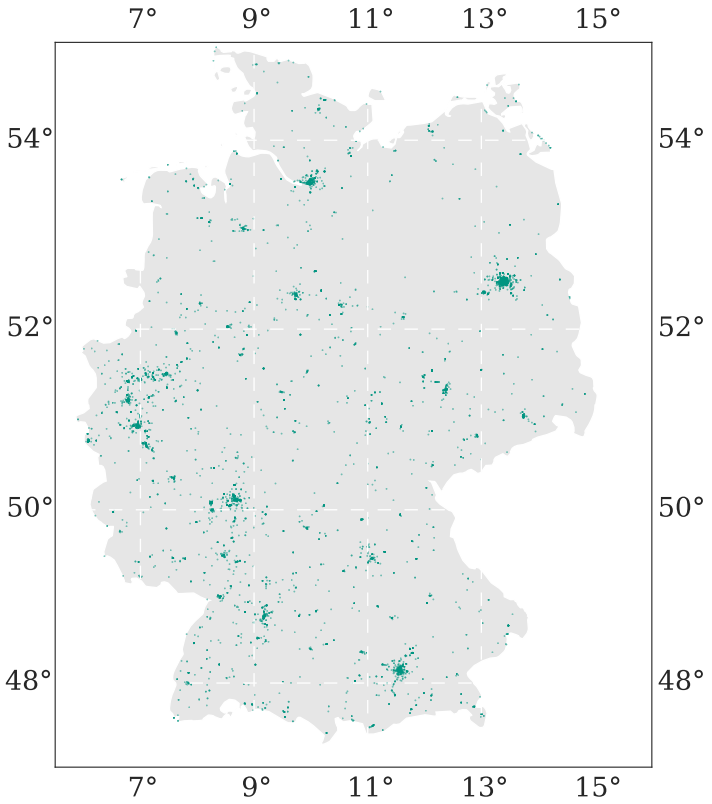

Figure 3.2: The typical spatial distribution of georeferenced tweets per day from locations in Germany (in Mercator projection)

However, for a statistical analysis of tweet volume it is not essential to understand the exact baseline generics, but to know the quantitative characteristics. Consequentially, no additional input data such as population density estimates or countries gross product are used as proxies in the approach, and instead it relies solely on the direct data, i.e. the georeferenced tweets. The fact that the set of georeferenced tweets is obviously not a representative sample of the whole population is acknowledged. But the approach will show that it can still be used as an indicative proxy for events.

On this global scale, the distribution of tweets shows no obvious recurring patterns and is rather "spiky" in nature. Accordingly, no smoothing methods are applied for modeling the spatial distribution, such as Kernel Density Estimation or regional averaging. These may lead to major errors in the baseline estimation.

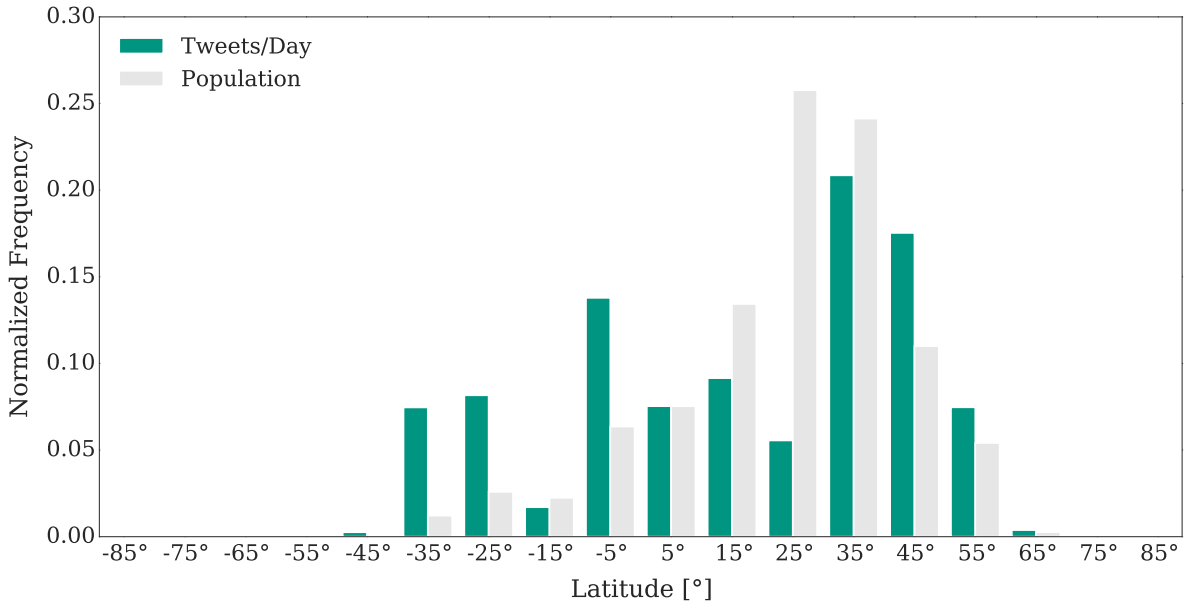

Figure 3.3: Comparison of the normalized histograms of daily sent georeferenced tweets and the population, in $10°$ latitude bins. The source of binned population data is provided by Kummu et al. (2011).

In terms of the spatial extent where georeferenced tweets occur, a rather clear constraint on a bounded latitudinal range is apparent. This is of course no surprise as it depicts approximately the zone inhabited by humans. In fact, more than 99.2% of the messages are sent between $40°$ south and $60°$ north and more than 99.6% of the world's population lives in this latitude range. Figure 3.3 shows the normalized histograms of the number of tweets sent per day and the population in $10°$ bins. The largest differences in the histograms can be observed at latitudes around $45°$, $25°$, $-5°$, $-25°$ and $-35°$. The low number of tweets around $25°$ is caused by the blockage of TWITTER in China and the relatively moderate usage in India. The remaining significant differences are in the other direction, i.e. many georeferenced tweets compared to the population (both normalized to the respective total number). The differences on the southern hemisphere can all be attributed to rather small areas of extreme tweet density – e.g. Indonesia ($\approx -5°$), southern Brazil with Sao Paulo, Rio de Janeiro and Curitiba ($\approx -25°$), the metropolitan area of Buenos Aires in Argentina and the metropolitan area of Montevideo in Uruguay (both $\approx -35°$). The difference on the northern hemisphere ($\approx 45°$) is caused by a generally higher TWITTER activity and several active regions, e.g. the northeastern United States of America (including the metropolitan areas of Chicago, Detroit, Cleveland and New York), northern Spain, northern Italy, northern Japan, and the metropolitan areas around Paris and Istanbul.

Figure 3.1 depicts each of the messages with a transparency value of $\alpha = 0.05$ so that it is possible to identify the most dense regions. In Figure 3.4 however, only messages with offshore coordinates are shown, and in a solid color. Obviously, there are in fact messages sent from almost all over the world including seas and oceans. Nonetheless, with only 0.5%, their number is rather insignificant compared to the overall amount.

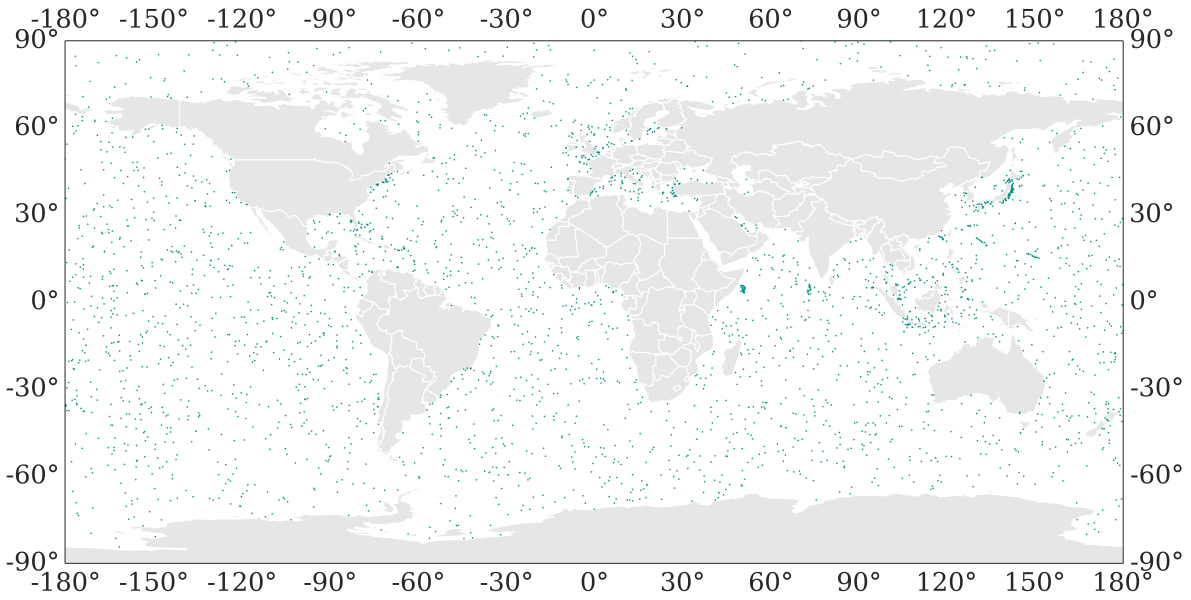

Figure 3.4: An exemplary depiction of the daily volume of georeferenced tweets from offshore locations.

3.2.2 Temporal Distribution

In order to understand the temporal characteristics of the georeferenced message flow on TWITTER, different temporal granularities are investigated to find recurring patterns. The data used to present the findings are tweets collected from the West Coast of the United States of America. The messages are constrained to one time zone to avoid skewing the statistics, as these are most likely related to daily routines of the users. The time zone is UTC−8 h, that is the PST.

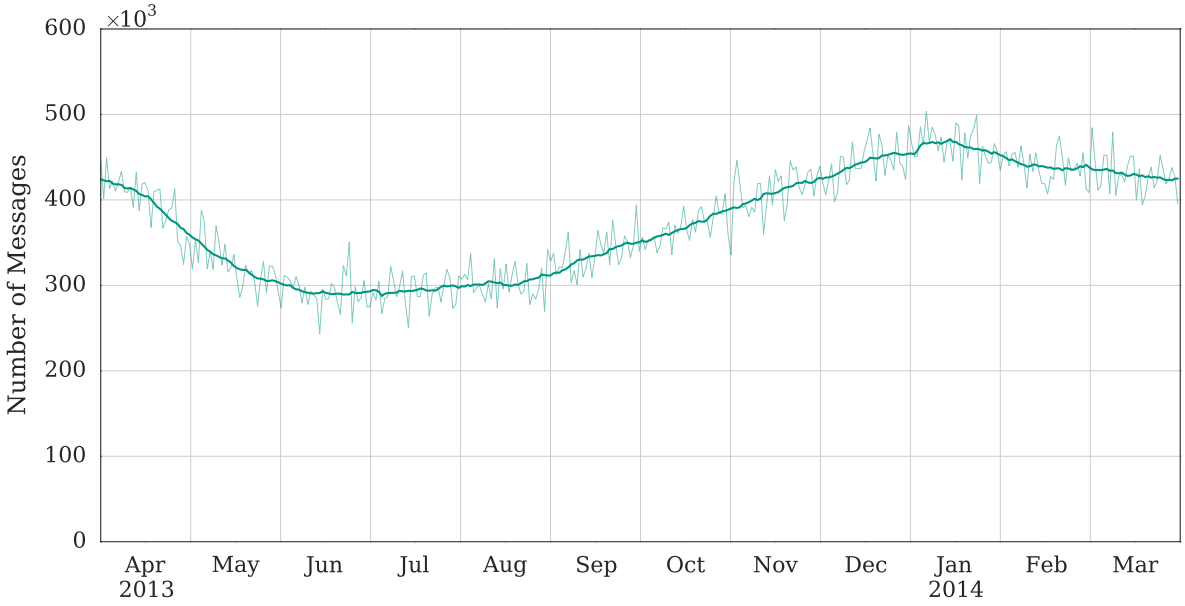

Figure 3.5: Daily amount of georeferenced messages in the PST timezone within the United States of America from the beginning of April 2013 until the end of March 2014. The dark green line is a smoothed version of the light green raw data, to show the general course.

Seasonal Effect The first visible aspect is a moderate seasonal effect of TWITTER volume over a year. December and January are the most active months, while the summer months from mid May to end of August show less tweeting activity. Figure 3.5 depicts the amount of georeferenced messages aggregated in one day periods from the beginning of April 2013 until the end of March 2014. This suggests a limited use of historical data when acquiring a robust baseline for the statistical model, i.e. the baseline generation should be adaptive to temporal change.

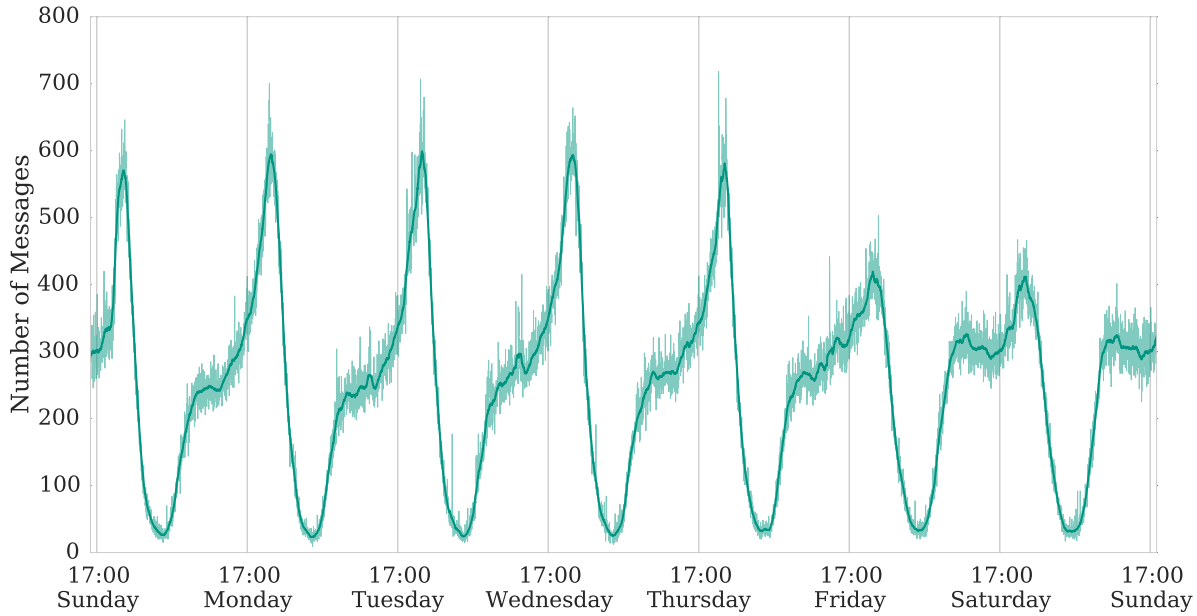

Figure 3.6: Example of the minute-by-minute amount of georeferenced messages in the PST timezone within the United States of America from Sunday 17:00 to 17:00 the following Sunday. The dark green line is a smoothed version of the light green raw data, to show the general course.

Daily Patterns The next finer temporal granularity that is investigated is one week from Sunday to Sunday. Here the expected clear pattern of daily tweet flow can be observed. Figure 3.6 shows the amount of georeferenced messages in the above described area in time steps of one minute. The dark green line is again a smoothed version of the raw data presented in light green, to emphasize the general similarities between the different days of the week. Two significantly differing patterns are visible – days during the week and days on the weekend[8].

A closer inspection yields that the normal day boundary, i.e. 24:00 or 0:00 respectively, is not suitable for decomposing the weekly message flow in periodical intervals. By looking at the time between Friday and Saturday, as well as Sunday and Monday, the time around 5 p.m. (local time) can be identified as a suitable boundary. Around this time, the daily tweet amount is very similar, no matter if the next day is a day during the week or not. That is why, the weekly flow of georeferenced messages can be compressed into two 24-hour models that are here called *weekday model* and *weekend model*, respectively. The former is applied from Sunday 5 p.m. to Friday 5 p.m., and the latter from Friday 5 p.m. to Sunday 5 p.m. (all local time).

[8]This can be generalized to days where the majority of users has to work or to go to school the next day, and days where the majority of users has the next day off. Taking into account each single holiday around the world would of course be out of scope for this work.

Figure 3.7 and Figure 3.8 depict the averaged amount of georeferenced tweets for the *weekday model* and the *weekend model* respectively (white line) – for the investigated area. The positive and negative 1σ-, 2σ- and 3σ-bounds are represented in shaded color of increasing transparency.

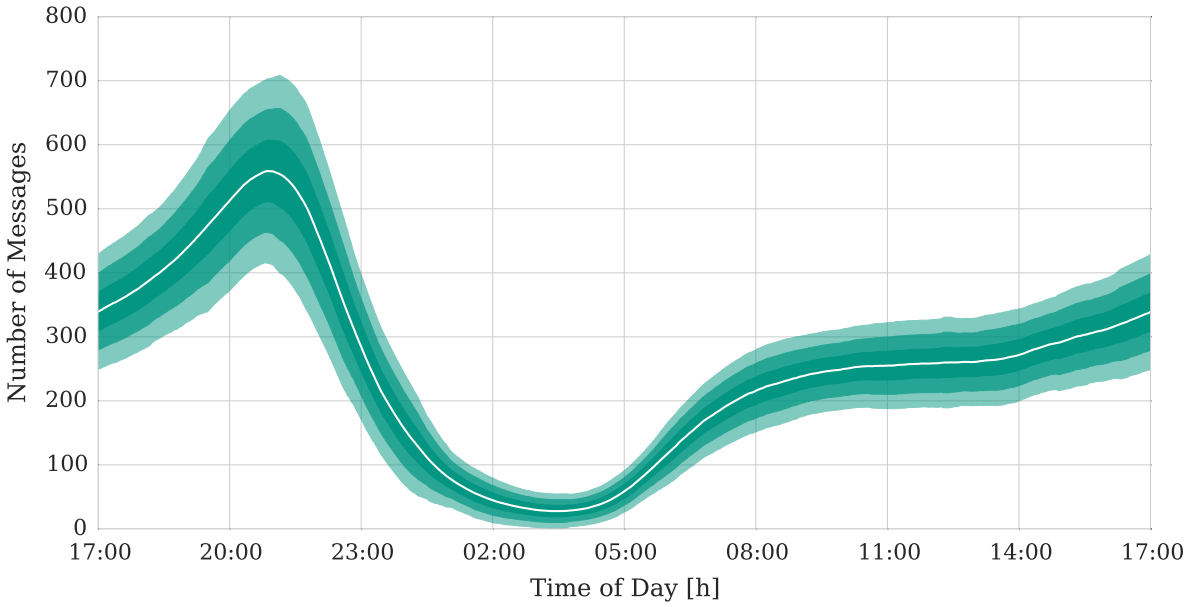

Figure 3.7: An averaged daily message flow on a *weekday* in the PST timezone within the United States of America (white line) with 1σ-, 2σ- and 3σ-bounds as shaded colors of increasing transparency

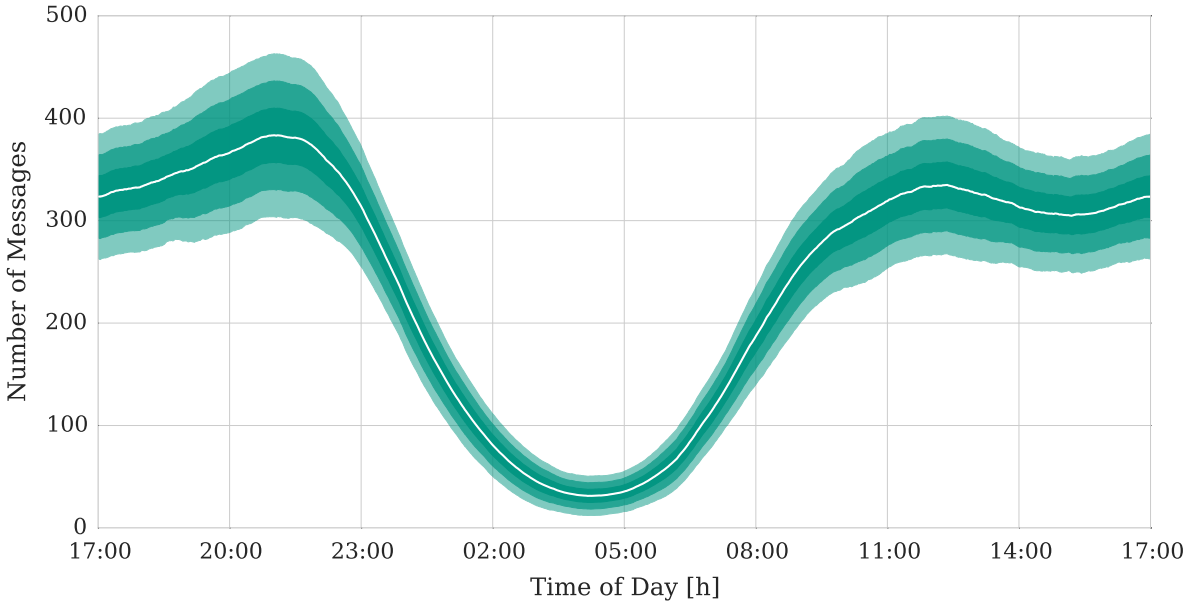

Figure 3.8: An averaged daily message flow on a *weekend* in the PST timezone within the United States of America (white line) with 1σ-, 2σ- and 3σ-bounds as shaded colors of increasing transparency

In a direct comparison as shown in Figure 3.9, the different characteristics of the two models can be seen and probable causes can be inferred. The peak of the *weekend model* is slightly shifted by approximately 30 minutes to the right and is much lower in terms of the absolute amount of messages per minute. This could be caused by a large percentage of users going out on evenings before a day off from work, for example, to eat in a restaurant, have a drink in a

bar, meet with friends or to go to the movies. The broad lows – most likely depicting the time most users are asleep – are also shifted against each other. This suggests that users stay up longer, and stay in bed longer on a weekend than on weekdays. The increased activity between 4 a.m. and 8:30 a.m. during weekdays also supports this. Accordingly, the phase between 8:30 a.m. and 4 p.m. represents the time when most users are at work or in school and thus the activity is higher on weekends during this time.

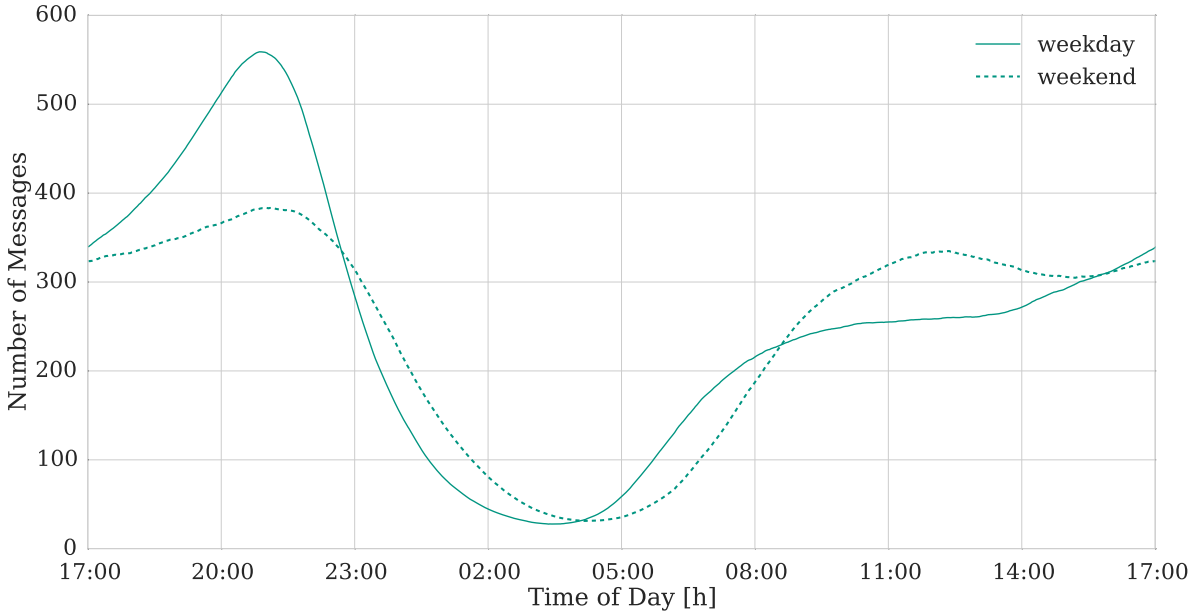

Figure 3.9: A comparison of the average daily flow of TWITTER messages for weekdays (solid line) and weekends (dashed line) in the PST timezone within the United States of America.

3.3 Spatio-Temporal Model for Global Event Detection

Based on the findings of the preceding sections concerning the spatial and temporal distribution of georeferenced messages on TWITTER, the spatio-temporal model for global event detection is introduced.

The specific characteristics of the input source are exploited to obtain the most efficient identification of significantly increased message volume in space and time. These are interpreted as first indicators for potential events. As identification conveys the sense of assigning an absolute time and location to an event – or in this case the time when users were first affected and the estimated areal location of these users – suitable discretization methods for both space and time are needed. After building a baseline model of *normal* message volume for each discrete region and each discrete time interval, the current amount of each region can then be compared to the respective, statistically derived threshold and thus potential events can be identified.

3.3.1 Spatial Discretization

The basis of the spatial discretization method is a constant global grid – basically a two-dimensional histogram – which is aligned with the meridians and parallels of the earth. This is somewhat similar to an equidistant cylindrical projection with the equator as the standard parallel, known as the *plate carée projection*. With this uniform tessellation, no implicit assumptions are made about how population or other factors introduce variance to the extent of an event (cf. Krumm et al., 2015). Figure 3.10 exemplifies such a grid with constant spacing of $10°$. The same grid is shown in Figure 3.11, but in the equal-area, pseudocylindrical projection from Mollweide that is based on ellipses and gives a more realistic view on the actual shapes of the cells on the earth's surface (cf. Snyder, 1993). Apparently, the grid distortion gets more severe when approaching the pole, i.e. the covered area of the cells gets smaller and the shape gets elongated in longitudinal direction. In fact, the equidistant rectangular tessellation suffers from singularities at the poles.

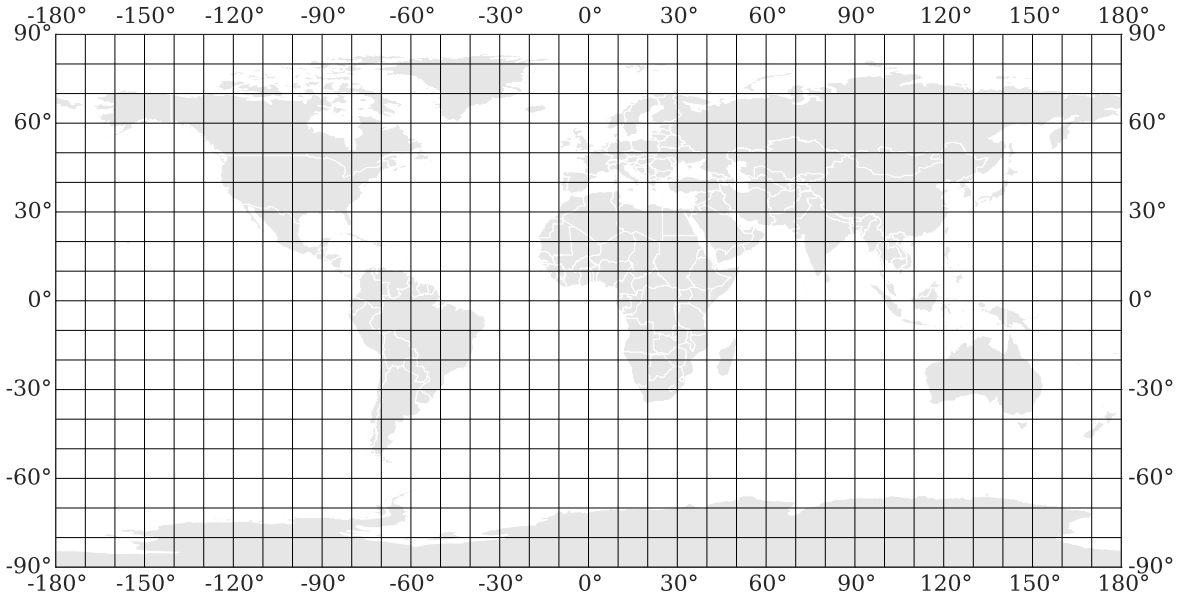

Figure 3.10: A depiction of a global regular grid with a spacing of $10°$ in latitude and longitude direction in *plate carée projection*, i.e. increasingly distorted area representation towards the poles.

However, as Section 3.2.1 revealed, the major focus of the framework, and therefore also its expected detection capabilities, lies mainly in the range from $-40°$ to $60°$. Moreover, I again emphasize the goal of this approach as defined in Section 2.1.1 where I postulated that a critical mass of people has to be affected by an event to be relevant in the scope of this work. Figure 3.12 shows the change in area size of the grid cells along the complete latitude range. In the critical range, the maximum area scaling factor reaches a value of approximately $1/2$. Thus, the grid cells at a lower latitude of $60°$ have approximately half the size of the $0°$ latitude cells.

Formal Interpretation Formally, the simple grid covering the earth is denoted by $\mathbf{G}'$, using matrix style notation, which means that a specific single cell is represented as $\mathbf{G}'(r, c)$. The indexes $r = 1, \ldots, m'$ and $c = 1, \ldots, n'$ represent the row and the column of the respective

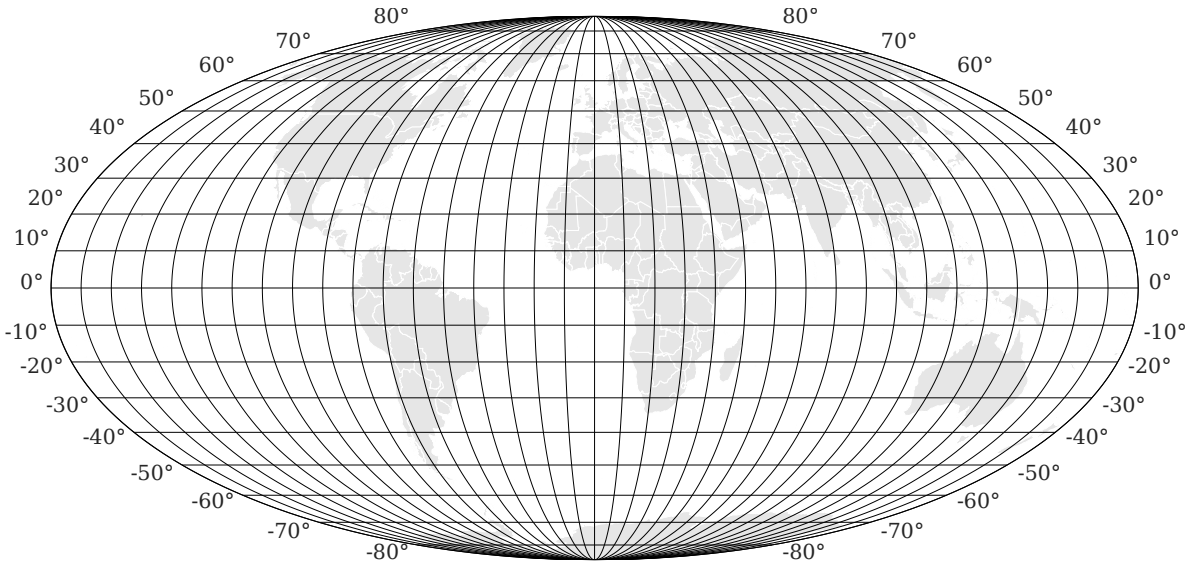

Figure 3.11: A depiction of a global regular grid with a spacing of 10° in latitude and longitude direction in Mollweide projection to better illustrate the changing shape and area of the grid cells towards the poles.

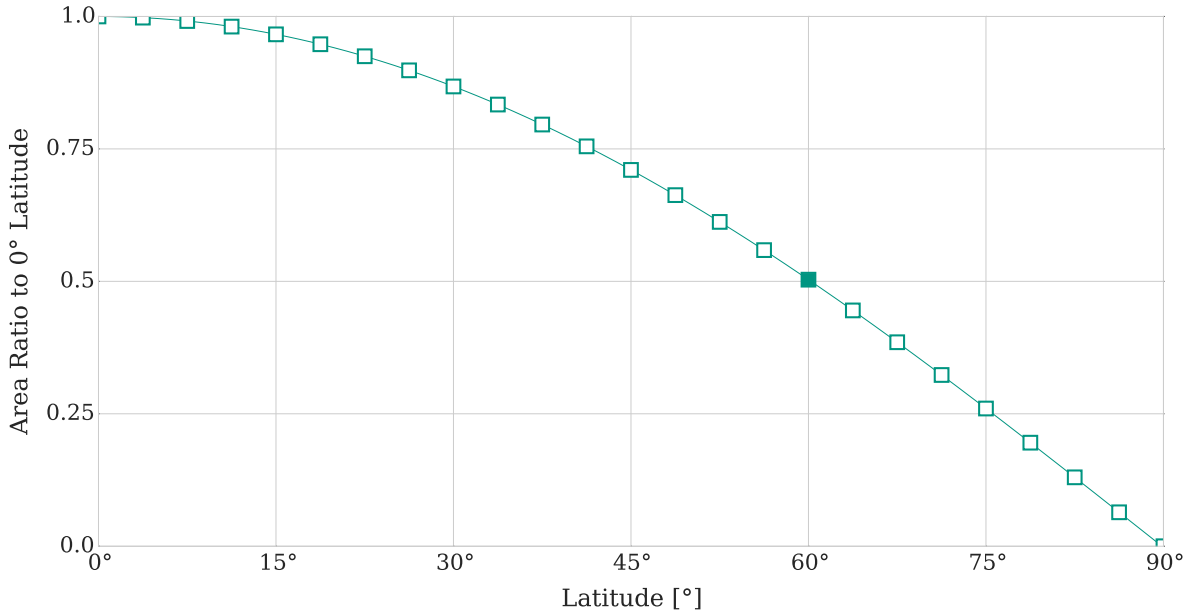

Figure 3.12: The ratio of the grid cell areas of different (lower) latitudes and the grid cell area at 0° latitude.

cell. The number of rows m' and the number of columns n' depend on the chosen cell size Δg, which is equally defined in units of degrees [°] for latitudinal and longitudinal spacing. As the extent of the earth is fixed to 360° and 180° in terms of longitude and latitude range respectively, m' and n' are given as

$$m' = \frac{180}{\Delta g} \quad \text{and} \quad n' = \frac{360}{\Delta g}. \tag{3.1}$$

The actual value of Δg is an application-dependent compromise of several factors. On the one hand, a suitable and representative spatial granularity of the targeted event type is needed. On the other hand, it needs to be taken into account if the input source is able to provide

enough messages in a certain space and time range to allow statistical analysis and eventually IR methods. For the operational prototype, Δg is accordingly set to $0.25°$, leading to $m' = 720$ and $n' = 1440$. In the latitudinal focus range this yields approximate cell areas from $387\,\mathrm{km}^2$ to $769\,\mathrm{km}^2$.

Oversampling The spatial discretization of the earth's surface with static boundaries exhibits a disadvantage concerning the detection capability. The problem arises if an event generates messages that are adversely distributed between neighboring cells in such a way that the respective increases are not significant enough on their own. As the cell boundaries are not correlated with the typical spatial distribution of messages, they may also split common clusters and distort the "real" underlying distribution to some extent. This is a common issue when quantifying data in discrete bins.

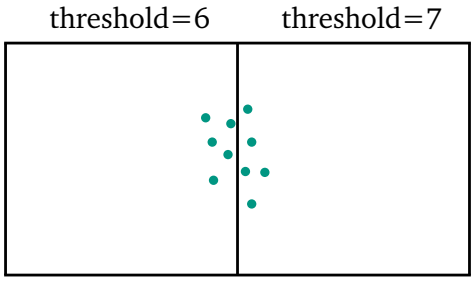

Figure 3.13: Depiction of an example event situation yielding messages in two adjacent grid cells without surpassing neither threshold.

Figure 3.13 illustrates the problem with two neighboring cells. Let us assume that an arbitrary event caused all ten displayed messages represented as dots. The event would not be detected as both thresholds are higher than their respective number of messages, here five. In order to still be able to achieve the detection of the event in the desired resolution Δg, oversampling is applied to acquire knowledge on the area around the edges.
The spacing used for oversampling – again in longitudinal and latitudinal direction – is $\Delta g/2$. Hence, the resulting cells are only covering (approximately[9]) $1/4$ of the original cells as depicted in Figure 3.14. They represent a new grid $\mathbf{G}''$ with $2m'$ rows and $2n'$ columns which represents an intermediate step during the data capturing.

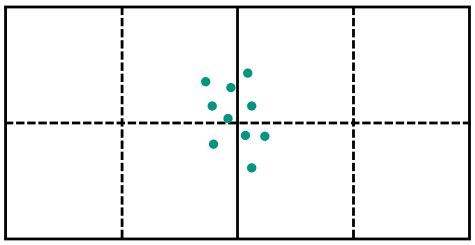

Figure 3.14: Depiction of oversampling the original grid by a factor of two in latitudinal and longitudinal direction to overcome the edge problem.

From $\mathbf{G}''$ the information can be retrieved that is needed to analyze all possible edge areas of the original cells – i.e. the edges between two vertical or horizontal neighboring cells, as well

[9] Exactly $1/4$ in geographical coordinates but not in terms of the covered area

as the area where four cells meet. In case of the simple fictitious example, it is now possible to detect the event because the threshold of the overlapping cell as depicted in Figure 3.15 is known.

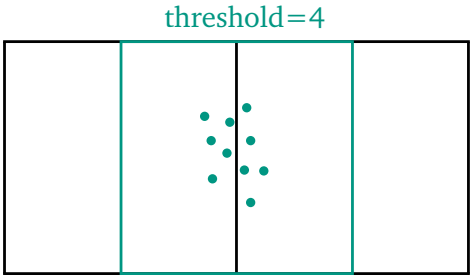

Figure 3.15: Depiction of the same event as in Figure 3.13, but detected due to the oversampling information for the region covering half of the respective original cells.

In order to obtain this information, a simple two-dimensional convolution is used – a fundamental mathematical operation often used in image processing, that produces linear combinations of input pixel values. The basic idea is to slide a so-called kernel or filter over the input grid through all positions where the kernel fits entirely into the input grid. The general formula for a linear convolution for discrete, two-dimensional functions is given as (adapted from Burger et al. (2008))

$$\mathbf{I}'(u,v) = \mathbf{I} * \mathbf{H} = \sum_{i=\infty}^{\infty} \sum_{j=\infty}^{\infty} \mathbf{I}(u+i-1, v+j-1) \cdot \mathbf{H}(i,j) \tag{3.2}$$

In this case the discrete function $\mathbf{I}$ equals the oversampled grid $\mathbf{G}''$ and the kernel $\mathbf{H}$ is given as a 2×2 matrix

$$\mathbf{H} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \tag{3.3}$$

Equation (3.2) can then be simplified to

$$\mathbf{G}(r,c) = \sum_{k=1}^{2} \sum_{l=1}^{2} \mathbf{G}''(r+k-1, c+l-1) \tag{3.4}$$

with the indexes now defined as $r = 1, \ldots, m$ and $c = 1, \ldots, n$ with $m = 2m' - 1$ and $n = 2n' - 1$. This is in fact a summation of the four cell values covered by the kernel at each possible position.

The resulting $m \times n$ grid $\mathbf{G}$ actually represents *overlapping* regions on the earth's surface of width and length Δg and no discrete cells anymore. They contain the sum of the respective cell values of $\mathbf{G}''$. Later on, the process accounts for possible multiple detections caused by the same messages.

The prototype setting of $\Delta g = 0.25°$ leads to $m = 1439$ and $n = 2879$ as the size of $\mathbf{G}$. Finally, $\mathbf{G}$ is the actual grid, which is used in terms of the statistical analysis in Section 3.3.3.

3.3.2 Temporal Discretization

As mentioned before, the second requirement for a spatio-temporal model to identify increased message volumes, is a bounded time interval that can be analyzed.
The selection of such a time interval, as well as the dimensions of the spatial cells described in the previous section, is subject to external conditions. With the most important one being the specific input source, i.e. its characteristics in terms of volume and velocity. The logical choice of a time interval in combination with a desired spatial resolution is definitely a compromise between statistical robustness and speed of detection. Simply put, a time interval that is too small for the chosen cell size will not yield enough messages for a solid analysis of the volume and the content of the messages. On the other hand, a time interval that is too long will not satisfy the criterion of a time efficient detection, and may lead to a system that is not sensitive enough for smaller events.
Another factor of a framework aiming at real-world functionality is the capacity of satisfying the definition of real-time systems given in Section 2.2. The definition specifically demands that the system has to yield results *before* the next analysis loop starts. Accordingly, even if the input source provided enough messages, there still were lower bounds for the applicable time interval. The maximum time the system needs to process the data from one time interval is denoted as t_{max}.

Formal Interpretation A certain length Δt is chosen for a time interval t, during which all incoming messages are mapped to their corresponding cells in $\mathbf{G}''(r, c)$. At the end of a time interval, each cell contains the number of messages that have occurred within its boundaries, i.e. the system simply increments the respective counter for each message arrival according to its location.
Although Δt is selected to be as short as possible considering the message volume, the detection time can still be accelerated. Therefore, a *temporal moving window* approach is applied with windows equal to the time intervals and fixed time steps ts. At the end of each time step ts_i, the respective time interval t_i is evaluated. Two constraints apply for the length of the time steps Δts:

$$\Delta ts > t_{max} \quad \text{and} \quad \Delta ts = \frac{\Delta t}{k}$$

where $k \in \mathbb{N}$ and for practical reasons $k \geq 2$. Consequently, the inequality

$$\Delta t > 2 \cdot t_{max}$$

has to be satisfied for the optimization to be applicable for a certain input source and implementation. Figure 3.16 depicts the process for 5 time intervals that cover 9 time steps in this example, i.e. $k = 5$.
In the prototype, Δt is set to $1\,\text{min}$ and Δts to $10\,\text{s}$, i.e. 8640 analysis steps per day. The time intervals are based on experience with the data source and the chosen spatial resolution. In contrast, the time steps are subject to the temporal performance capacity of the current implementation of the prototype.

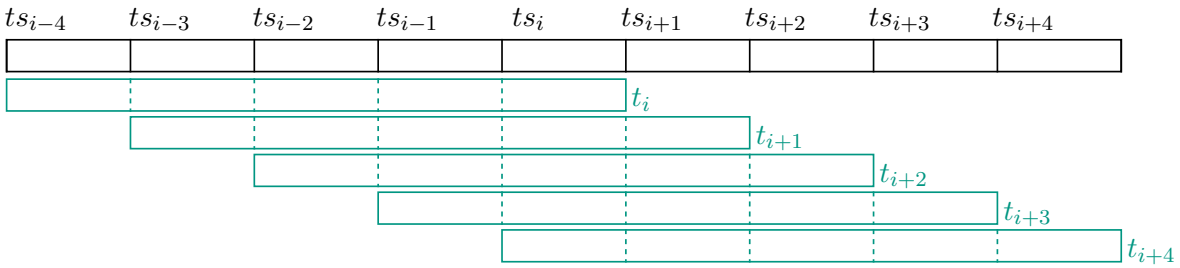

Figure 3.16: Abstract visualization of the temporal moving window approach to optimize the detection efficiency. The original time interval is split into an integer-valued number of time steps at which the preceding time interval is processed.

This *temporal moving window* idea has similarities to the oversampling of space. Assuming that an event produces messages distributed over the boundary of two time intervals, then this event might not be detected with the use of temporally disjoint analysis entities. But now the single time intervals are overlapping each other, i.e. possible increases not large enough to be significant, can often still contribute to a detection in the next time interval.
The advantage in contrast to the spatial case is the fact that the message volume of one cell is temporarily much more stable than the volumes of neighboring cells in space. Therefore, a weighted average can be performed, instead of capturing the baseline statistics for each time step. The weighted average is simply a linear interpolation of the values in the two respective disjoint time intervals. In the majority of cases, the introduced error stays in the magnitude of one to two messages.

Time Zones In Section 3.3.2 it was already briefly mentioned that temporal data from different time zones should not be mixed. Otherwise the resulting model would be blurred in some parts.
The problem is that by introducing different models for specific days during a week, the system always needs to know which model has to be applied to which parts of the world, in other words to which parts of the grid. For example, at Friday 5 p.m. UTC, the *weekday* model has to be applied to all cells within time zones west of UTC, that is UTC$-1\,\mathrm{h}$ to UTC$-12\,\mathrm{h}$. At the same time the *weekend* model has to be applied to the rest of the cells, i.e. UTC to UTC$+14\,\mathrm{h}$.
As two models were identified to be necessary to represent the daily tweet volumes reliably, two time spans have to be handled that traverse the grid from right to left during one week and depict the time when the *weekday* model has to be applied to some part of the grid and the *weekend* model to the other part.
The first time span starts at Friday 3 a.m. UTC, when the weekend model starts in the time zone UTC$+14\,\mathrm{h}$ and goes on until Saturday 5 a.m. UTC, i.e. when the weekend model started everywhere.
The second time span starts at Sunday 3 a.m. UTC, i.e. when the weekend model ends in the time zone UTC$+14\,\mathrm{h}$ and goes on until Monday 5 a.m. UTC, i.e. when the weekend model ends everywhere.

So far the approach has not taken into account daylight saving time – that is the advancing of the clocks by one hour during summer months. However, as the two models are rather close in

the transition phases, the introduced inaccuracies are not too large.
Technically, so-called masked arrays in PYTHON are used to assemble the current grid from the two separate models. The masks are pre-compiled and are adapted in $15\,\mathrm{min}$ time steps, as this is the smallest existing difference between time zones.
The time zone information is obtained from the Internet Assigned Numbers Authority (IANA) time zone database via the PYTHON library PYTZWHERE[10]. Figure 3.17 illustrates Friday 13:00:00 UTC, i.e. when all timezones with a positive offset larger or equal to $4\,\mathrm{h}$ are represented by the *weekend* model. Hence, all cells where the weekend has already started are green and the rest is not colored. The grid resolution is the one used in the prototype at the oversampling rate of $\Delta g/2$, i.e. the cells are $0.125° \times 0.125°$.

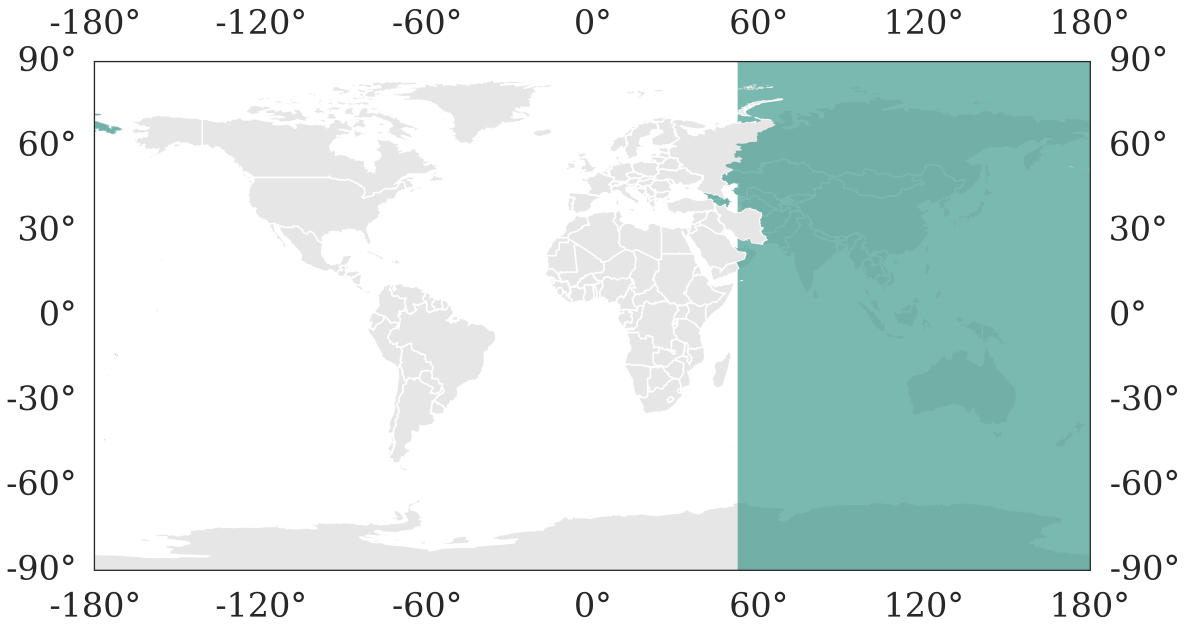

Figure 3.17: Visualization of a so-called masked array for time zone handling at Friday, 13:00:00 UTC in a spatial resolution of $0.125° \times 0.125°$. The *weekend* cells are green and *weekday* cells are not colored.

3.3.3 Frequency Analysis

Now that suitable discretization methods for time and space are established, the respective thresholds have to be modeled in order to detect significant message increases, i.e. potential event cells. Consequentially, appropriate distributions to model the typical number of messages and its variability have to be chosen. So far, only increases are considered to be indicative to an event, as a significantly decreased tweet volume could be the result of several unknown and mostly non-detectable causes.

Count Data Models An important aspect for choosing the right statistical model is the nature of the data that is analyzed. In this case, the variable to model is the number of messages in

[10] Available from https://github.com/pegler/pytzwhere and based on work done by Eric Muller – an up-to-date shapefile of the timezones of the world available from http://efele.net/maps/tz/world/.

a cell that arrived during a time interval[11]. This is usually referred to as *count data*, i.e. the realization of a nonnegative integer-valued random variable.

The most common way to model count data is the Poisson distribution (cf. Cameron et al., 2013), a discrete probability distribution, given as

$$P_\lambda = \frac{(t\lambda)^k}{k!} e^{-t\lambda} \tag{3.5}$$

with $t, \lambda > 0$ and $k \in \mathbb{N}$. The distribution has only one parameter λ that represents the expected value[12] as well as the variance, this characteristic is called equidispersion. Here it represents the expected amount of messages in a cell during a certain time interval and is also called intensity or rate. The parameter t denotes the exposure or the length of time the messages get counted. Here, t can be set to 1 as it is always the exact same exposure and thus no adjustment for varying exposures is needed. Equation (3.5) is then simply given as

$$P_\lambda = \frac{\lambda^k}{k!} e^{-\lambda} \tag{3.6}$$

Hence, the Poisson distribution provides the probability of the occurrence of a certain number k of messages in the time interval. In Figure 3.18 a Poisson distribution is shown for a typical cell with $\lambda = 1.5$ with its **cumulative distribution function (CDF)**. As the variable is discrete, the CDF is discontinuous at the possible variable values and constant in between.

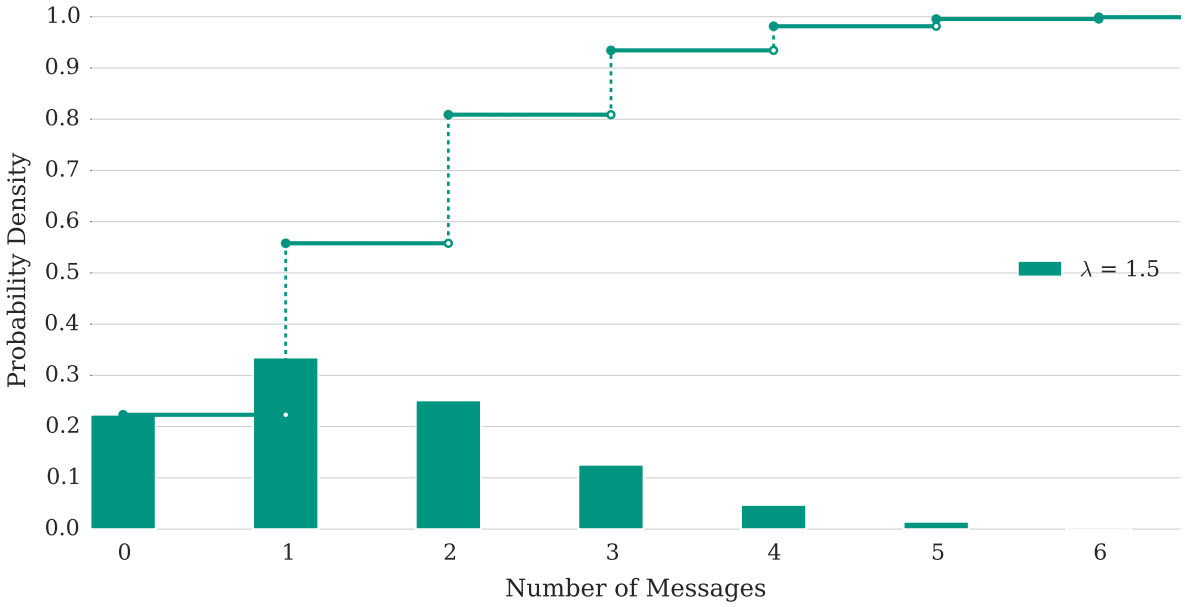

Figure 3.18: A typical Poisson distribution of one cell for a specific one minute time interval with $\lambda = 1.5$.

Due to its limitation of one parameter, the Poisson distribution is obviously not very flexible. In order to test if the data really follows a Poisson distribution, a Pearson's χ^2 goodness-of-fit test is conducted with a significance level of 1% for the historic data in each cell and each time

[11] To avoid skewing the message frequency by so-called tweet bots (cf. Chu et al., 2010), only one message per user per time interval is considered.

[12] The Maximum-Likelihood estimate for the parameter λ of the Poisson distribution is given by the arithmetic mean.

interval. This is the standard way of testing if a data sample comes from a specified *discrete* probability distribution. Approximately 5% of cases can be identified in the data, where the null hypothesis is rejected, i.e. where the data most likely does not follow a Poisson distribution.

The deviation from the expected Poisson distribution has two possible causes that are handled separately. The first is called *overdispersion* and refers to the situation when the sample variance exceeds the sample mean. Formally, this is defined as an index of dispersion (also variance-to-mean ratio) that is greater than 1. In a Poisson model, the index of dispersion is equal to one, or at least very close in real data.
However, there is another discrete probability distribution, namely the negative binomial distribution, which has two parameters and a variance greater than its mean. This distribution can be interpreted as a continuous mixture of Poisson distributions, also called a compound probability distribution. The mixing distribution of the Poisson parameter λ follows a gamma distribution. Thus, the negative binomial distribution is also known as gamma-Poisson mixture distribution and given as

$$P_{r,p} = \frac{\Gamma(r+k)}{k!\,\Gamma(r)} p^k\,(1-p)^r \tag{3.7}$$

For r approaching infinity, the negative binomial distribution converges to the Poisson distribution. The parameters $r \in \mathbb{R}^+$ and $p \in [0,1]$ do not have a straightforward physical interpretation anymore, but they can still be used in a model of the number of messages in a cell during a time interval. In Figure 3.19 a clearly elongated tail of the distribution, and a higher amount of zeros can be observed, which both depict a larger variance. The parameters of the negative binomial distribution in the graph approximately correspond to a sample mean of 1.8 and a sample variance of 2.9.

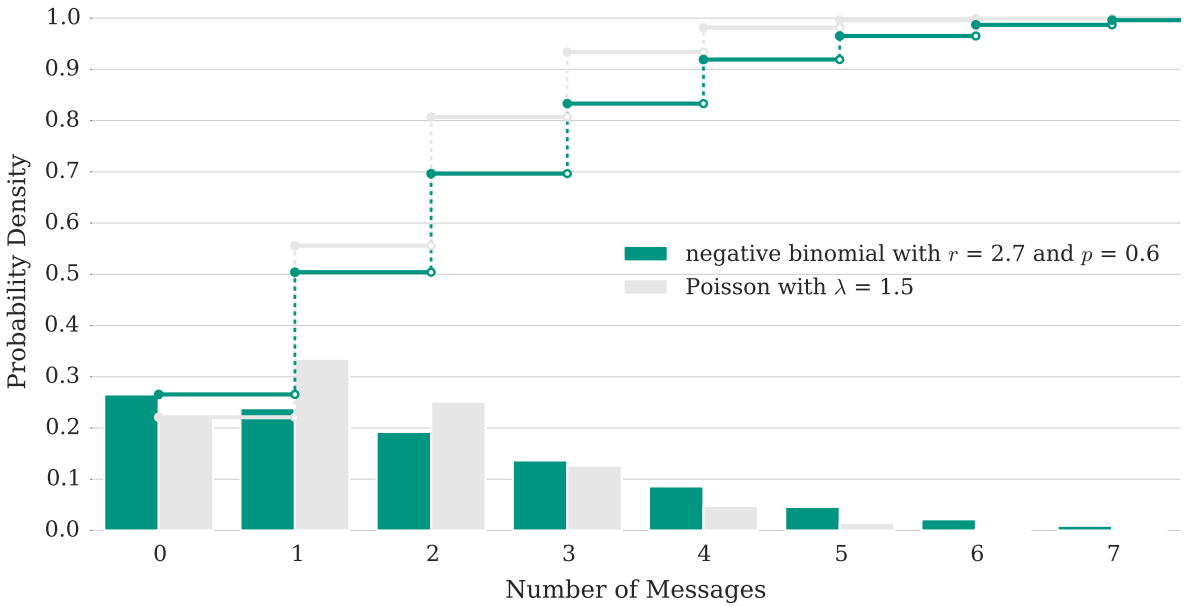

Figure 3.19: A negative binomial distribution (green) of a typical cell for a specific one minute time interval exhibiting overdispersion, in comparison with a typical Poisson distribution (gray) with $\lambda = 1.5$.

The second reason for a rejection of the Poisson assumption is called *zero-inflation*. The term refers to the excessive occurrence of zero-valued observations compared to the expected

amount according to a Poisson distribution. In these cases the approach relies on the so-called **zero-inflated Poisson (ZIP)** model.
In the ZIP model it is assumed that the process generating the messages has two states (cf. D. B. Hall, 2000), a state from which only zero values are generated and a Poisson state from which all other values are generated (possibly also zero). Which state the model takes is determined by the result of a Bernoulli trial. An illustrative interpretation of the generation of observations from a ZIP model is given by Rochford (2015):

> A weighted coin with a probability of π of yielding heads is flipped repeatedly. In case the result is head, the observation is zero, and in case of tails, the observation is generated from a Poisson distribution with parameter λ. Hence, there are two ways such a model can produce a zero observation. Either the coin shows heads or the coin shows tails and the Poisson process generates a zero.

Formally, the probability mass function of a ZIP model is given as

$$\begin{aligned} P(X = 0) &= \pi + (1 - \pi)e^{-\lambda} \\ P(X = k) &= (1 - \pi)e^{-\lambda}\frac{\lambda^k}{k!} \quad \text{for} \quad k > 0 \end{aligned} \tag{3.8}$$

Figure 3.20 depicts again a Poisson distribution with $\lambda = 1.5$ and a ZIP cell with $\lambda = 1.5$ and $\pi = 0.5$. The excess zero observations are clearly visible and decrease the probabilities of the other observations generated by the Poisson process. The reason for these excess zero observations is hard to determine in the case of social media messages. It can be due to local network failures that are not equally likely across the world.

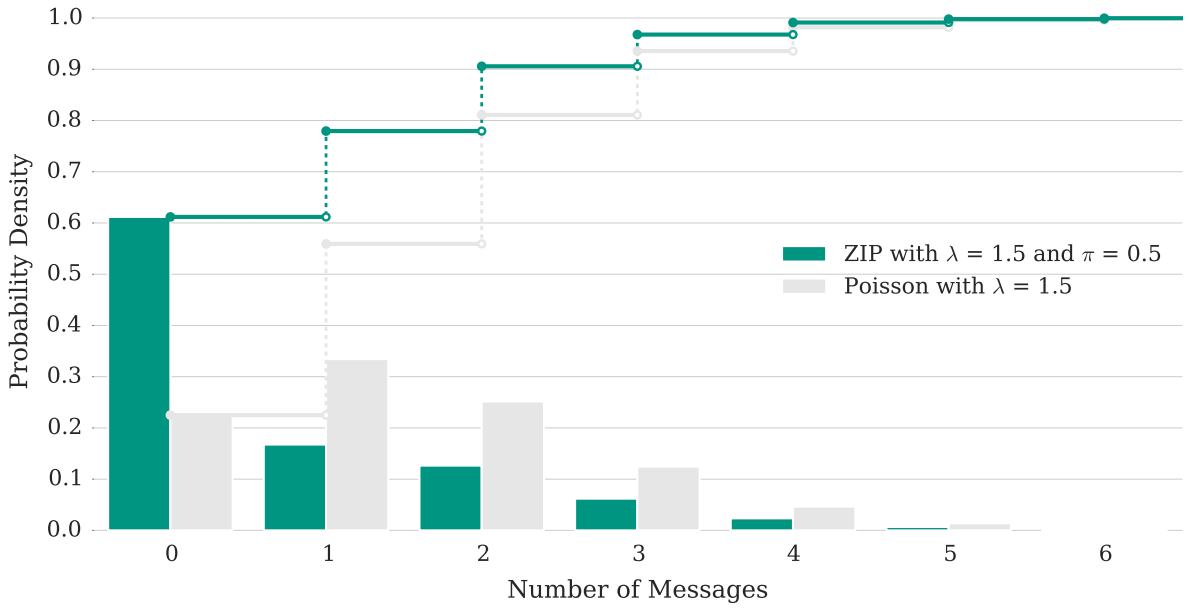

Figure 3.20: A ZIP distribution (green) of one cell for a specific one minute time interval exhibiting excessive zeros, in comparison with a typical Poisson distribution (gray) with $\lambda = 1.5$.

In all of the three count data models the CDF is exploited to derive the thresholds. For the Poisson and the negative binomial distribution, calculating the CDF is straightforward. Let $f(x)$ denote the probability mass function of a discrete random variable X, then

$$F(x) = \sum_{x_i \leq x} f(x_i) \tag{3.9}$$

is the CDF of X. However, in case that X follows a ZIP model, Equation (3.9) needs to be extended.

$$F_{ZIP}(x) = \pi + F_{P_\lambda}(x) \cdot (1 - \pi) \tag{3.10}$$

with $F_{P_\lambda}(x)$ being the CDF of the respective Poisson part of the ZIP model.

Dynamic Baseline Data Acquisition The following explanations are always meant to be applied to each cell for each time interval.
Before the initiation of the actual real-time analysis, the thresholds of the cells are derived for each time interval, as well as separated in *weekday* and *weekend* instances. In order to build a robust and reliable, initial statistical model to predict the number of messages, 3 months of baseline data is collected – that means approximate sample sizes of 60 observations for *weekdays* and 24 observations for *weekends*. After this initial phase, the framework continues to rely on a historic baseline of a maximum length of 3 months, and continuously collects new baseline data. No older data is used as it increasingly skews the data due to the described seasonal effects (see Section 3.2.2). Moreover, the models and thus their applied thresholds are automatically updated during runtime on a weekly basis. Consequently, the framework never applies "outdated" baseline data to predict current message volume.

Identification of Potential Event Cells In order to set a plausible threshold to detect significant increases, a value of $p_1 = 0.95$ and $p_2 = 0.99$ of the CDF of the fitted count model are heuristically chosen. These thresholds are inspired by the 2σ-bound and 3σ-bound, often referred to as *significant* and *highly significant*, respectively. For the count models, the thresholds depict the amount of messages that have a probability of only 5% and 1% to be exceeded.
For the framework, these heuristic thresholds are a good compromise between sensitivity and robustness – i.e. they are not too sensitive and thus yielding too many potential event cells for the next step, but they are sensitive enough to also not miss interesting events.
In case of natural disasters, the recall rate is of course more important than the precision, i.e. no important event should be missed and therefore a certain amount of false positive alarms is accepted. Moreover, as the detection of increased message volume is just the first step in a process chain to an actual alarm, a rather sensitive threshold is set as long as the defined real-time constraints are not violated. The thresholds are applied at the end of each time step ts. They are derived from the models whenever the current time step is synchronous with a time interval t available in the baseline data. If the current time step is in between two baseline time intervals, the threshold is derived by applying a weighted average. Let an example clarify the process:
The baseline data is sampled in disjoint time intervals of length $\Delta t = 1\,\mathrm{min}$. In order to evaluate the data at time steps of size $\Delta ts = 10\,\mathrm{s}$, six time steps per time interval are needed.

The thresholds available for the exemplary time intervals t_1 from 16:50:00 to 16:51:00 and t_2 from 16:51:00 to 16:52:00, are denoted as th_1 and th_2. When analyzing t_1 at 16:51:00, th_1 can be directly retrieved from the stored models. However, at 16:51:10 the system needs to analyze the time interval $t_{1,10}$ from 16:50:10 to 16:51:10. In order to apply plausible thresholds to the cells, a weighted average is used to calculate $th_{1,10}$.

$$th_{1,10} = \frac{5}{6} \cdot th_1 + \frac{1}{6} \cdot th_2 \tag{3.11}$$

Finally, at the end of each time step, the cells exceeding their respective thresholds are identified and denoted as potential event cells. This set of cells builds the basis for the thematic classification described in Section 3.4.

3.3.4 Alternative Spatial Discretization Methods

In a purely scientific view on the spatial discretization, there are obviously more complex and advanced methods than a regular grid. However, with one main goal of this work being the global, real-time applicability of the event analysis, the number of possible approaches decreases. Nonetheless, I want to very briefly discuss three alternatives.

Hierarchical Triangular Mesh The first approach is the one used by Krumm et al. (2015), a HTM. This is a two-dimensional tessellation of the surface of the earth in equilateral triangles of nearly equal size. Several discrete levels of resolution are applicable. A new level arises by inscribing a reversed triangle within an existing one, i.e. by dividing each triangle into four smaller triangles. Krumm et al. (2015) do not use the HTM in different resolution in different areas but only investigate four different resolutions as discrete nets for local event detection. Compared to the grid the triangles are less distorted near the poles, as they are not aligned with the meridians and parallels of the earth. Hence, they do neither suffer from anisotropy nor create singularities when subdividing the sphere. If this has any significant impact on the detection capabilities in the relevant latitude range (i.e. from $-40°$ to $60°$) is doubtful.
The implementation and resulting management of the geographic data stored in this way is quite complex and introduces even more complexities for following analysis steps. The HTM has not been applicable in a global real-time approach so far as Krumm et al. (2015) state themselves. Moreover, solving the edge problem (cf. Section 3.3.1) and the correct handling of time zones would also raise more non-trivial issues.

Quadtree Another possible method for discretizing space is the well known quadtree method of Finkel et al. (1974). Basically, it is also a hierarchical mesh just like the HTM but using rectangles (also sometimes squares) instead of triangles[13]. I implemented and tested a quadtree algorithm that is aligned with the meridians and parallels of the earth and makes use of the inherent hierarchy of the quadtree idea – i.e. according to a certain threshold the quadtree has a higher granularity in areas that yield more messages. The threshold describes the maximum

[13] Sometimes the HTM is categorized into a broad notion of quadtrees according to Samet (1984).

number of messages in one rectangle before it is split up into four smaller ones along its two centerlines in longitudinal and latitudinal direction.
After defining a baseline tree for a certain time of day, a current quadtree can be compared to the baseline tree. Consequentially, potential event areas can be detected by extracting rectangles where the current tree has a higher resolution than the baseline tree. Tests show that this comparison is actually possible within approximately $0.3\,\mathrm{sec}$ for a global, one minute quadtree of TWITTER messages. Thus, the quadtree approach would be feasible concerning the goals of this work. Figure 3.21 depicts a one minute quadtree from Thursday, July 9th 2015 at 2:35 UTC.

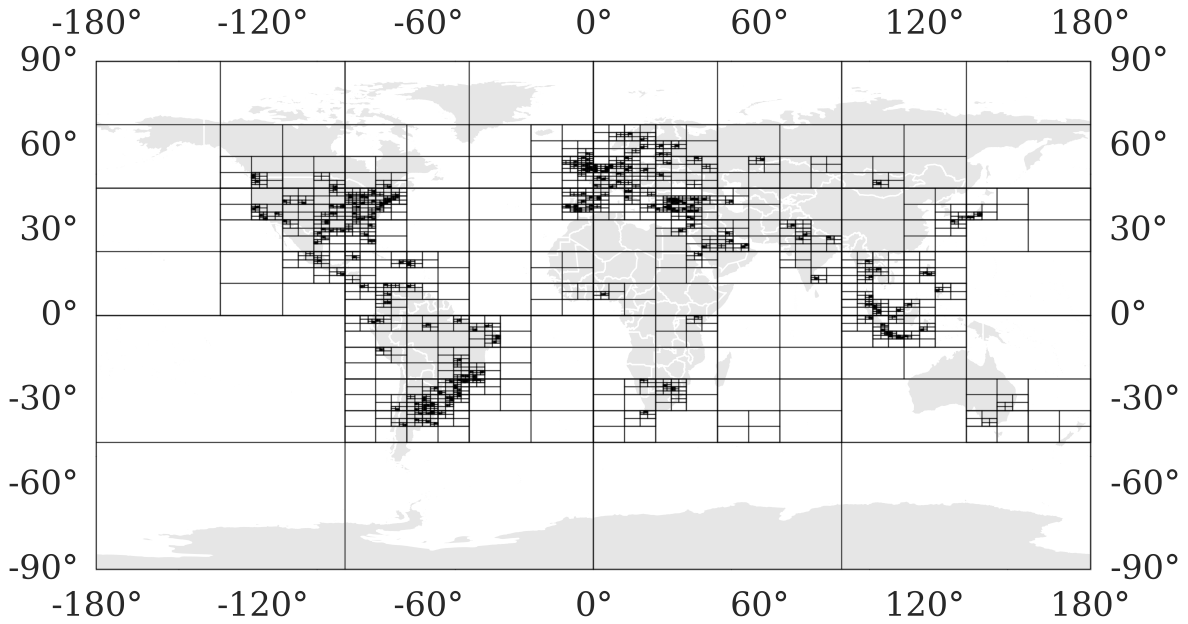

Figure 3.21: The one minute global quadtree of georeferenced tweets from Thursday, July 9th 2015 at 2:35 UTC down to a maximum of one tweet per quadrant.

However, there are some drawbacks as well. First, the storage size of a global quadtree model representing one minute amounts to approximately 2MB compared to the grid with only 20 to 25KB. Secondly, the edge problem is not as easily resolvable as with a grid. The adjacent rectangles can be distributed over several depth levels of the tree. Third, the incorporation of time zones is far from trivial.
Last and most important is the question of defining a sensible baseline tree for a time interval from historic data. The chosen approach for the quick feasibility test is straightforward. All messages are aggregated from the n historic instances of the respective time interval and build one quadtree with a maximum threshold of n messages per rectangle. The current quadtree is then assembled by using a maximum threshold of one message per rectangle. Thus, it is not a mathematically strict but highly plausible basis for the comparison. However, there is also no obvious correct way of quantifying the increase and thus evaluating its significance.

Congruent cells The last and very interesting approach is an idea presented in the broad field of GNSS and specifically in the handling of multipath effects as main error source in static and kinematic GNSS measurements. In order to optimize the consideration of site-specific conditions when generating so-called multipath stacking maps, Fuhrmann et al. (2014)

introduce the idea of *congruent cells*.
The cells have a constant increment in latitudinal direction but variable longitudinal resolution, i.e. there is a variable number of cells per latitudinal increment. The goal is to generate cells with very similar shape and size to possibly capture more accurate area specific characteristics over nearly the whole latitude range.
In their formulae, Fuhrmann et al. (2014) assume a spherical approximation of the earth's shape. The direct extension for ellipsoidal models – e.g. the WGS84 ellipsoid on which the coordinates are based – is non-trivial because a constant increment in latitudinal direction in degrees on an ellipsoid does not generate constant arc lengths. But even when applying the spherical model, the maximum relative area deviation with respect to the ellipsoidal surface can be reduced to only 1.25% in the relevant latitude range from $-40°$ to $60°$. Moreover, the cells are not systematically elongated in longitudinal direction when approaching the poles.
Nonetheless, this approach also leads to a much more complex setup for data storage structures. This has negative implications for the efficient assignment of messages to cells as well as the retrieval of messages from specific cells. Additionally, the handling of the edge problem poses another non-trivial problem compared to the grid approach. However, in contrast to the HTM and the quadtree, the congruent cells approach could be made aware of time zones, although not as straightforward as the grid.

Synopsis Overall, it is questionable if approaches that in parts strongly increase complexity at various levels of the analysis, are worth the effort that needs to be put into making them globally and real-time applicable. Especially, when a much simpler approach is just as effective for the event detection, and even largely outperforms the more complex ones concerning the storage and retrieval efficiency, the speed of implementation, and the maintenance workload.

3.4 Thematic Classification

At this point of the event analysis, possible event locations can be narrowed down to one or several cells in the global grid for the previous time interval, based on statistical models and respective appropriate thresholds for message volumes. Consequentially, the next step is to derive the cause of the increased message volume, i.e. to determine the topic of the messages from the respective cells. Therefore, the textual content of the messages that generated the statistical detection is analyzed. However, instead of assigning an arbitrary topic to each of the cells, the similarity to a predefined set of topics from a domain of interest is calculated – i.e. only with domain relevance provided, a concrete class label is derived.
In order to explain the thematic classification approach, I will rely on examples of the use-case scenario of natural disasters throughout this section.

3.4.1 Modeling the Domain

For the abstraction of the domain of interest a hierarchical tree structure is employed for the different levels of sub-domains – a *domain taxonomy*. Thus, the desired topical granularity of

the classification is controlled.
In order to keep the approach as generic as possible, the exact contents and characteristic of the taxonomy are decoupled from the rest of the framework. So a domain change is feasible in a plug-and-play fashion without any other changes to the overall system[14].

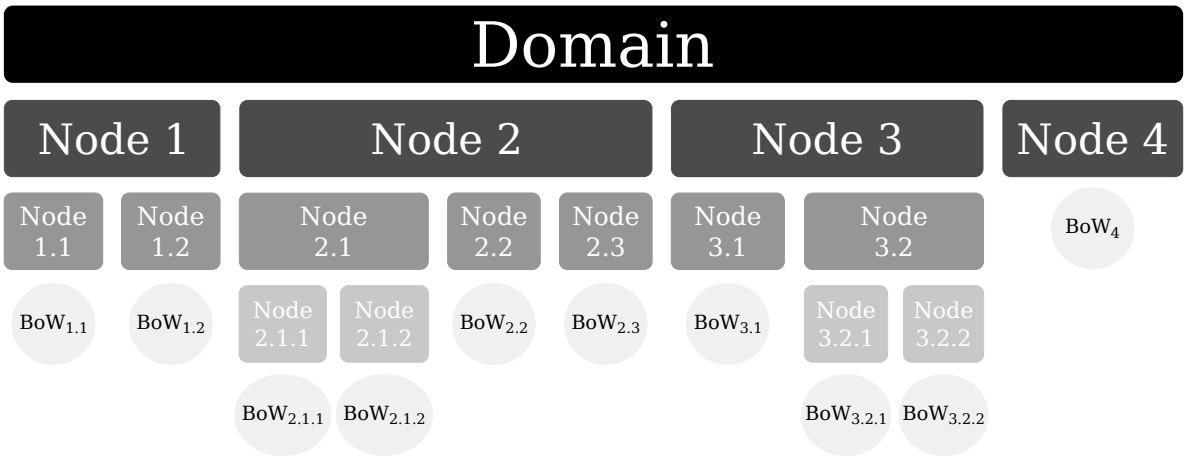

Figure 3.22: Example for a general domain taxonomy with BoW for each leaf node.

An abstract depiction of a potential taxonomy is shown in Figure 3.22. The nodes in this structure represent the sub-domains of the respective granularity level and the root node represents the complete domain. The different branches in the taxonomy do not need to have the same depth, which means that the topical granularity of the sub-domains can vary. Especially considering the used input data, it might not be possible to discriminate between some sub-domains. The main reason is the unspecific reporting style in social media data, i.e. the users might not distinguish between formally distinct sub-domains, either because of indifference or due to ignorance.
Each leaf of the taxonomy contains a set of distinct terms that is virtually unambiguous for the specific sub-domain. These terms are initially based on common sense heuristics and expert knowledge and represent the event as BoW model. The BoW for a non-leaf node is derived during runtime through the union of the BoWs from the respective leaves. For example, the BoW for node 2, i.e. BoW_2 is given by

$$BoW_2 = BoW_{2.1} \cup BoW_{2.2} \cup BoW_{2.3} \tag{3.12}$$

with

$$BoW_{2.1} = BoW_{2.1.1} \cup BoW_{2.1.2} \tag{3.13}$$

For the sake of clarity, I will go through the process of the thematic classification using the natural disaster taxonomy presented in Figure 3.23, which is a simplified and adapted version of Table 2.1 based on Below et al. (2009).

The initial distinct terms for the leaves are mainly the different event names and event-related terms that are otherwise rather rare in natural language. When several events of a certain type could be detected, a careful, manual extension of the distinct terms might be applied. In case of

[14] Except for the domain dictionary described in Section 3.4.2 for a multi-lingual coverage.

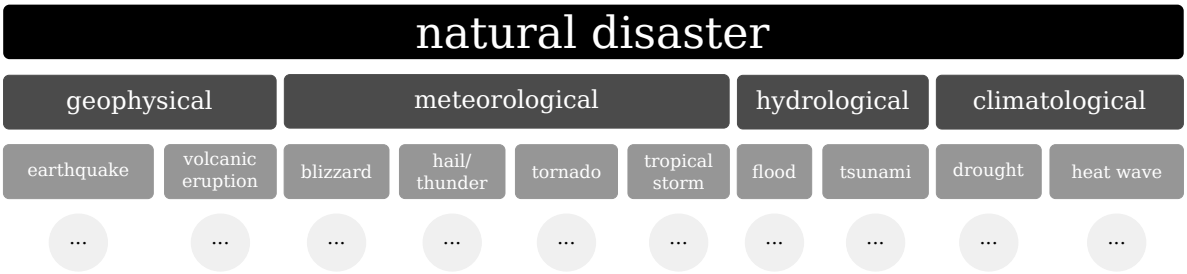

Figure 3.23: The employed taxonomy for the domain of natural disasters. The BoWs are excluded from this depiction.

the disaster type *earthquake*, for example, the terms *richter scale* and *epicenter* were manually added after some event experience, whereas the term *shaking* was removed as it proofed to be too frequent in general usage. Only terms with a high discriminative power should be kept.
In the prototype the taxonomy is loaded as a standard JSON file which can reflect the tree structure. Listing 3.3 shows the natural disaster taxonomy in JSON-format including the terms in their respective BoWs.

Analogous to the procedure in Equations (3.12) and (3.13), the BoW for the node *hydrological* can be derived by uniting the BoWs from the nodes *flood* and *tsunami*, and consequently contains the terms:

flood, *flooding*, *inundation* and *tsunami*.

3.4.2 Additional Pre-Processing

In Section 2.3.3, it was explained why certain pre-processing steps are necessary to the effective adoption of models that are based on the number of term occurrences. Hence, the steps are tokenization[15], stop word removal[16] (and exclusion of pure digits, i.e. numbers, and non-alphanumeric characters), as well as linguistic normalization of surface word forms (i.e. *lemmatization*[17]) when possible.
These steps, as well as the ones explained in the following, are mainly necessary to reduce the size of the vocabulary space and to yield more reliable similarity values between the documents and the BoW models for events.

Apart from these typical methods, the special characteristics of social media data has to be accounted for. These texts commonly suffer from spelling mistakes, (uncommon) abbreviations and acronyms, colloquial terms and lexical variants, mixed language use, etc. There are two ways of handling these issues according to Eisenstein (2013): *normalization*[18] (also *cleansing*) and *domain adaption*. Whereas the former is described as adapting text to fit the tools, the

[15] A custom tokenizer based on regular expressions is used, which basically removes all non-word characters and then splits the resulting string on word boundaries.
[16] Custom lists were compiled, inspired by the lists of Doyle (2016), MySQL (2016), and PostgreSQL (2008).
[17] *morphy*, a tool for morphological transformations integrated in the WordNet lexical database for English (Miller, 1995) is used; available at `https://wordnet.princeton.edu/`.
[18] This is not the same as *linguistic normalization* in Section 2.3.3.

```
{name:"natural disaster",
    geophysical:{name:"geophysical",
        earthquake:{name:"earthquake",
            terms:["earthquake","quake","aftershock",
            "epicenter","foreshock","mainshock","quake",
            "richter scale","seismic","seismicity",
            "tremor"]},
        volcanic_eruption:{name:"volcanic eruption",
            terms:["volcano","volcanic","caldera","crater",
            "eruption","eruptive","lava","magma",
            "stormvolcano"]}},
    meteorological:{name:"meteorological",
        blizzard:{name:"blizzard",
            terms:["blizzard","snowstorm","winterstorm"]},
        hail_thunder:{name:"hail/thunder",
            terms:["hail","lightning","hailstorm",
            "thunderstorm"]},
        tornado:{name:"tornado",
            terms:["tornado","whirlwind"]},
        tropical_storm:{name:"tropical storm",
            terms:["cyclone","typhoon","hurricane"]}},
    hydrological:{name:"hydrological",
        flood:{name:"flood",
            terms:["flood","flooding","inundation"]},
        tsunami:{name:"tsunami",
            terms:["tsunami"]}},
    climatological:{name:"climatological",
        drought:{name:"drought",
            terms:["drought"]},
        heat_wave:{name:"heat wave",
            terms:["heat wave"]}}}
```

Listing 3.3: Natural disaster taxonomy represented as JSON file to reflect the tree structure for a fast traversal of the different granularity levels.

latter can be seen as adapting the tools to fit the text. Baldwin, Cook, et al. (2013) showed that, although social media is indeed noisy, it is possible to cleanse it using existing NLP tools. Moreover, cleansing often is less complex than domain adaption but still equally efficient.
The major step is the detection and correction of possible spelling mistakes (also called *lexical normalization*[19]). Frequent mistakes include

- *expressive lengthening*, e.g. “touchdoooooown” for “touchdown”
  – often indicating subjectivity and sentiment according to Brody et al. (2011)
- *vowel dropping*, e.g. “tlking” for “talking”
  – to save time on standard keyboard writing (cf. Eisenstein, 2013) and adopted for mobile devices

[19] This, again, is not the same as *linguistic normalization* in Section 2.3.3.

- *miscellaneous*, e.g. “ggame”, “hame” or “gam” for “game”
  – usually unintentionally

These spelling mistakes can be reduced using an implementation based on word lists and dictionaries, and the *Damerau-Levenshtein distance*. The *Damerau-Levenshtein distance* (cf. Damerau (1964) and Levenshtein (1966)), is a metric (also called edit distance) to quantify the dissimilarity of two sequences, here strings. This is achieved by counting the minimum number of operations required to convert one sequence into the other. The operations defined for that distance are insertion, deletion, substitution, and the transposition of two adjacent characters. Figure 3.24 depicts an example of a spelling mistake that needs two steps to be corrected, i.e. the original and the correct term have an edit distance of two. First, the letters *u* and *o* need to change their respective position with each other and then the last *o* needs to be deleted.

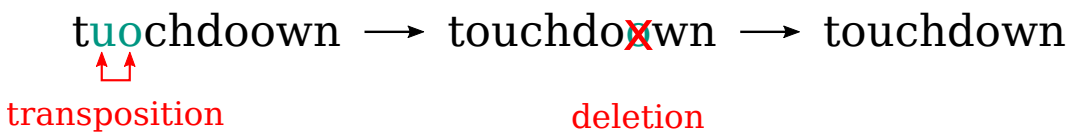

Figure 3.24: Example for a spelling correction with an edit distance (Damerau-Levenshtein) of two – one transposition of two adjacent characters and one deletion.

For spelling correction, a given word is first checked for existence in the dictionaries and only if it does not exist, corrections are applied. The dictionary consists of extensive word lists mainly from WordNet, GNU Aspell[20] and a collection of English Wikipedia[21] article titles. The approach also accounts for modern slang terms[22] and typical social media abbreviations[23]. However, no transformation of so-called *phrasal abbreviations* such as *lol* for *laugh out loud* or *shmily* for *see how much I love you* into their full form is conducted, as they usually do not contain important keywords.
In order to limit the results of the possible correct versions of a term, a maximum edit distance of two is set. Nonetheless, several possibilities can remain. This is the reason why the results are ranked based on an algorithm of Ratcliff et al. (1988). The idea is to retrieve the longest contiguous matching subsequence of two strings. Then the same procedure is applied recursively to the parts of the strings on the right side and on the left side of the matching substring. The final score is calculated as the ratio of twice the number of characters in common, and the total number of characters of the two strings. This algorithm tries to emphasize the similarity of the overall *Gestalt*[24] of the two strings (see Ratcliff et al., 1988).
Expressive lengthening is partially cleansed (for English messages) before spelling correction is applied. The occurrence of three or more consecutive equal characters is reduced to only two of them, i.e. *touchdooooown* will be mapped to *touchdoown*. Thus, the necessary edit distance to the true form of the term decreases. Otherwise, terms with expressive lengthening would be excluded or at least ranked very low by the standard procedure possibly leading to false corrections.

[20] Available at http://www.aspell.net.
[21] Dumps of all English Wikipedia article titles are available at https://dumps.wikimedia.org/enwiki/latest/.
[22] Available at http://onlineslangdictionary.com/word-list/0-a/.
[23] Partly derived from http://www.netlingo.com/ and semi-automatically extended.
[24] The approach refers to the *Gestalt* laws of grouping originally introduced by Wertheimer (1923).

The following minor cleansing steps are also conducted mostly by using regular expressions:

- replace specific TWITTER glossary such as @ and #,
  @David is at #StonesConcert ⇒ *David is at StonesConcert*
- remove any form of URL and **H**yper**T**ext **M**arkup **L**anguage (HTML) code,
- split camel caps expressions and terms concatenated by underscores,
  BigData ⇒ *Big Data*
  Big_Data ⇒ *Big Data*
- remove multiple consecutive whitespace characters,
- replace multiple consecutive punctuation,
  !!!?!?! ⇒ *!*
- transform words with all upper letters to capitalized form,
  IT'S JUSTIN BIEBER ⇒ *It's Justin Bieber*
- append missing sentence ending punctuation.

Multilingual Aspects When analyzing georeferenced tweets on a global scale, various languages that can occur in social media messages have to be considered. As Leetaru et al. (2013) showed in their extensive study, more than 40% of georeferenced tweets are in English and 88.82% are written in one of the languages in Figure 3.25 (including English) – hence, a main focus on the English language is justified. Nonetheless, I try to account to some extent for the most common languages. Therefore, all terms in the natural disaster taxonomy are compiled in the 64 languages listed in Table 3.2. The list contains the most common languages used in georeferenced tweets as well as other languages with a large number of native speakers[25].

Table 3.2: The 64 languages represented in the disaster dictionary of the translation engine.

Albanian	French	Khmer	Serbian
Arabic	Georgian	Korean	Slovak
Armenian	German	Laotian	Slovenian
Azerbaijanian	Greek	Macedonian	Spanish
Belarusian	Gujarati	Malaysian	Sundanese
Bengal	Hausa	Maltese	Swedish
Bosnian	Hindi	Marathi	Tagalog
Bulgarian	Hungarian	Nepali	Tamil
Cebuano	Icelandic	Norwegian	Telugu
Chinese	Indonesian	Pashto	Thai
Croatian	Irish	Persian	Tok pisin
Czech	Italian	Polish	Turkish
Danish	Japanese	Portuguese	Ukrainian
Dutch	Javanese	Punjabi	Urdu
English	Kannada	Romanian	Vietnamese
Finnish	Kazakh	Russian	Zulu

[25] Estimates of the number of native speakers of different languages are available from http://www.ethnologue.com/statistics/size

The study of Leetaru et al. (2013) also yielded that another 8.39% of georeferenced tweets can not be clearly assigned to one language (referred to as *Other* in Figure 3.25).
The employed tokenization approach based on regular expressions is language agnostic for segmented languages. Hence, even if a message shows a mixed language usage or the language cannot be identified, it can still be tokenized (in case it is a segmented language) and relevant keywords can be found in the 64 languages. In order to be able to retrieve meaningful tokens from unsegmented languages such as Japanese[26], Korean[27], Thai[28] and Chinese[29], specialized tokenizers are employed.

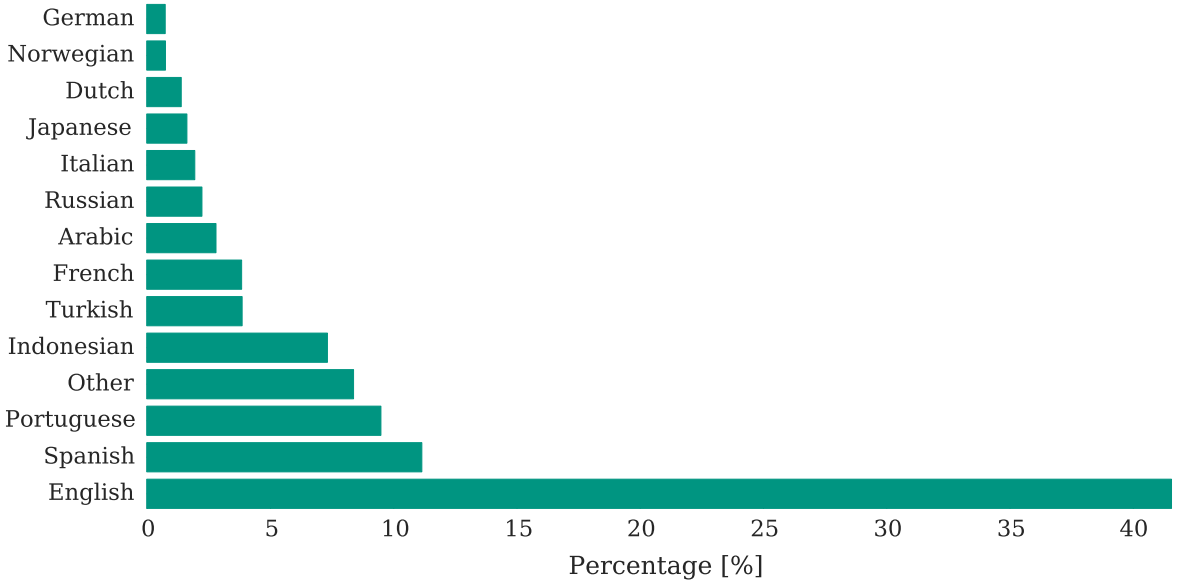

Figure 3.25: Language frequencies in georeferenced tweets according to Leetaru et al. (2013).

Concerning stop word removal, respective lists for all languages in Table 3.2 are applied. These lists are available in the aforementioned sources for 40 of the 64 languages. For the rest, custom short lists of articles (*the*, *a(n)*), prepositions, pronouns (personal, possessive, etc.) and conjunctions (e.g. *and*, *or*, *but*) are compiled[30].
The spelling correction approach works for all languages with available dictionaries from GNU ASPELL[31]. Moreover, the check for existence is of course optimized for English, as the slang dictionary, the WIKIPEDIA titles and the WORDNET database are in English. So, spelling correction is employed only if the language of the message can be identified. TWITTER has an in-built language identification system on its servers, i.e. each tweet incorporates a field

[26] A PYTHON port by Masato Hagiwara (available at https://pypi.python.org/pypi/tinysegmenter) of TINYSEGMENTER, a compact Japanese tokenizer originally written in JAVASCRIPT by Taku Kudo is used (available at http://chasen.org/~taku/software/TinySegmenter/).
[27] TWKOREAN, a PYTHON wrapper from Jaepil Jeong (available at https://github.com/jaepil/twkorean/) is used for the SCALA/JAVA library TWITTER-KOREAN-TEXT from TWITTER (available at https://github.com/twitter/twitter-korean-text).
[28] PYTHAI, a PYTHON library from Herman Schaaf (available at https://github.com/hermanschaaf/pythai) based on the C library LIBTHAI is used (available at http://linux.thai.net/projects/libthai/).
[29] PYMMSEG, a PYTHON interface from Chiyuan Zhang (available at https://github.com/pluskid/pymmseg-cpp) to the RUBY utility RMMSEG is used (available at https://github.com/pluskid/rmmseg-cpp).
[30] Different online translation engines were used: http://www.worldlingo.com/en/products_services/worldlingo_translator.html, https://translate.google.com/ and http://freetranslation.paralink.com/.
[31] That excludes Albanian, Bosnian, Cebuano, Chinese, Georgian, Hausa, Japanese, Javanese, Kazakh, Khmer, Korean, Lao, Nepali, Sundanese, Thai and Urdu from Table 3.2.

specifying its automatically derived language. However, this is a black-box system and the TWITTER API terms of service forbid benchmarking[32]. Nonetheless, Lui et al. (2014) report that the proprietary solution does not outperform the best off-the-shelf tools, and even lacks a wide language coverage. In order to overcome this issue, the off-the-shelf language identification tool LANGID.PY of Lui et al. (2012)[33] is used, which supports 97 languages including all but three of the 64 languages in Table 3.2 (Hausa, Cebuano, Sundanese). Thus, the system is also guaranteed to stay independent from one specific social media source.

Beside the *cleansing* measures to prepare the multilingual input for the tf-idf weighting, a simple domain-restricted translation engine is applied. The translation engine detects all terms in the BoW for the complete taxonomy in all languages listed in Table 3.2 and maps them onto the English equivalent. In case a term can have several English translations, the count is equally split across the English terms – e.g. the Arabic term زلزال can be translated as *quake* or *earthquake*, thus both terms would get 0.5 added to their respective counters.
The translation in one system language (here English) allows a consistent classification output for events and in addition, it accounts for messages in different languages from the same cell. Oftentimes, the messages from one cell are in (one of) the main languages of the specific country, with some occurrences of English messages. In cells close to country borders as well as in countries with several frequently used languages, these cases are quite common. Figure 3.26 depicts an example from a cell in India where messages in the three Indian languages Urdu, Telugu and Hindi occurred together with English messages. Due to the translation engine, an inter-lingual aggregation can be performed for the equivalent terms – ژرگ (Urdu), ఉరుము (Telugu) and आंधी (Hindi) – in English and eventually the event can be classified as *hail/thunder*.

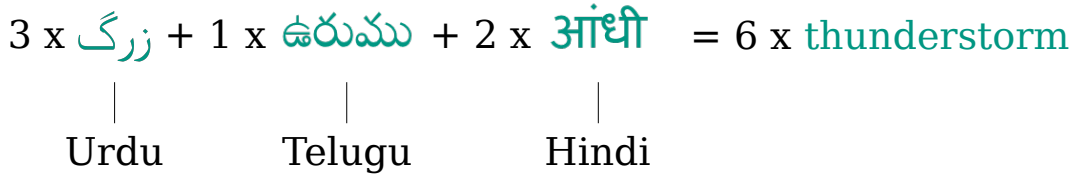

Figure 3.26: Example of inter-lingual aggregation of equivalent terms (here *thunderstorm*) in three common languages in India: Urdu, Telugu and Hindi.

The optimal way would be a complete translation of all messages into one single system language upon arrival in the system. However, this is not a feasible approach with the currently available translation APIs or software tools. Besides the constraints concerning the usual rate limits of these services, the translation capabilities generally neither achieve the desirable quality nor the temporal efficiency for the large range of different languages.

[32] "Be a Good Partner to Twitter", Part I. Section 6. Paragraph e. Phrase iv. of the TWITTER Developer Policy available at https://dev.twitter.com/overview/terms/agreement-and-policy.
[33] An up-to-date version is available from https://github.com/saffsd/langid.py.

3.4.3 Classification Process

The initial situation of the thematic classification is a certain set of cells that exhibited a significantly increased message volume for the preceding time interval. The following classification process is then conducted separately for each of these potential event cells. The methods employed to derive the topic are based on the ideas of IR introduced in Section 2.3. Although one final class label is assigned for each cell, the idea of a fuzzy class membership is adopted based on the similarity scores of the current documents with the sub-domains in the taxonomy.

Document Aggregation As described in Section 2.3.1, an important aspect for VSM is an appropriate definition of a document unit. Concerning social media data as input, the single messages are quite short compared to the usual length of documents in IR. In case of TWITTER, for example, the encountered messages have an average length of 10 ± 6.4 words per message[34]. By defining one message as one document, the topic will be scattered and the weighting based on *tf* will be unstable.
That is why I decided to introduce the idea of aggregating several messages into one document. Naturally, this aggregation can not be random but has to be guided by suitable constraints. Fortunately, the framework already provides a message partitioning system that fits this purpose – that is the messages from one cell during one time interval. When adapting the idea of Tobler's first law of geography[35], the assumption in this case is that messages, which are sent from the same area and during the same (short) time interval, are very likely to be topically related, especially when a significant event has occurred.

Dynamic Document Collections In order to be able to apply the tf-idf weighting scheme detailed in Section 2.3.2, a collection of documents is needed instead of a single document. This collection should represent a sort of up-to-date baseline of the typical topics discussed in the respective cell. I compile a dynamic document collection *on-the-fly* accordingly, instead of using pre-built collections. Due to the topically variable nature of TWITTER, only historic data of up to one hour for each cell in the grid is used. So each time a potential event cell is analyzed, the system can rely on a sufficient collection of documents for comparison. Given that the length of a time interval Δt is set to one minute, a maximum of 60 documents in total is available. Another advantage of not using pre-built models is that the framework stays more independent from the domain.
In cases where the cell already yielded a detection of an event relevant to the domain in the previous hour, the document representing that time interval is excluded from the collection. Otherwise, this document would skew the results as it does not represent a baseline usage of the terms in the vector space.

With that document collection at hand, the tf-idf weighting scheme can be applied to the documents and the BoWs representing the event types – i.e. both are represented in the

[34] Calculated over a total of 0.2B georeferenced tweets.
[35] "[e]verything is related to everything else, but near things are more related than distant things". (cf. Tobler, 1970).

same vector space and the cosine similarities can eventually be calculated as described in Section 2.3.4. Here the analogy to IR systems or search engines applies, as the documents can be interpreted as indexed websites and the event BoWs are similar to search queries provided for example by user input.
The documents in the collection are then ranked according to their similarity scores. The *current* document – i.e. the aggregated tweets from the statistical significant cell of the previous time interval – should yield a significantly higher score than the other documents representing the baseline and hence be ranked on top. Otherwise, it is not accepted as event-related. In practice, either all documents frequently exhibit similarities very close to zero in case the increase was not event-related, or the increase was in fact event-related, then often all but the current document exhibit values close to zero.

Classification Granularity Instead of only testing the different event BoWs stored in the leaves of the taxonomy, I additionally try to account for rather unspecific event descriptions in the messages. This is achieved via a stepwise classification process through the different levels of the taxonomy from top to bottom.
In this process, the different nodes of the taxonomy are represented by the union of the BoWs of all their leaves. In terms of the use case scenario, the first step is the calculation of the similarities as described above, between the documents and the BoW for the first level (the root node), i.e. *natural disaster*. If the relevance of the current document for the domain in general can be established based on the similarity ranking, the next deeper levels of the taxonomy are analyzed recursively. From there onwards, there are several nodes per level whose scores for the current document are compared if necessary. The comparison becomes necessary if the current document is ranked on top for more than one node at the current level. In case the nodes have equal scores for the current document or no node ranks the current document on top, the classification process terminates and yields the preceding node as event type.
Let me illustrate the procedure with an example. Let d_1 denote the current document and d_i with increasing index $i = 2, \ldots, 60$ the other documents in the collection. The term frequency analysis for the tf-idf weighting scheme of the VSM yields the following results for keywords contained in the disaster taxonomy.

d_1 6 occurrences of the term *tornado*
2 occurrences of the term *whirlwind*
1 occurrences of the term *lightning*
7 occurrences of the term *thunderstorm*
$\vdots$
d_9 1 occurrence of the term *thunderstorm*
$\vdots$
d_{17} 1 occurrence of the term *earthquake*
$\vdots$
d_{35} 1 occurrence of the term *tornado*
$\vdots$
d_{60} ...

A possible document ranking in the first step for the natural disaster taxonomy is then depicted in Table 3.3.

Table 3.3: Example of a document ranking according to the similarity with the BoW representing the first level *natural disaster* in the taxonomy for the domain of natural disasters

Rank	Document	Similarity
1	d_1	0.696
2	d_{17}	0.138
3	d_9	0.032
4	d_{35}	0.015
⋮	⋮	⋮
60	d_{48}	0.0

In the second step, the document rankings are determined for the nodes *earthquake*, *meteorological*, *hydrological* and *volcanic eruption*, with their respective BoWs. This could result in the rankings depicted in Table 3.4.

Table 3.4: Example of a document ranking according to the similarity with the BoWs representing the second level nodes in the taxonomy for the domain of natural disasters – i.e. (a) *geophysical*, (b) *meteorological*, (c) *hydrological* and (d) *climatological*

Rank	Document	Similarity
1	d_{17}	0.258
2	d_{35}	0.0
3	d_9	0.0
4	d_1	0.0
⋮	⋮	⋮
60	d_{48}	0.0

(a) node *geophysical*

Rank	Document	Similarity
1	d_1	0.824
2	d_9	0.038
3	d_{35}	0.018
4	d_{17}	0.0
⋮	⋮	⋮
60	d_{48}	0.0

(b) node *meteorological*

Rank	Document	Similarity
1	d_{35}	0.0
2	d_1	0.0
3	d_{17}	0.0
4	d_9	0.0
⋮	⋮	⋮
60	d_{48}	0.0

(c) node *hydrological*

Rank	Document	Similarity
1	d_1	0.0
2	d_{35}	0.0
3	d_{17}	0.0
4	d_9	0.0
⋮	⋮	⋮
60	d_{48}	0.0

(d) node *climatological*

The following step only tests the sub-nodes of node *meteorological* as this yielded the highest score of the nodes which ranked d_1 on top. In fact, it is the only node that ranks d_1 on top, because the nodes *hydrological* and *volcanic eruption* are not taken into account as all scores equal zero and thus their rankings are arbitrary.
The next nodes in consideration are *blizzard*, *tornado*, *drought/heat wave*, *tropical storm* and

hail/thunder. In the example, all nodes except for *tornado* and *hail/thunder* exclusively yield zero scores. Thus, Table 3.5 only depicts the two relevant nodes.

Table 3.5: Example of a document ranking according to the similarity with the BoWs representing the third level nodes in the taxonomy for the domain of natural disasters – i.e. (a) *tornado*, (b) *hail/thunder*

Rank	Document	Similarity
1	d_1	0.582
2	d_{35}	0.025
3	d_9	0.0
4	d_{17}	0.0
⋮	⋮	⋮
60	d_{48}	0.0

(a) node *tornado*

Rank	Document	Similarity
1	d_1	0.582
2	d_9	0.054
3	d_{35}	0.0
4	d_{17}	0.0
⋮	⋮	⋮
60	d_{48}	0.0

(b) node *hail/thunder*

At this point, the two nodes can obviously not be distinguished as they both rank d_1 on top and also with the same similarity score. In this case, the preceding node in the taxonomy hierarchy is set as label for this event – i.e *meteorological*. Hence, the system accounts for the unspecific content of the current document with equal similarity scores for two formally different events. Nonetheless, the respective scores are kept for the current document of the lower level for later usage.
In such cases, the users who disseminated the messages might either have interpreted the weather conditions differently or it actually was a combination of the two events.

3.5 Spatial-Thematic and Temporal Clustering

At this point of the event analysis, cells of the preceding time interval with significantly increased message volume have been identified and they have been tested if the topic of the respective messages can be assigned to the domain of interest in general or even to a specific subtype.
In order to enable this large-scale detection and classification of events in the first place, I introduced discretization in the framework for both space and time (see Section 3.3.1 and Section 3.3.2). By now, however, the amount of data has been reduced to a size that is much more processible and allows for steps that, in some sense, reverse the discretization – i.e. the events can now get clustered with respect to space and time.

3.5.1 Spatial-thematic Clustering

First, the process of spatially clustering event cells of the same time interval is introduced. Here, the event cells that are highly likely to be triggered by the same real-world event based on their spatial proximity and the class label assigned by the thematic classification should be identified. For this reason, the aim is a spatial-thematic clustering.

This step is necessary, as the resolution of the grid is chosen – in combination with additional factors – to avoid being too sensitive to very small-scale events (depending on the domain of interest). Nevertheless, there might be events of interest that exceed the extent of single cells and spread over larger areas. The framework should account for these cases by aggregating the respective cells into one spatial-thematic event cluster.

Image Segmentation As the space is represented as a rectangular grid, methods from digital image processing can be applied. Here the idea of image segmentation according to Haralick et al. (1992) is adopted, i.e. partitioning an image into a set of non-overlapping regions, whose union is the entire image, in order to decompose the image into parts that are meaningful with respect to a particular application. In this case, the meaningful parts are obviously the areas which are affected by the same event in the same time interval.

Region Growing The method applied for clustering cells is based on region-oriented or contextual segmentation algorithms which consider the local neighborhood of cells – commonly referred to as pixels in image processing. The general goal is to identify objects in images as these are usually represented as connected regions. The specific approach I adapt for this purpose is called *region growing*.

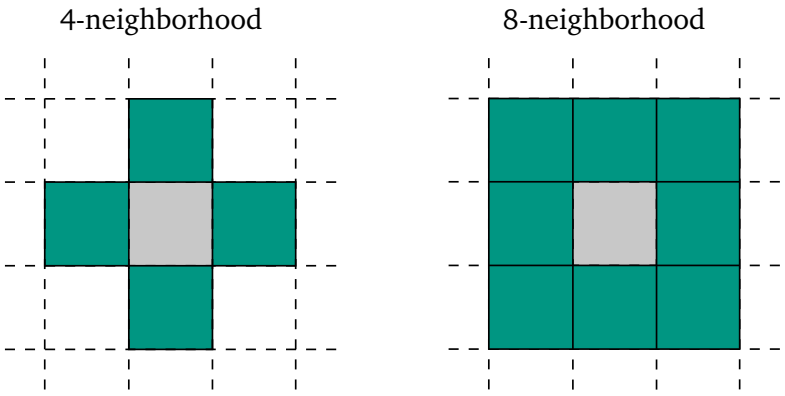

Figure 3.27: Depiction of a typical 4-neighborhood (left) and 8-neighborhood (right) of the central pixel (gray).

The basic idea is to initialize a starting position on the image, i.e. a first pixel, as the current region. Then the adjacent pixels (either constrained by a 4-neighborhood or 8-neighborhood depicted in Figure 3.27) are tested against a certain criterion of homogeneity based on the characteristics of the starting pixel or dynamically adapted characteristics of the region. Pixels that satisfy the criterion of homogeneity are added to the current region. Thus, the region grows iteratively until no more pixels meet the criterion or all pixels have been tested.
The main difference introduced in the method is the admission of *disjoint* image components being aggregated in the same cluster. The constraint of pixel connectivity is therefore relaxed from direct adjacency to a more general notion of spatial proximity – i.e. larger cell neighborhoods than in image processing are allowed. The reason for this alternative approach is motivated by the non-continuous nature of human settlements. In contrast to general objects in image processing applications, these are rather scattered and not represented as one continuous region in the grid, but still can be affected by the same event – i.e. they belong to the same

"object" in image processing terms.
The specific neighborhood could, of course, be adjusted on a per event type basis. However, if not constrained to the first sub-domain level of the taxonomy, this causes a pointless increase in complexity of the clustering method. Moreover, as I aim for a rather generic approach in general, the clustering should not be tailored too much to particular needs of a specific domain.

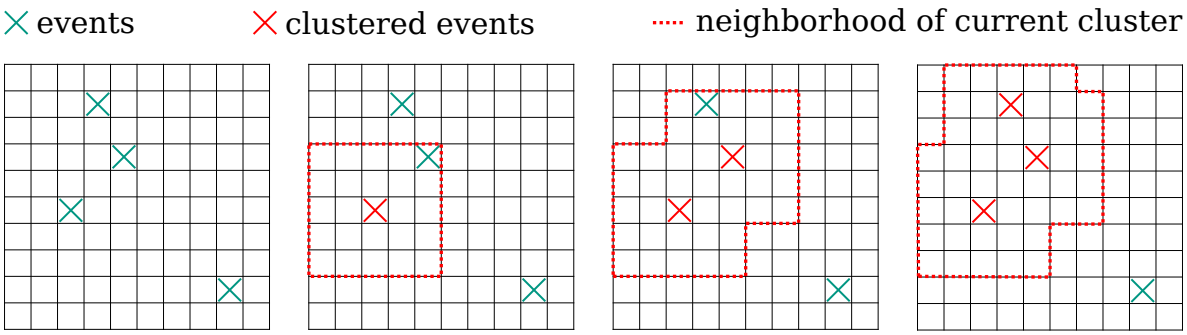

Figure 3.28: Example of a spatial-thematic clustering of event cells based on region growing.

Figure 3.28 shows an example of the clustering process using a 24-neighborhood as currently implemented in the prototype. This setting matches an influence radius for each event cell of approximately twice its size. Here the cells represent disjoint regions in real space, which is a simplified view for the sake of clarity. The oversampling approach introduced in Section 3.3.1 results in cells that represent overlapping regions in real space. Thus, the system actually has to search a 80-neighborhood to achieve the results exemplified in Figure 3.28.

An advantage over the common algorithm is the a priori knowledge which cells could potentially be clustered – that is only the classified cells of the time interval. Thus, the process can be significantly accelerated. Instead of searching the complete neighborhood at each step, just the remaining candidates have to be tested if they fall into the current neighborhood region of the cluster or not.
Another adaption is applied that concerns the starting position, which is often randomly chosen in region growing algorithms. For the situation here, the classified event cells are sorted according to their depth in the taxonomy in decreasing order – i.e. the clustering starts from the cell that has the longest direct path to the root node. This means that the system tries to cluster the cells with a high classification granularity first. Based on this condition, the criterion of homogeneity is defined as:

> An event cell in the current neighborhood is added to the cluster, if
>
> > it has the same class label as the starting cell
>
> or
>
> > its class label is on the direct path from the label of the starting cell to the root node in the taxonomy and the similarity scores *overlap*,
>
> or

its class label has the same direct parent in the taxonomy as the starting cell and the similarity scores *overlap*.

In the natural disaster taxonomy (see Figure 3.23) a possible scenario is

$$\textit{tornado} \quad \rightarrow \quad \textit{meteorological} \quad \rightarrow \quad \textit{natural disaster}$$

as direct path and e.g. *tornado* and *tropical storm* as nodes with the same direct parent, i.e. *meteorological*. *Overlapping* similarity scores should denote that, using the example above, the cell classified as *tornado* yielded a similarity score for *tropical storm* larger than zero or vice versa – i.e. both cells shared similar content to some extent and thus have a kind of fuzzy membership in both classes.

At the end of the clustering process, the label with maximum depth is kept as the final class label for a cluster. However, in case this is ambiguous, the system again reverts to the similarity scores for the labels in question of the respective cells. Eventually, the label with the highest average similarity score is set as the final class label for the cluster.
Let me provide a simple example for this procedure. Let us assume 4 cells – 1 classified as node *meteorological*, 1 classified as *tornado*, and 2 classified as *tropical storm*, then Table 3.6 presents the state of things in a structured way.

Table 3.6: Example of a cell cluster of four cells with differing final class labels.

Cell	Label
1	*meteorological*
2	*tornado*
3	*tropical storm*
4	*tropical storm*

Hence, *tornado* and *tropical storm* are in consideration for the final class label. Therefore, their respective similarity scores in the cells 2, 3 and 4 are taken into account (depicted in Table 3.7).

Table 3.7: Comparison of a cell ranking according to the similarity scores for the two different labels (a) *tornado*, (b) *tropical storm*

Cell	Similarity
2	0.55
3	0.31
4	0.27
average	0.38

(a) Similarity scores for *tornado*

Cell	Similarity
2	0.13
3	0.33
4	0.30
average	0.25

(b) Similarity scores for *tropical storm*

Although, there are more cells labeled as *tropical storm* in total, on average over the whole cluster the label *tornado* is obviously more prominent and hence set as final class label. However, just like for single cells, the individual averaged scores are kept. On the one hand

they provide a possible human operator with more context knowledge to assess the reliability of the classification result. On the other hand these scores can be exploited for the next step in the framework – the temporal monitoring of events described in Section 3.5.2.

Theoretically, the system may be confronted with as many different event types as present in the applied taxonomy – e.g. 15 in the disaster taxonomy (10 leaf-nodes, 4 sub-domain nodes and one root node). However, the experience with the prototype shows that there are rarely more than three different class labels per time interval. Yet, depending on the domain, this number might be higher and more diverse.

Real-World Example Figure 3.29 depicts a real scenario of the detection of an earthquake which exhibits a typical case of spatial-thematic clustering in the prototype. The earthquake took place between Stillwater, Oklahoma and Oklahoma City on April 19th, 2015 at 5:27:14 UTC and had a magnitude of 4.2 with its epicenter at 97.332°W and 35.953°N. It is depicted as red dot in Figure 3.29. The framework detected the event for the time interval from 5:27:10 to 5:28:10. The analysis of this time interval finished at 5:28:16 and the automatic e-mail alert arrived at 5:28:20.

The majority of the clustered cells (green shaded rectangles) were classified as *earthquake* and some (on the bottom of the map) were classified as *geophysical* because of some tweets containing the term *eruption*[36]. As *geophysical* is on the direct path from *earthquake* to *natural disaster* in the taxonomy, it is added to the cluster based on the above defined criterion of homogeneity.
In this case, I tried to visualize the fact that the cells in the grid for the statistical analysis represent overlapping regions in real space. Therefore, the dashed gray lines depict the oversampled grid and the clustered cells are depicted in a transparent green shade (always encompassing four oversampled cells) – i.e. the darker the green the more overlapping cells were clustered. As a side-effect, this visualization approach yields a suitable depiction for an impact area of the event based on TWITTER user activation.

This specific event also illustrates the importance to account for disjoint cells being affected by the same real-world event. Here, the TWITTER users of Stillwater and Oklahoma City both felt the earthquake. The epicenter, however, actually does not lie in an event cell, which is most likely due to its low population. At this point, another characteristic of the approach becomes apparent again – i.e. it neither necessarily detects, nor is it aiming for, the exact location of an event (especially concerning natural disasters), but where it affects a critical mass of people.

[36] The usage of this "wrong" term might be due to people in Oklahoma not yet being accustomed to earthquakes. Oklahoma has been a rather quiet region concerning earthquakes before 2009 but it surpassed California in the total number of earthquakes with a magnitude higher than 3.0 in 2014 (derived from data of the **U**nited **S**tates **G**eological **S**urvey (USGS) earthquake archives available at `http://earthquake.usgs.gov/earthquakes/search/`). According to Keranen et al. (2014), the increase is likely caused by fluid migration from wastewater disposal wells of unconventional oil and gas production facilities – i.e. so-called *induced seismicity*.

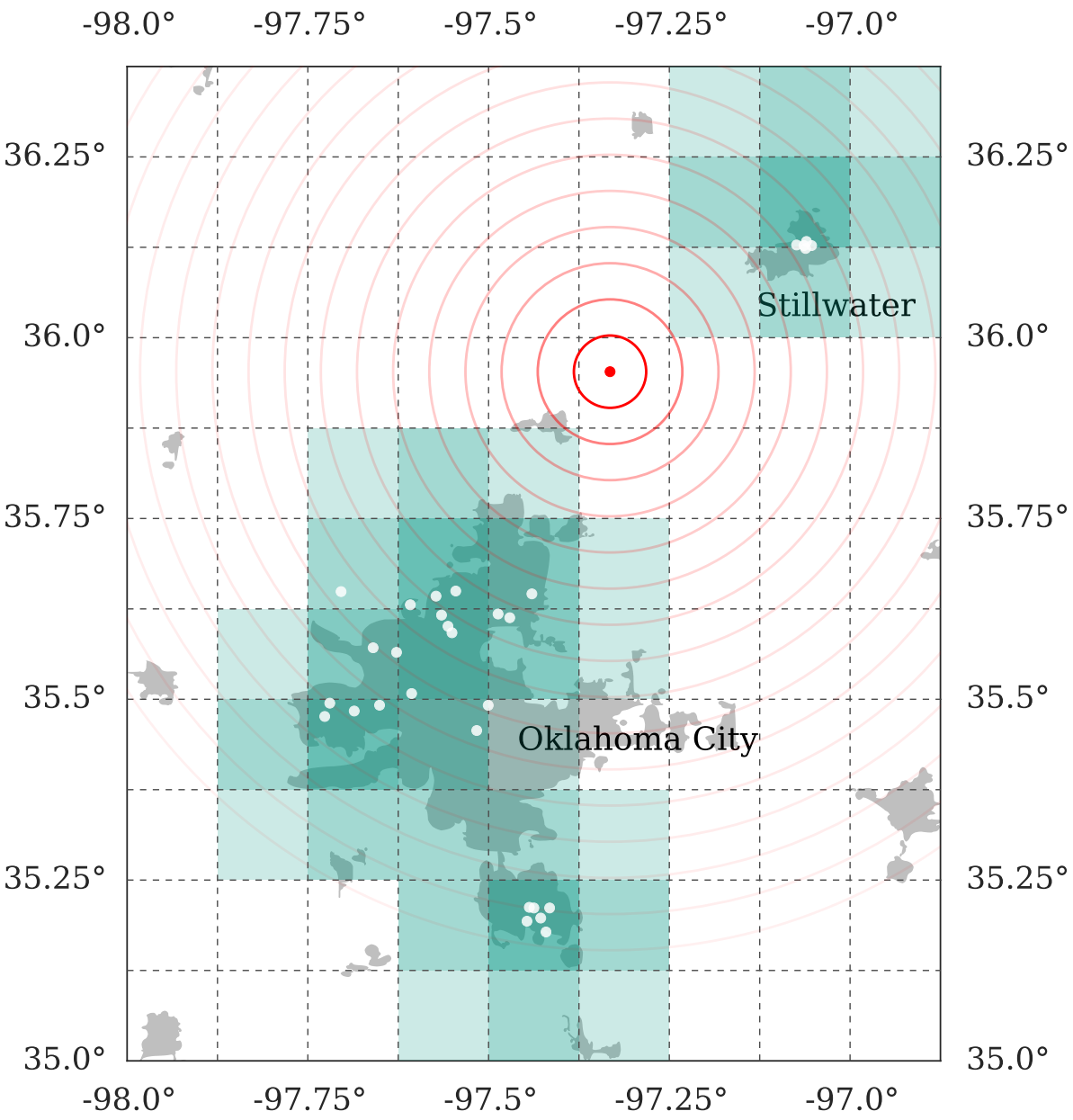

Figure 3.29: Result of spatial-thematic clustering of an earthquake event. The map depicts the cells of the oversampled grid (dashed gray lines), the tweets (white dots), urban areas (medium gray), clustered cells (green shaded rectangles) and the earthquake epicenter (red dot)

3.5.2 Temporal Monitoring as Clustering

Now that events which spread over larger areas than covered by the cells are taken into account, long lasting events are also considered – i.e. events that, because of their characteristics or severity, affect people and potentially trigger them to disseminate messages on social media over a longer period of time. With respect to the formal approach, this translates to expanding the spatial-thematic clustering over a certain number of time intervals. This number could be adapted for different event types of a domain (e.g. by incorporating domain experts and/or heuristics). The information already persistently stored in the event database is extended and updated accordingly, e.g. by adding the new messages of the current event cluster to the

database and linking them to the original event.
The general clustering approach I employ is straightforward:

> For each current event cluster, the database is queried for another cluster,
>
> > whose impact area overlaps with the current one,
>
> and
>
> > which is not older than the set number of time intervals to be considered,
>
> and
>
> > that meets the criterion of homogeneity for the spatial-thematic clustering concerning its final class label.

However, in the very unlikely case that the query yields more than one potential *historic* cluster, a fast process is needed that yields the most plausible match. Therefore, a heuristically driven algorithm is implemented to establish a decision based on how well the historic clusters agree with the database query. Hence, the system has to prioritize the three query parts. The ranking was chosen as (i) criterion of homogeneity (general) over (ii) temporal distance over (iii) size of spatial overlap.Within the first, another prioritization has to be introduced. The attributes are prioritized as follows:

1. same class label
2. direct path
3. same direct parent

Hence, Algorithm 3.1 depicts the decision process after querying the database. In case the query yields no results, the current cluster is classified as a new event, is stored in the event database and generates an automatic alert.

3.6 Operational Aspects of the Prototype

As mentioned throughout this thesis, I implemented a prototype that exploits the global real-time stream of georeferenced tweets and analyzes specific natural disasters, according to the methods explained in the preceding sections of Part I. The prototype is called **T**witter **E**vent **N**otification and **A**nalysis **S**ervice (TENAS) and shows the feasibility of the methods for a real-world application. As the explanation of the approach already mostly relied on the prototype, now some of the more technical issues of the system are presented.

Algorithm 3.1. Matching a current event cluster to a historic event

```
R ← set of historic event clusters matching the database query
l ← class label of the current cluster
E ← current event cluster
function BESTMATCH(R, l, E)
    sameLabel ← all clusters from R with label l
    directPath ← all clusters from R whose label is on direct path from l to root
    sameParent ← all clusters from R whose label has the same direct parent as l
    if R = {} then
        return {}                                          ▷ New event detection
    end if
    if sameLabel ≠ {} then
        if len(sameLabel) > 1 then
            return SUBRANKING(sameLabel, E)
        else
            return sameLabel
        end if
    else if directPath ≠ {} then
        if len(directPath) > 1 then
            return SUBRANKING(directPath, E)
        else
            return directPath
        end if
    else if len(sameParent) > 1 then
        return SUBRANKING(sameParent, E)
    else
        return sameParent
    end if

end function
function SUBRANKING(C, E)
    mostCurrent ← getMostCurrent(C)
    if len(mostCurrent) > 1 then
        largestOverlap ← getLargestSpatialOverlap(mostCurrent, E)
        if len(largestOverlap) > 1 then
            highestScore ← getHighestScore(largestOverlap, E)
                                    ▷ cluster with highest similarity score for label l
            return highestScore
        else
            return largestOverlap
        end if
    else
        return mostCurrent
    end if
end function
```

3.6.1 Computational Resources

TENAS's main modules are implemented in JAVA[37]. For some of the numerical calculations involving the spatial grid, the framework issues system calls to MATLAB[38] processes. Most

[37] A general-purpose, concurrent, class-based, object-oriented programming language (cf. Gosling et al., 2015).
[38] A proprietary (interactive) computing environment for numeric computation, visualization, and data analysis (cf. Cavers, 1998), developed by MATHWORKS.

of the computer linguistic methods are integrated via system calls to PYTHON[39] scripts. The framework runs 24/7 on a server with an Intel Core i7-3820 processor (3.6 GHz, Octa Core) and 64GB of RAM running OpenSuse 13.2.
In order to exploit the full capacity of the hardware, the code makes use of JAVA's straightforward threading capabilities and task scheduling. Hence, the system continuously collects and processes incoming messages and, in parallel, the framework analyzes the preceding time interval at scheduled time steps. Consequently, the workload is well distributed over all CPU cores.

MONGODB, a document-oriented database system is employed as main storage technology and database back end. It is part of a relatively new group of database architectures, which are not based on a relational schema. This group is often referred to as NoSQL databases. However, some can be queried with **S**tructured **Q**uery **L**anguage (SQL) and thus the acronym is rather meant as *Not only* SQL.
MONGODB uses the JSON format. Hence, with TWITTER as input source, no data transformation steps are necessary, but each tweet is stored *as-is*. Moreover, MONGODB offers two-dimensional spatial indexing for efficiently querying georeferenced data and enables regular expressions for fast keyword searches. Additionally, an index is created on the field holding the time stamp, as the system often has to retrieve space-time slices from the data – i.e. the data of a cell from a specific time interval. I also make use of MONGODB's *capped connection* setting, which enables the database to "forget" data after a specific time. In the prototype, the ongoing message flow is only kept for one hour, mainly to be accessible for the thematic classification (see Section 3.4.3). Without deleting older data, the system would need to accommodate for a daily amount of more than 40GB of tweets and an additional 5GB for the indexes.

In order to provide real-time dictionary look ups and spelling correction, the capabilities of ELASTICSEARCH[40] are exploited. It is basically a powerful search engine that also provides storage facilities in JSON format. The integration of custom functions for tokenization, stop word removal and tf-idf weighting is also possible.

3.6.2 Alert Mechanism

At the moment of writing, the alert mechanism is implemented as an automatic e-mail message issued to recipients which can subscribe to the service on a per event type basis. So far it is an internal service for other researchers in CEDIM, where the main part of this research is based. In the future, this service could be opened up and might issue the alerts back to its data source – i.e. disseminate the extracted event information as tweet. This way any interested individual or organization could follow TENAS updates without any changes on the service side.

The alert message incorporates the most important event information that the system derives from the messages in a compact format. As an example, Figure 3.30 shows the information

[39] An interpreted, higher programming language suited for rapid development, production deployments, and scalable systems (cf. Gorelick et al., 2014).
[40] Avaliable from `https://www.elastic.co/products/elasticsearch`.

that was sent after the detection of an earthquake in Guayaquil, Ecuador on October 24, 2013. The mail arrived at 0:39:08 UTC, with a processing time of 3 s, that means a 6 sec delay is caused by the mail delivery process. The use of TWITTER could often reduce this delay to below 2 sec (see Section 3.1.2). The information includes the date and time the event was detected, the coordinates of the centroid of the messages, the relevant similarity scores, the detected terms and their frequency, and the number of messages as well as the 95% threshold from the respective count data model(s).
According to the USGS earthquake archive, the event happened at 0:37:20 UTC with its epicenter at 79.790°W and 2.060°S[41]. TENAS detected the first reactions related to the earthquake 1 min and 46 sec after it occurred. The messages' centroid was at an approximate distance of 15.2 km from the epicenter.

```
TWITTER EVENT!!!

Event ID:    52686c29e4b0780f20cd64ca
Date:        24.10.2013
Time:        0:39:00 UTC
Place:       Guayaquil, Guayas, Ecuador
             longitude   -79.90152
             latitude    -2.14005

Similarity Scores:
    earthquake  0.67

Terms:
    temblor     39

Statistics:
    number          55
    95%-threshold   15
```

Figure 3.30: The e-mail notification for the earthquake in Guayaquil, Ecuador on October 24, 2013 at 0:37:20 UTC with its epicenter at 79.790°W and 2.060°S.

In order to make the location information more comprehensible for a human operator, reverse geocoding is applied to the geographical coordinates of the centroid. The NOMINATIM web service[42] is used, which provides the *Place* information in a structured format.

3.6.3 Ad Hoc Visualization

As a small add-on for a human operator to get a first glance of the impact area, the alert mail includes a link to a custom web service. The service is implemented in JAVASCRIPT using the library NODE.JS for the web server and the library LEAFLET.JS to visualize an interactive map based on OSM data. The single tweets are depicted as markers. By clicking on a marker a small

[41] Event information available from http://earthquake.usgs.gov/earthquakes/eventpage/usc000km1p.
[42] The search engine for **O**pen **S**treet **M**ap (OSM) data accessible via the base-URL http://nominatim.openstreetmap.org/reverse?.

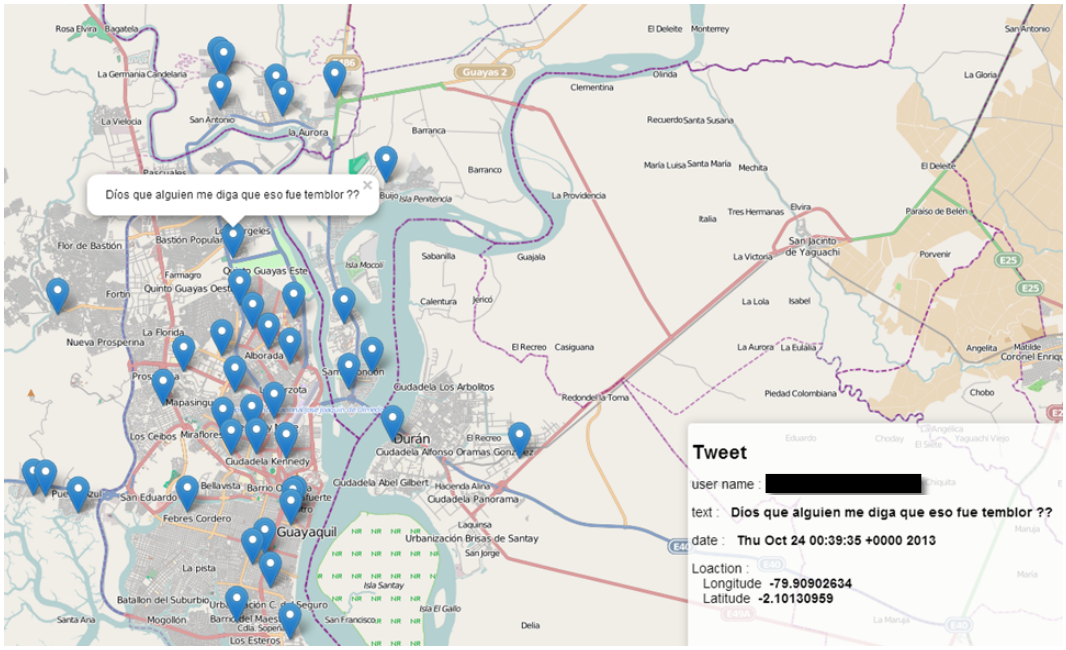

Figure 3.31: Ad hoc impact area visualization in a browser with OSM data as background and overlain by tweet metadata.

window shows the text, the date and time, and the coordinates of the message. Figure 3.31 shows the map corresponding to the earthquake alert mail presented in the preceding section.

3.7 Summary

In this chapter, the used methods for real-time event analysis using social media data have been explained.

First, the exploited input data and its attributes in general, as well as the relevant parts used in the approach has been introduced. Then the spatial and temporal characteristics have been investigated to enable an informed decision for the choice of suitable discretization approaches for space and time. The detection of significantly increased message volumes on a per cell and per time interval basis using different count data models has been described.

The idea of a thematic classification based on domain-dependent document similarities and a domain taxonomy has been detailed. The similarities allowed for the implementation of a class label ranking rather than only an assignment of fixed class labels. Moreover, the spatial-thematic and temporal clustering approaches have been described adapting a common algorithm from the field of image processing. Eventually, operational aspects of the prototype have been provided, including the automatic alert mechanism and the ad hoc visualization.

The next chapter, will evaluate the detection capabilities of the approach based on an earthquake ground truth dataset, with respect to the detection rate, the temporal efficiency and the spatial proximity to the epicenter.

4 Experimental Results

In this chapter, the detection performance of the developed algorithm described in the preceding sections will be numerically evaluated. Eventually, Section 4.3 concludes this chapter as well as the first part of this thesis with a critical discussion and a short summary.

4.1 Experimental Setup

For a numerical evaluation of the developed system, it is necessary to acquire so-called *ground truth* data, in this case, trustworthy and verified information on the time, the location and the type of large-scale events. Due to the focus of the prototype implementation, the evaluation will be concerned with the capabilities of the system related to natural disaster events. To exemplify a robust evaluation, I rely on historical earthquake data, which is, in contrast to other types of disasters, consistently available on a global scale and supplied with informative metadata. Moreover, earthquakes can be rather accurately assigned to a position in space and time. Thus, they allow for an appropriate comparison of real earthquakes and the events which the system detects and classifies as earthquakes.

4.1.1 Evaluation Set

As a resource for ground truth data, the **A**dvanced **N**ational **S**eismic **S**ystem **Com**prehensive **Cat**alog (ANSS ComCat) is used, which is an online accessible earthquake database combining earthquake source parameters and other products generated by a large group of contributing seismic networks[1]. The database can be queried via a web-interface[2] or accessed through its API.
The ANSS ComCat contains several pieces of information for the stored earthquake events. The most important pieces for the evaluation are

- *time*
  the time when the event occurred,
- *longitude*
  the longitude of the earthquake's epicenter in decimal degrees (negative values for western longitudes),
- *latitude*

[1] A list of all contributing networks is available at http://earthquake.usgs.gov/earthquakes/map/doc_aboutdata.php#contributing-networks.
[2] Accessible at http://earthquake.usgs.gov/earthquakes/search/

the latitude of the earthquake's epicenter in decimal degrees (negative values for southern latitudes),

- *magnitude*

  the magnitude for the event,

- *depth*

  the depth of the event in kilometers, i.e. vertical distance from epicenter to hypocenter.

For some events there is even more information available such as the horizontal distance from the epicenter to the nearest station or the total number of seismic stations used to determine the earthquake's location. For the evaluation, only the consistently populated fields in the set are used. However, additional information is derived, such as the local time[3], the country[4], the distance to the closest major city[5] (population larger than 100K), and the maximum *intensity radius* – the approximate maximum radius in which the earthquake can be felt by humans. These values will be considered in the evaluation to analyze missed events.

Felt Earthquake Intensities The mentioned maximum *intensity radius* is based on the Modified Mercalli Intensity (MMI) scale for felt earthquake intensities (see U.S. Geological Survey, 2000). The MMI scale categorizes the felt intensities of earthquake events in twelve classes labeled with roman letters from **I** to **XII**. Only intensities of **II** and above can be felt by humans, so the radius excludes areas of lower intensities.
Depending on the magnitude and the depth of the earthquake the intensities for different distances from the epicenter can be roughly approximated. The following are the most common formulas according to Dr. James Daniell (personal communication, December 10, 2015), a designated expert in the field of earthquake engineering. The formulas are given in Shebalin et al. (1997) with

$$mmi = 2m - 0.2 - 3\log(s) - 0.0008s, \tag{4.1}$$

in Ambraseys and Douglas (2000) with

$$mmi = \frac{1}{0.65}\left(\left(\frac{(m-1.176)}{0.817}\right) + 1.54 - 0.0029s - 2.14\log(s)\right), \tag{4.2}$$

and in Ambraseys (1985) with

$$mmi = 1.5m - 0.5 + 0.15\log(d) - 2.85\log\left(\frac{s}{d}\right) - 0.0024(s-d), \tag{4.3}$$

where $s = \sqrt{(r^2 + d^2)}$, m is the magnitude, r is the distance from the epicenter in kilometers and d is the depth in kilometers.
Again relying on external domain expertise of Dr. James Daniell (personal communication,

[3] The time zone information is again obtained from the IANA time zone database via the PYTHON library PYTZWHERE (see Section 3.3.2).
[4] Again, the NOMINATIM web service, the search engine for OSM data is used, which is accessible via the base-URL `http://nominatim.openstreetmap.org/reverse?` (see Section 3.6.2).
[5] Data on populated places at global scale is available from Natural Earth at `http://www.naturalearthdata.com/downloads/110m-cultural-vectors/110m-populated-places/`.

December 10, 2015), Equation (4.1) is used for shallow earthquakes ($d < 75\,\mathrm{km}$) in active regions, and Equation (4.2) for intermediate-depth earthquakes ($d \geq 75\,\mathrm{km}$) in active regions. For stable regions in the dataset, Equation (4.3) is used. The decision whether an earthquake occurred in a stable or active region is based on the information provided in the global seismic hazard map[6] of Giardini et al. (2003). The map provides probabilistic **P**eak **G**round **A**cceleration (PGA) values in $\mathrm{m\,s^{-2}}$ on a global scale. The values have a 10% probability to be exceeded for the duration of fifty years which corresponds to a 475 years repeat rate. Following Dr. James Daniell, active regions are defined where the PGA provided in the global seismic hazard map exceeds $1\,\mathrm{m\,s^{-2}}$. These regions have a 10% probability to undergo an event with an intensity as high as **V** with respect to the MMI scale in the next fifty years (cf. Wald et al., 1999).

Table 4.1: Temporal extent of the global evaluation set in three disjoint time periods.

	start date	end date
time period #1	2015-06-13 23:00:00 UTC	2015-06-15 08:49:00 UTC
time period #2	2015-06-18 11:00:00 UTC	2015-06-21 10:00:00 UTC
time period #3	2015-07-06 23:00:00 UTC	2015-07-12 12:05:00 UTC

Global Evaluation Set The global evaluation set was acquired for three distinct time periods. During these periods, the prototype was running without interruption. The time periods are listed in Table 4.1. This set contains all earthquakes around the world which took place during one of the three time periods and had a magnitude of larger than or equal to 3.0. Again according to the U.S. Geological Survey (2015), earthquakes with a magnitude below 3.0 commonly exhibit maximum intensities of **I** on the MMI scale and are thus excluded. The frequencies of the earthquakes with respect to their magnitude is depicted in Table 4.2 together with the total number and the respective numbers for events with onshore and offshore epicenters.

Concerning the spatial distribution with respect to political boundaries, the ground truth dataset contains earthquakes from approximately 60 different countries. Table 4.3 shows countries with five or more earthquakes in the dataset, again for the total number as well as split in onshore and offshore events.

Finally, Figure 4.1 depicts all earthquakes in the evaluation set on a global map with increasing size according to their magnitude range. Here I used larger steps for the magnitude ranges for a clearer visual arrangement of the legend. The events are depicted in different colors for onshore and offshore epicenter locations.

[6] Available at `http://www.seismo.ethz.ch/static/GSHAP/global/`.

Table 4.2: Number of earthquakes in total and split in onshore and offshore events, grouped by different magnitude ranges (Note: ranges are upper-bound exclusive!)

magnitude range	#EQs total	#EQs onshore	#EQs offshore
3.0 - 3.2	34	20	14
3.2 - 3.4	23	10	13
3.4 - 3.6	14	8	6
3.6 - 3.8	11	6	5
3.8 - 4.0	10	3	7
4.0 - 4.2	54	23	31
4.2 - 4.4	97	31	66
4.4 - 4.6	77	12	65
4.6 - 4.8	53	13	40
4.8 - 5.0	33	2	31
5.0 - 5.2	9	1	8
5.2 - 5.4	9	0	9
5.4 - 5.6	4	1	3
5.6 - 5.8	5	0	5
5.8 - 6.0	3	0	3
6.0 - 6.2	0	0	0
6.2 - 6.4	1	0	1
6.4 - 6.6	1	0	1
6.6 - 6.8	1	0	1
$\sum$	439	130	309

Table 4.3: Number of earthquakes per country with at least 5 events in total. The numbers are also given for onshore and offshore events.

Country	#EQs total	#EQs onshore	#EQs offshore
United States of America	83	44	39
Indonesia	40	4	36
Japan	36	5	31
Papua New Guinea	34	10	24
Solomon Islands	29	0	29
Tonga	19	0	19
Fiji	18	0	18
Chile	16	8	8
New Zealand	14	0	14
South Sandwich Islands	14	0	14
Russian Federation	12	3	9
Mexico	12	3	9
Afghanistan	11	11	0
China	7	7	0
Peru	6	5	1
British Virgin Islands	6	0	6
Dominican Republic	5	0	5

4.2 Results

In order to quantify the performance of the system with respect to the detection of earthquakes, all event detections are taken into account which the system yielded in the respective time

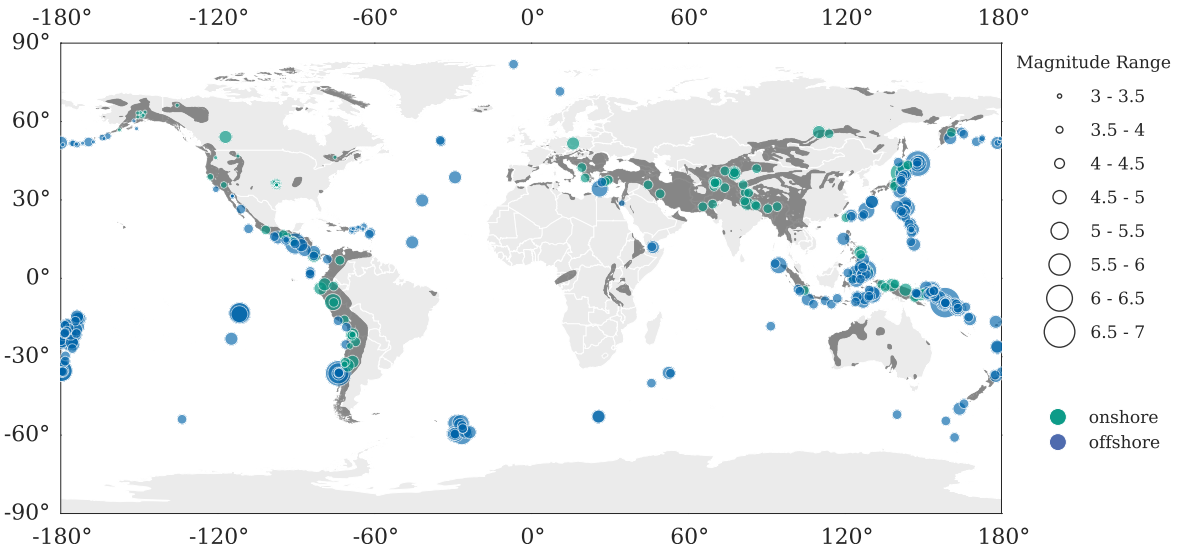

Figure 4.1: Visualization of the earthquake evaluation set acquired from the ANSS ComCat. The marker size is adapted according to the magnitude range of the earthquakes. Onshore events are depicted in green and offshore events in blue. Additionally, dark gray denotes areas classified as seismically active according to the global seismic hazard map of Giardini et al. (2003).

periods and which were classified as *earthquake*. The detection rate is the main metric used, i.e. the ratio of detected earthquakes divided by the total number. The detection rate with respect to several other event attributes will be reported and analyzed. Moreover, the temporal efficiency of the detection as well as the spatial proximity to the earthquakes' epicenter will be presented.
A detection is only considered within $20\,\mathrm{min}$ of the earthquake occurrence and within a distance from the epicenter of twice the maximum intensity radius.

Detection Rates The overall detection rates for the global evaluation set in total and split for onshore and offshore events are given in Table 4.4. As expected, offshore events are less likely to be detected.

Table 4.4: Overall detection rate for the global evaluation set in total and split in onshore and offshore events.

Total	Onshore	Offshore
52.62%	70.77%	44.98%

As another indicator why events were missed, the ratio of the intensity radius and the distance to the next major city are determined. Table 4.5 shows that in general earthquakes that have a small ratio, i.e. combined a small intensity radius with a large distance to a major city, are less likely to be detected. In contrast, the local time had no significant influence on the detection performance.

In Figure 4.2, the detection rates are presented for the different magnitude ranges existing in the dataset. First, the rates increase with the magnitude range as expected. Then however, a significant drop in the detection rate can be observed for events with magnitudes ≥ 4.0 and

Table 4.5: Median ratios of the intensity radius and the distance to the next major city in total and split in onshore and offshore events.

	Total	Onshore	Offshore
detected	0.207	0.311	0.136
missed	0.041	0.165	0.038

< 4.4 and also lower rates for magnitudes of ≥ 4.4 and < 4.8. The trends for onshore and offshore events are qualitatively rather similar.

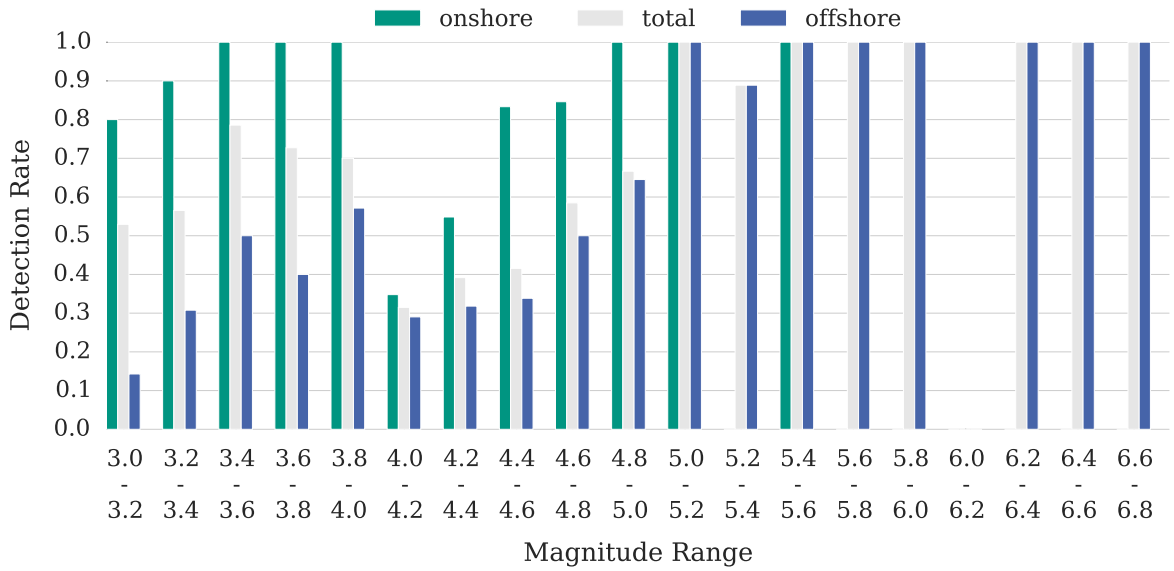

Figure 4.2: The detection rates per magnitude ranges in total (gray) and split for onshore (green) and offshore events (blue).

Focusing on this problematic magnitude range for onshore events with a detection rate of only 58.2%, the reasons for the low rates can be identified by looking at the respective country statistics. The bar chart in Figure 4.3 depicts the number of onshore earthquakes detected and in total, for countries that experienced at least 3 onshore events with magnitudes ≥ 4.0 and < 4.8. The countries are ordered according to their detection rate from top to bottom. Clearly, in some countries the earthquake detection capabilities of the system is much lower than in others. Moreover, especially for the three countries with the most events in this category (Afghanistan, Papua New Guinea and China) the detection rate is rather low.

Additionally, the figure depicts the average number of daily georeferenced tweets per 1K inhabitants in the countries[7]. The value will be referred to as average TWITTER activity in the following. Not surprisingly, the detection rate strongly correlates with the average TWITTER activity of the country. However, for the problematic category in total – i.e. $4.0 \leq$ magnitude < 4.8 and onshore epicenter – 27 out of the 33 missed events occurred in countries with less than 1 georeferenced tweet per day per 1K inhabitants. In contrast, for the remaining onshore events – i.e. magnitude < 4.0 or ≥ 4.8 – the detection rate was 90.2% and the events originate

[7]The population data per country is derived from the World Development Indicators, The World Bank, accessible through http://data.worldbank.org/data-catalog/world-development-indicators.

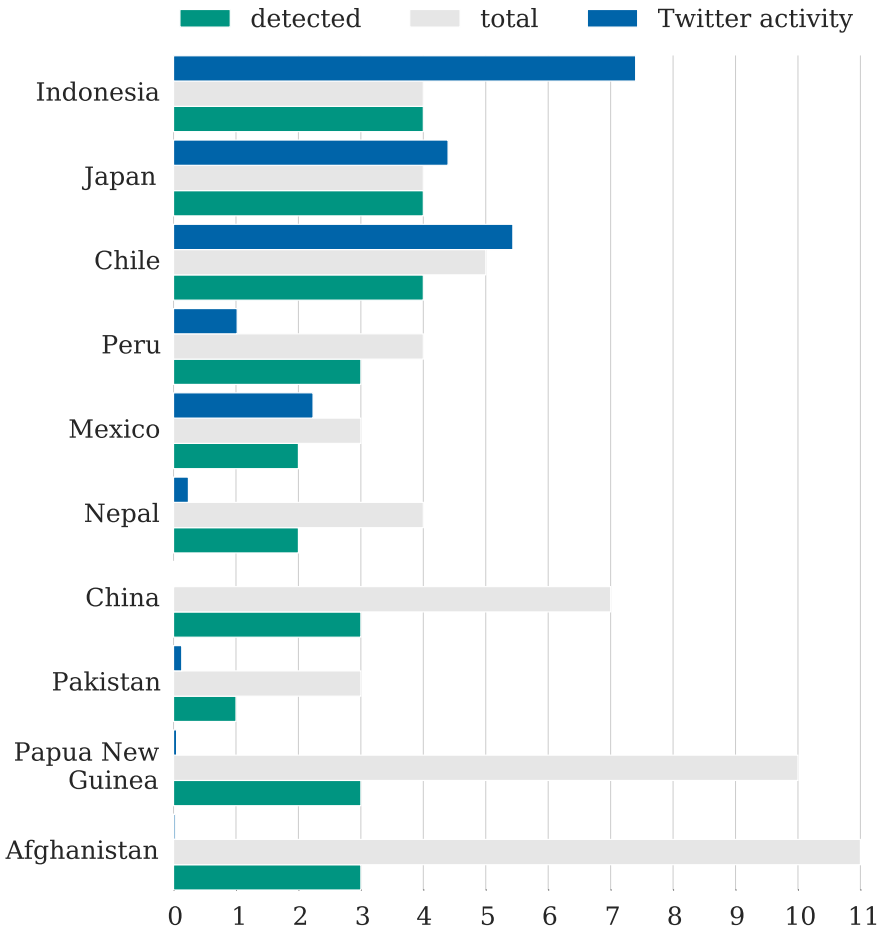

Figure 4.3: The number of onshore events with magnitudes of ≥ 4.0 to < 4.8 in total (light gray) and the number of detections (green) per country that exhibited at least 3 events. The blue bars depict the average number of daily georeferenced tweets per 1K inhabitants.

from 7 different countries from which each exceeds 1 georeferenced tweet per day per 1K inhabitants (see Figure 4.4).
Following the MMI scale, earthquakes that have a magnitude of ≥ 5.0 have the potential to actually generate damage, corresponding to intensities of **VI** and above. In this group, the detection rates are even 96.97% (total), 100% (onshore) and 96.77% (offshore).

Temporal Efficiency The *detection time* is the time that passed between the occurrence of the earthquake and the respective detection in the system. Hence, it is a measure of the temporal efficiency.
The overall median detection time for the evaluation set is $4\,\text{min}\ 2.4\,\text{s}$ with a median absolute deviation of $2\,\text{min}\ 24.4\,\text{s}$. More than 94% of the events could be detected in less than 10 minutes. The results for the different magnitude ranges are shown in Figure 4.5. The box plot depicts the median, the interquartile range (IQR), the 5th percentile (lower whiskers) and the 95th percentile (upper whiskers), and where necessary the outliers.

For a very general comparison of my system with established real-time earthquake notification services, I use the Earthquake Notification Service (ENS) of the USGS. For the evaluation set, the ENS sent notifications for 24 events, three of which my system did not detect. The

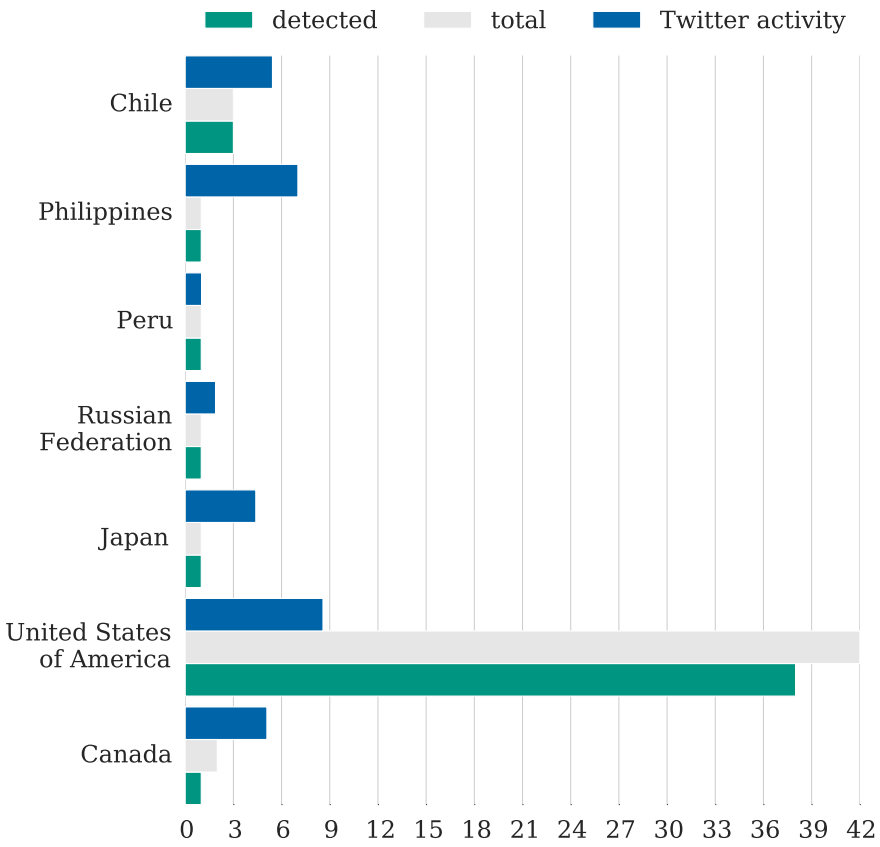

Figure 4.4: The number of onshore events with magnitudes of < 4.0 or ≥ 4.8 in total and the number of detections per country. The blue bars depict the average number of daily georeferenced tweets per 1K inhabitants.

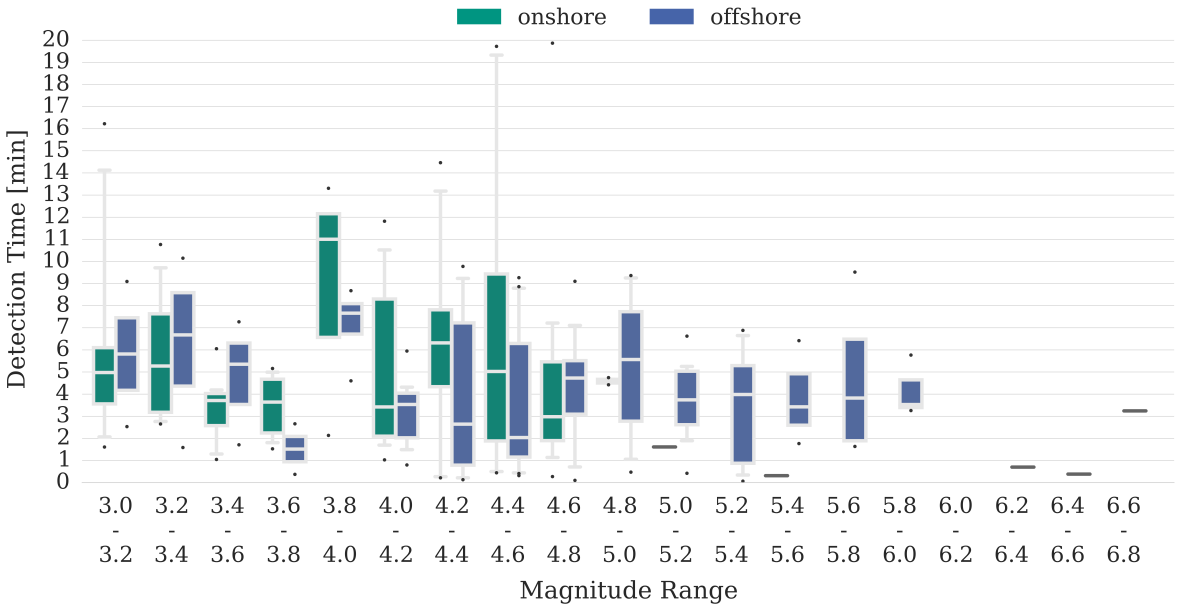

Figure 4.5: The median detection time per magnitude range (light gray line in boxes), split in onshore (green) and offshore (blue) events, together with the IQR (colored boxes), the 5th percentile (lower whiskers) and the 95th percentile (upper whiskers) and outliers (black dots). For the sake of visual clarity, magnitude ranges with only one value are highlighted in a darker gray.

respective detection times for these events by ENS, were 63 min, 198 min and 2954 min (i.e. more than 47 h). Concerning the remaining 21 events, the system consistently outperformed the ENS. The detection times are presented in Figure 4.6. However, due to the low number of instances, this comparison might not be completely representative.

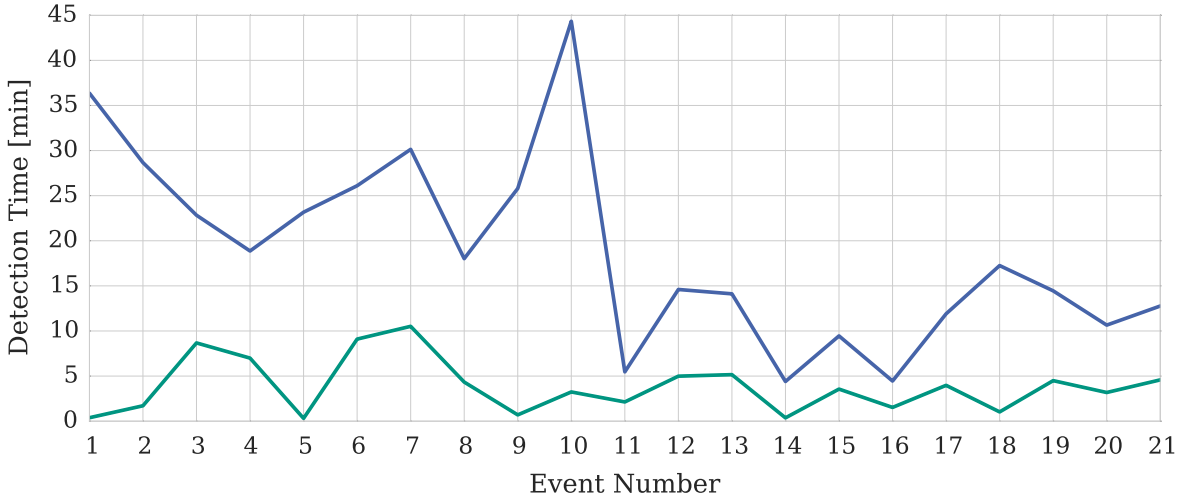

Figure 4.6: A comparison of the detection times for 24 earthquakes from the evaluation set which were automatically reported by the ENS of the USGS. TENAS, the system developed in this work is depicted in green and the ENS in blue.

Spatial Proximity to Epicenter The centroids of the messages of the first detection by the system are considered as the estimated earthquake impact location. These values are compared to the estimated epicenter locations provided in the evaluation set. The results for all detections and split for onshore and offshore events are presented in Table 4.6 as mean and standard deviation.

Table 4.6: Overall detection distances for the global evaluation set in total and split in onshore and offshore events

Total	Onshore	Offshore
32.04 km±40.48 km	20.03 km±30.09 km	40.11 km±44.24 km

The estimated maximum intensity radii can be considered as a rough spatial constraint for a plausible detection. Out of the 229 detected earthquakes in total, 87.3% were detected within their respective maximum intensity radius. Except for two events, the remaining events were all detected within a 15% limit above the respective radius. Splitting in onshore and offshore events, the percentages are 91.3% and 84.7% respectively, wherein the two events outside of the 15% limit happened offshore. Figure 4.7 depicts the detected events and their corresponding estimated maximum intensity radius with the individual 15% add-on, and their detection distances.

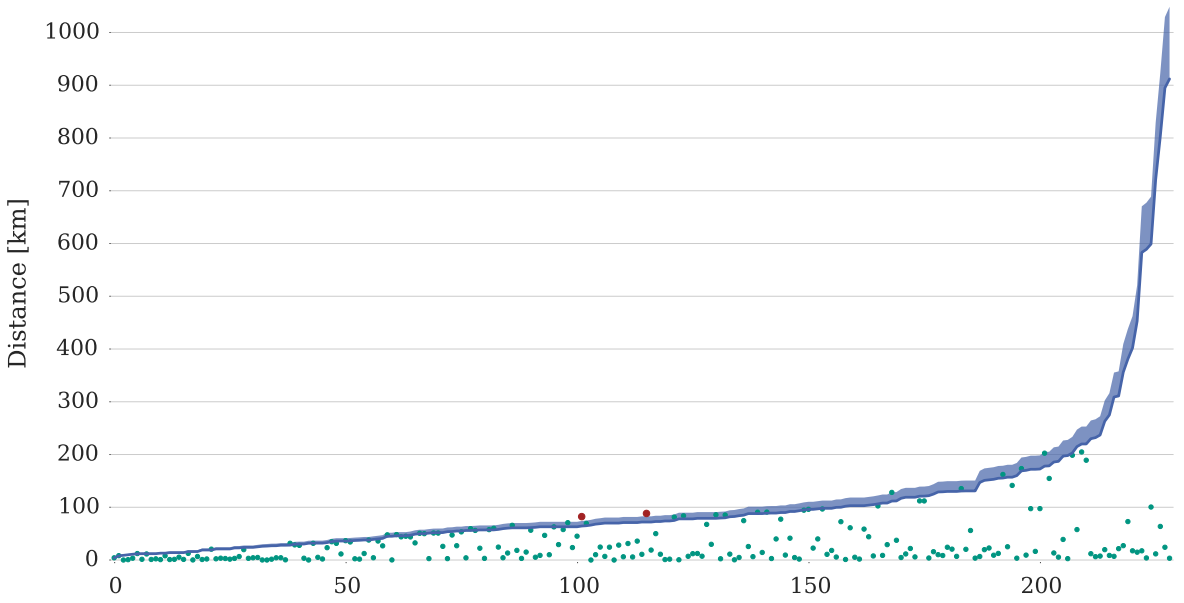

Figure 4.7: The detected earthquakes (green dots) ordered from left to right by increasing maximum intensity radius. The vertical axis encodes the detection distance to the epicenter. The dark blue line depicts the respective maximum intensity radius together with an individual 15% add-on (visualized as light blue shade). The two events detected outside of that range are colored in red.

4.3 Discussion

The results show that the detection of specific events is feasible at a high detection rate on a global scale together with good temporal efficiency and plausible spatial proximity to the earthquakes' epicenter.

As expected, the results indicate that the proximity to large urban areas is beneficial for a successful detection. In terms of spatial coverage, the results revealed that the system suffers from drawbacks in some political regions where the exploited social media platform TWITTER is either not very popular (e.g. Russian Federation) or (partly) blocked (e.g. China), or where internet access in general is limited (e.g. Afghanistan, Papua New Guinea). In countries with a high TWITTER penetration in contrast, the detection rates are very high (e.g. Japan, United States of America, Indonesia, Chile, Turkey, Argentina). The overall detection rate for onshore events is slightly skewed due to the predominance of events in the magnitude range ≥ 4.0 to < 4.8. Within this range, a significant part of the missed events took place in countries with very low TWITTER activity. Concerning events that are likely to have damaging impact however, the detection rate is almost perfect with 96.97%.

Considering temporal aspects, the majority of events could be detected in less than 10 minutes. For a subgroup of the evaluation set, real-time notifications from USGS were available. The comparison revealed that the developed system is faster than the official service ENS in all but 3 cases.

Although the aim is not to detect earthquake epicenters, but the main impact area with respect to affected people, the results show that the detections occurred in reasonable spatial proximity to the epicenter. Moreover, the comparison to the estimated maximum intensity radii also

demonstrates that the detection locations are within a plausible distance – i.e. people on site were potentially able to experience the earthquake themselves.

Concerning the covered time periods of the evaluation set, the system yielded not a single false alarm. For earthquakes in general, all alarms could be assigned to a real event by manual post-processing. In case of other natural disasters the retrieval of information for local events on a global scale is not feasible and mostly even unavailable. Nonetheless, since the beginning of this research, a couple of incidents took place which were obvious false alarms. The most frequent cause with three occurrences, was an alarm classified as a *volcanic eruption* in Turkey. By manual investigation, however, it could be identified as a soccer-related issue – the goalkeeper of the Turkish national soccer team is called Volkan Demirel, and his first name happens to be the Turkish word for volcano. Similar cases are not entirely impossible but so far have not been occurred.

4.4 Conclusion

In this chapter, the detection capabilities of the developed system were evaluated based on a global earthquake ground truth data set. The detection rate with respect to different attributes as well as the temporal efficiency and spatial proximity to the earthquakes' epicenter were presented and critically discussed. As exemplified on available cases, the system outperformed even an official, dedicated earthquake notification service.

Other Event Types In general, the system naturally shows its strengths in terms of temporal detection efficiency with events that exhibit a sudden and rather unexpected impact such as earthquakes or volcanic eruptions from the natural disaster domain. Events that do not have a real climax but a very long start-up period are sometimes not detected at all, for example due to a small effect on people. Sometimes these are only detected at an arbitrary time during their existence. However, according to the definition of an event given in Section 2.1, the first detection symbolizes the status when the event activated a critical mass of users to react to the event rather than a domain specific definition – such as the forming of a tornado.
During the last three years, two different cases of unexpected but interesting results were observed in the framework. The first case can occur when the population is warned by the authorities beforehand that a disaster event such as a hurricane is expected to hit at some specific time and in a specific area. This has led to two incidents, where the system issued an alert *before* the event actually took place. This of course was not due to any amazing capability of predicting natural disasters. The system was simply triggered by the significantly increased TWITTER traffic yielded by users within the area, who were already worrying or just chatting about the upcoming event.
The second case has only occurred once so far. On October 12, 2013 the category 4 cyclone *Phailin* with approximated wind speeds of 222 $\frac{\text{km}}{\text{h}}$ hit the coast south of the city of Brahmapur in the Indian federal state of Odisha (Mühr et al., 2013). However, the Indian Meteorological Service (IMD) issued warnings days ahead of landfall and approximately 1.7M people were

evacuated prior to the event (IFRC, 2013). Thus, one of the largest evacuations in Indian history (see Mühr et al., 2013), not only made the system completely miss this major event, but also caused a significant *decrease* in tweet volume in some affected areas.

The chapter concludes the first part of this thesis "Real – Time Event Analysis". A comprehensive recapitulation of this part together with the recapitulation of the second part is given in Section 8.1 in the form of a combined concluding summary.

Part II

SPATIAL INFORMATION EXTRACTION FROM TEXT

5 Introduction

This starting chapter of Part II introduces the field of textual spatial information in general, as the developed methods are completely independent from the joint use case of natural disasters. In fact, the core methods are not even limited to social media input, but can handle English text no matter what kind of source it originates from. The focus on social media obviously just complicates the task and calls for additional processing and pre-processing steps, as will be mentioned throughout Chapter 6.

First, Section 5.1 gives a short overview and details the most common encoding for textual spatial information in English according to the literature (cf. Herskovits, 1986; Landau et al., 1993; Miller and Johnson-Laird, 1976) – so-called *locative expressions*. Then the research gap will be identified based on related work and fundamental literature concerning spatial prepositions will be discussed (Section 5.2). Finally, in Section 5.3, important methods of NLP will be introduced that are used in the extraction and disambiguation process.

5.1 Textual Spatial Information

Spatial language understanding is a complex field drawing on different disciplines. The ability of a computer to recognize and interpret textual spatial information, such as place descriptions (*I'm waiting in front of the train station*) or route descriptions (*turn right at the post office*), is a particular challenge that has recently attracted attention from a wide range of diverse research communities such as computer science, robotics, computer linguistics, spatial cognition and **G**eographic **I**nformation **S**cience (GIScience). People use these kind of expressions in their daily life when making decisions and verbalizing their spatial knowledge about their surroundings; enabling computers for spatial language understanding will therefore be beneficial for human-computer interaction relating to everyday decision support, search, and similar generic information. But the automatic extraction and disambiguation of spatial information in particular, e.g. from large textual data streams, is a promising field for possible applications.

5.1.1 Locative Expressions

Textual spatial information can be encoded in a variety of syntactic categories, such as prepositions, adverbs, nouns and verbs, or any combination of them. In English, as well as many other natural languages, a very common means for people to express their spatial knowledge is LEs, as described in their prototypical form by Herskovits (1985).

LEs are spatial expressions incorporating a preposition, its object, and the entity the prepositional phrase modifies, i.e. the subject.

[5.1] the spider is on the wall.

In [5.1], the object *wall* of the preposition *on* is called the *relatum*, ground, anchor or landmark, and the entity the prepositional phrase modifies, i.e. *spider*, is called the *locatum*, figure, theme or trajectory in the various literature (cf. Levinson (1996) and Retz-Schmidt (1988)). The terms relatum and locatum will be used throughout this work, even in cases with non-spatial preposition uses.

However, apart from the spatial domain, these prepositions can occur in a wide range of senses (e.g., temporal, modal, causal) as well as in semantically transformed senses (e.g., metaphors and metonymies[1]). Existing practical approaches usually disregard semantic transformations or falsely classify them as spatial, although they represent the majority of cases[2]. The next section will detail this shortcoming of current approaches.

5.2 Related Work

5.2.1 Identifying the Research Gap

Applications which involve LEs in verbal interaction are still a major challenge for Artificial Intelligence (AI). These include, e.g., extracting spatial information from streaming data, robots following verbal wayfinding instructions, dialog-driven geo-services, emergency response/assistance systems, or querying search engines with spatial language. From a GIScience point of view, LEs represent a so-far largely untapped resource of spatial information. In particular with the rapid advent of social media platforms in the last decade, the amount of potentially valuable spatial information has been increasing by the second.
To utilize this information in intelligent systems, it needs to be correctly identified, extracted and modeled in a suitable machine readable format first. This task represents an essential step in the complex processing pipeline from unstructured spatial information to structured spatial knowledge and its potential iconic representations on sketch-maps (cf. Vasardani, Timpf, et al., 2013) or georeferenced maps. In approaches dealing with the actual interpretation of spatial relations in a geometric or geo-spatial sense (e.g. qualitative distances and directions), the extraction and disambiguation is assumed to be given (cf. M. Hall and Jones, 2012; M. Hall, Smart, et al., 2010; Lucas, 2012). However, the task is far from trivial and needs a dedicated approach based on methods from NLP, computer linguistic and machine learning.

[1] In contrast to metaphor, metonymy is a cognitive process mapping structures *within* one domain, and not *across* different domains (see Boers, 1996).
[2] Lakoff et al. (1980) demonstrated that metaphors are not just pervasive in language, but in thought and action and that our conceptual system is fundamentally metaphorical in nature. Herskovits (1986) later explicitly re-emphasizes that spatial metaphors pervade language; they are necessary to conceptualize various semantic domains, in particular abstract domains.

Spatial prepositions can occur in a multitude of different senses apart from their spatial usage. Although several approaches exist to automatically disambiguate prepositional senses (cf. Boonthum et al., 2006; Litkowski and Hargraves, 2007; O'Hara et al., 2003; Srikumar et al., 2013; Villanueva et al., 2013) the disambiguation of spatial prepositions from their extended uses in metaphors, metonymies, idioms and related figures of speech (e.g. [5.2], cf. Section 6.1.1), generally named semantic transformations (see Gärdenfors, 2014), has been so far largely disregarded.

[5.2] the thought in the back of my mind.

In the spatial extension of the linguistic ontology **G**eneralized **U**pper **M**odel (GUM) by Bateman et al. (2010), for instance, the "[C]lauses with idiomatic or metaphorical uses of spatial terms were not considered" for the inter-annotator agreement. Khan et al. (2013), however, discovered these cases of semantic transformations as the main reason for high false positive rates in the identification of spatial language utterances. Some advanced approaches for spatial information extraction from text, such as the SpatialML scheme from Mani et al. (2008) have a related goal of identifying and annotating spatial information in text, but focus on geographical and culturally-relevant landmarks, i.e. named places or toponyms instead of spatial relations. This is often called *geoparsing* (cf. Gelernter et al., 2013; Oliveira et al., 2015; Ritter, Clark, et al., 2011).

In general approaches to sense disambiguation, semantically transformed cases are often classified as spatial, in contrast to temporal, manner and other common senses. In Srikumar et al., 2013, the authors identify 32 distinct classes of prepositions, but still classify metaphoric cases as spatial amongst others. Furthermore, approaches to automatic spatial relation extraction often completely omit these cases (Kordjamshidi, Frasconi, et al., 2012; Kordjamshidi et al., 2010, 2011; H. Li et al., 2006; Mani et al., 2008; Pustejovsky et al., 2011, 2013; Shen et al., 2009; C. Zhang et al., 2009; X. Zhang et al., 2011). The main reason for omission often lies in the choice of corpora, for example by using corpora that are manually preselected to have a high rate of spatial language utterances (cf. Bateman et al., 2010). In an unrestricted natural language environment however, the number of expressions misleadingly identified as being spatial will be significant. This high rate of false positives has a direct negative impact on the accuracy, on the processing speed, and ultimately on the feasibility of an automatic system.

I am therefore aiming to provide an intelligible and fast approach to extract and disambiguate spatial from non-spatial uses of prepositions that are generated by any kind of semantic transformation.

5.2.2 Fundamental Literature on Spatial Prepositions

Due to the richness of approaches to the general topic of spatial prepositions, only selected views on spatial prepositions that influenced my approach are provided.

In semantic theory in general, two types of approaches can be identified: full specification and minimal specification. In full specification, every meaning of a word has a distinct representation in a lexicon. In minimal specification one meaning is seen as central, and from this all others are supposed to be derived by context or semantic transformations such as metaphors and metonymies (cf. Gärdenfors, 2014).
In terms of research on the semantics of prepositions, the minimal specification view is also called the localist view. The localists argue for the spatial sense being the central meaning of prepositions. The first to promote this view was Leibniz (1765), stating that prepositions "are all taken from space, distance and movement", and many others followed (e.g., Coventry et al., 2004; Herskovits, 1986; Landau et al., 1993; Miller and Johnson-Laird, 1976). Herskovits (1986) argued for an ideal meaning, and called any divergence from it a sense-shift, but still as based on the spatial meaning. However, later theories claimed that there is more to prepositional meaning, for example the notion of a control relation as in [5.3] (cf. Garrod and Sanford, 1989).

[5.3] John is in a bad mood.

Neuro-scientist and Nobel-Prize winner John O'Keefe described the non-spatial relationships as higher dimensional axes additional to the first dimensions of space and time and called them metaphorical (O'Keefe, 1996). Coventry et al. (2004) supports this view but states that such extended uses are direct extensions of the spatial meaning of the terms rather than novel metaphorical uses.

For an efficient GIScience approach, a rather practical common sense separation in spatial and non-spatial uses of prepositions is targeted. With the main purpose being to increase the accuracy and efficiency of language parsers to identify physically spatial LEs. Hence, I argue that for the scope of this thesis, the existence of a core meaning is not essential. Section 6.3.2 shows that the claimed core meaning, i.e. the spatial meaning, is not necessarily the most frequent one. However, the deep linguistic and cognitive analysis of the cited work (and many more) represents the basis of the current understanding of prepositions, and nurtured the idea for a feasible disambiguation approach.

5.3 Natural Language Processing

According to Bird et al. (2009), NLP, in a wide sense, covers any kind of computer manipulation of natural language, i.e. languages spoken by humans. This includes simple tasks such as comparing word frequencies as well as complex algorithms for "understanding" complete human utterances, at least to a useful extent, e.g. for automatic question answering.
In this section, typical methods in NLP are introduced that are employed for the extraction and disambiguation approach.

5.3.1 Part-Of-Speech Tagging

Usually the first step after tokenization in a typical NLP pipeline, is **P**art-**O**f-**S**peech (POS)-tagging. It is the process of classifying words into their parts of speech and labeling them accordingly (see Bird et al., 2009). The mentioned parts of speech are also referred to as word classes or lexical categories. The set of labels used for a specific task is known as a *tagset*.

Table 5.1: The 45 Penn Treebank tags for word classes in English with examples from Jurafsky et al. (2009).

Tag	Description	Example	Tag	Description	Example
CC	coordin. Conjunction	and, but, or	SYM	symbol	+,%, &
CD	cardinal number	one, two	TO	"to"	to
DT	determiner	a, the	UH	interjection	ah, oops
EX	existential 'there'	there	VB	verb base form	eat
FW	foreign word	mea culpa	VBD	verb past tense	ate
IN	preposition/sub-conj	of, in, by	VBG	verb gerund	eating
JJ	adjective	yellow	VBN	verb past participle	eaten
JJR	adj., comparative	bigger	VBP	verb non-3sg pres	eat
JJS	adj., superlative	wildest	VBZ	verb 3sg pres	eats
LS	list item marker	1, 2, One	WDT	wh-determiner	which, that
MD	modal	can, should	WP	wh-pronoun	what, who
NN	noun, sing. or mass	house	WP$	possessive wh-	whose
NNS	noun, plural	houses	WRB	wh-adverb	how, where
NNP	proper noun, sing.	IBM	$	dollar sign	$
NNPS	proper noun, plural	Carolinas	#	pound sign	#
PDT	predeterminer	all, both	"	left quote	' or "
POS	possessive ending	's	"	right quote	' or "
PRP	personal pronoun	I, you, he	(	left parenthesis	[, (, {, <
PRP$	possessive pronoun	your, one's	)	right parenthesis	],), }, >
RB	adverb	quickly, never	,	comma	,
RBR	adverb, comparative	faster	.	sentence-final punc	. ! ?
RBS	adverb, superlative	fastest	:	mid-sentence punc	: ; ... – -
RP	particle	up, off			

A widely used tagset is the so-called Penn Treebank tagset described in Santorini (1990) and developed at the University of Pennsylvania with 45 different tags. Table 5.1 shows the tags for all word classes in English accompanied by their names and examples from Jurafsky et al. (2009).

State-of-the-art POS-taggers for English achieve reliable accuracies of approximately 97 % according to Jurafsky et al. (2009) and Manning (2011), but have stagnated since 2003. Most of the taggers either use rule-based approaches or supervised learning techniques. The learning approaches usually employ **H**idden **M**arkov **M**odels (HMM), maximum entropy conditional sequence models, or other techniques like decision trees, that can deal with the sequential nature of the tagging problem (cf. Toutanova et al., 2003).

An example of a sentence with its POS-tags is depicted in Figure 5.1 visualized with the BRAT[3] visualization and annotation software described in Stenetorp et al. (2012). Here the problems that can arise with ungrammatical input are already visible. As the sentence completely misses any verb, the term *Road* at the beginning is falsely tagged as NNP instead of NN – most likely because it is capitalized. Nonetheless, these tags are highly valuable input features for learning approaches involving natural language.

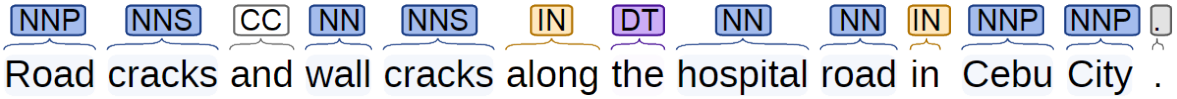

Figure 5.1: Result of POS-tagging for an English sentence.

Concerning social media input, standard taggers exhibit decreased accuracies as they are usually trained on well-formed corpora like news texts, for example from the Wall Street Journal. However, when specifically trained for mixed input, they can still reach between $85\,\%$ to $89\,\%$ accuracy according to Derczynski et al. (2013).

5.3.2 Named Entity Recognition

Another typical NLP tool that is used to benefit the approach is NER. It is the process of assigning a specific categorical label to mentions of so-called *named entities*. What exactly constitutes a named entity type is application dependent according to Jurafsky et al. (2009), but commonly includes people, places, and organizations. In other fields more specific entities might be of interest, e.g. the names of genes and proteins in NLP for bio-medical applications (Cohen et al., 2014).
For my purpose the common tags PERSON, TIME/DATE/DURATION, ORGANIZATION and LOCATION are sufficient. They will provide valuable hints for the disambiguation approach, to evaluate the involved entities in a LE. Figure 5.2 depicts the result of an named entity recognizer.

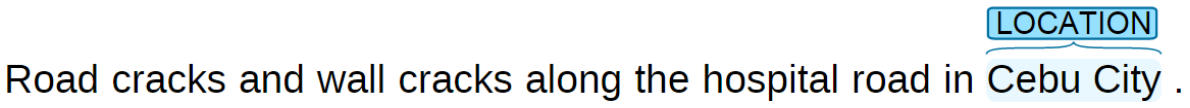

Figure 5.2: Result of NER for an English sentence.

5.3.3 Dependency Parsing

The probably most powerful method from the NLP toolbox that will be used is *dependency parsing*. It is the process of analyzing the grammatical structure of a sentence by establishing binary asymmetric relationships (semantic or syntactic) between so-called "head" words and their dependents, i.e. words which modify those heads (see Bird et al., 2009). According to Jurafsky et al. (2009), these relationships are often called lexical dependencies. In English, the tensed verb is commonly the head of a sentence, and all remaining words are either direct

[3] Available from `http://brat.nlplab.org/`.

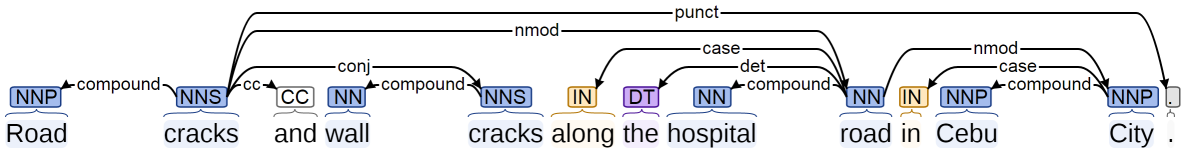

Figure 5.3: Result of a dependency parser for an English sentence.

dependents or are connected to the head via a path of dependencies. Consequently, the result of a dependency parser can be represented as labeled directed graph. The nodes then denote the lexical items, here words, and the labeled arrows denote the dependency relations from heads to dependents, i.e. the edges of the graph. Figure 5.3 shows an example result of a dependency parser using the *universal dependencies*[4] which mainly evolved from work of Marneffe et al. (2014), Petrov et al. (2012), and Zeman (2008).
The arrow from the word *City* to the word *in* indicates that *in* modifies *City*, and the label *case* assigned to the arrow classifies the relation as case-marking. Case-marking elements such as prepositions and postpositions, are regarded as dependents of the noun or clause they attach to or introduce (see Marneffe et al., 2014). In this work however, the dependency graph is used as an undirected graph, because the directions of the relations are not only irrelevant for the approach described here, but in fact limit the necessary connectivity of the nodes that I want to exploit.
In the following, a *relations path* refers to the sequence of relations that connects two words in this graph. For example the *relations path* between *along* and the first mention of *cracks* is given by

$$[\mathit{Start} \quad -\,] \quad \mathit{case} \quad - \quad \mathit{nmod} \quad [\,- \quad \mathit{End}\,]$$

where the bracketed parts are automatically added for further processing steps explained in Section 6.5.

For all three of the above described NLP tasks, the StanfordCoreNLP library is used in version 3.5.2. It is a toolkit by Manning, Surdeanu, et al. (2014) that offers, in addition to POS-tagging, NER and dependency parsing, also a high-quality lemmatizer and a syntactic parser for English. It is originally written in Java but here the Python wrapper stanford-corenlp-python from Dustin Smith is used[5].

5.3.4 Word Sense Disambiguation

The majority of words in natural languages have multiple possible meanings[6] such as the word *road*, which can mean *an open way (generally public) for travel or transportation* but also *a way or means to achieve something* as in *the road to fame*. Computers do not have the advantage of

[4] A detailed description of the currently 40 dependency relations can be found in the online documentation available at http://universaldependencies.org/u/dep/index.html.
[5] The library is available at https://github.com/dasmith/stanford-corenlp-python.
[6] This is commonly referred to as polysemy according to Jurafsky et al. (2009) and Villanueva et al. (2013).

a human's experience of the world and language. So assigning the correct sense to a word is challenging (cf. Banerjee et al., 2002). The automatic process of distinguishing these different senses from one another is called **W**ord **S**ense **D**isambiguation (WSD).
Many algorithms rely on machine readable dictionaries and exploit the shared vocabulary between the definitions of words. The prototypical approach for this idea has been introduced by Lesk (1986). Later, Banerjee et al. (2002) have optimized the approach by using WordNet instead of traditional dictionaries. WordNet is a lexical database of English, containing nouns, verbs, adjectives and adverbs, which are grouped into sets of cognitive synonyms, each expressing a distinct concept (see WordNet, 2015).
Moreover, WordNet offers a wide range of semantic hierarchies between concepts such as hyponymy and hypernymy. Concerning the term *road* for example, a hyponym would be *side road* – *side road* ***is a kind of*** *road* – and a hypernym would be *way* – *road* ***is a kind of*** *way*. The all-comprising concept in WordNet is the *entity* no matter where the hypernym hierarchy is entered. *Entity* has two hyponym concepts, which are *abstract entity* and *physical entity*, respectively given as *a general concept formed by extracting common features from specific examples*, and *an entity that has physical existence*. The hypernym path for the term *road* in its first sense given above, is depicted in the following list. When combining the output of the WSD and the hypernym hierarchy, a strong indicator can be given if a noun is a physical or abstract entity.

road, *route* – an open way (generally public) for travel or transportation
⇒ *way* – any artifact consisting of a road or path affording passage from one place to another
⇒ *artifact*, *artefact* – a man-made object taken as a whole
⇒ *whole*, *unit* – an assemblage of parts that is regarded as a single entity
⇒ *object*, *physical object* – a tangible and visible entity; an entity that can cast a shadow
⇒ *physical entity* – an entity that has physical existence
⇒ *entity* – that which is perceived or known or inferred to have its own distinct existence (living or nonliving)

6 Developed Extraction and Disambiguation Process

This chapter introduces the approach for efficiently extracting and disambiguating spatial information encoded as LEs from unrestricted natural language. As an additional task, the approach needs to account for the noisiness of social media texts. For this reason, the investigated corpus (cf. Section 6.3) comprises texts from different social media sources to represent a variety of modern language usage.
The main contributions of the approach are

- the applicability to *noisy* language such as social media texts,
- the coverage of all potentially spatial English prepositions,
- the automatic extraction of the locatum and relatum, and
- the capability of disambiguating spatial from non-spatial preposition usage.

The chapter is structured in six methodical sections and a short summary. First, I outline the scope of the approach, i.e. what exactly it aims to accomplish and which prepositions are investigated. Section 6.2 then details the conceptual decision schema based on the preceding section and introduces its three rules for a manual disambiguation of spatial and non-spatial uses of prepositions.
The mixed social media corpus used in this study is described in Section 6.3. Then, the last three sections explain how this schema can be implemented in code – i.e. the methods to teach a computer to get from a raw English utterance to a structured output of a LE. These are again filter steps to generate the desired output. Section 6.4 deals with the identification of LE candidates based on the existence of a preposition. Subsequently, Section 6.5 details the step of extracting triplets of the form

locatum — preposition — relatum

as essential preparation for the final process of disambiguating these triplets – that is classifying them as spatial or non-spatial (Section 6.6). In some cases, these triplets are missing the locatum. In particular in short messages the implicit *I* or *you* is sometimes omitted. These are so-called degenerated triplets as described by Khan et al. (2013)[1]. Here, these are explicitly excluded as non-spatial.

[1] Khan et al. (2013) us the term degenerated locative expression, however, as they can be non-locative, that is non-spatial, the name degenerated triplets is a better fit in the scope of this work.

Concerning the use case of natural disasters, these descriptions of spatial scenes can hold valuable on-site information. Extracting this kind of information efficiently from large, real-time social media streams is then essential to be able to contribute to the situational awareness in a disaster scenario. Still, I will rather focus on the scientific methods and algorithms than on the question how to further process the output in a real disaster or general event.

6.1 Scope

This section describes the scope and theory of the extraction and disambiguation approach. As there is no universally valid definition of what is really meant by *spatial*, a description is given to explain the notion of a spatial preposition in the scope of this thesis. Thereupon, general constraints on syntax and types of objects in consideration will be explained. Finally, this section is concluded by introducing the resulting choice of prepositions for this study.

6.1.1 Spatial vs. Non-Spatial

The notion of spatial in the approach is *locating in physical space*. Hence, spatial prepositions in LEs can be defined as:

> describing the location (e.g., inclusion, proximity) or movement (e.g., origin, path, endpoint) of physical entities [and events] relative to other physical entities, actual places or locations.

The definition provides a good basis as to which LEs should be detected and which should be excluded in this practical approach. The involved entities in the scope should support the notion of identifiability as well as being physically anchored. In small-scale environments entities should be accessible for direct or indirect physical interaction, often involving changing their location in space. Following the definition above, some types of expressions should explicitly be excluded as non-spatial. These expressions can incorporate spatial prepositions of which the sense has changed or shifted from their spatial meaning through semantic transformations. For humans, these non-spatial meanings are relatively easy to identify, due to their understanding of the involved entities and their world knowledge in general.

[6.1] The thought *in the back of* my mind.

In [6.1], it is generally understood that the relation *in the back of* is not a physically spatial one, i.e. the *thought* is not physically in the back of the *mind*. The locatum is an abstract entity and as such it can not be assigned to an explicit location in physical space. It is also known that thoughts *being in one's mind* is just the way people conceptualize this abstract relation because of its resemblance to the general spatial relation of an object in a container. These metaphors are denoted as *container metaphors* and are probably the most wide-spread in everyday usage, with

conceptualizations such as STATES ARE CONTAINERS as in [6.2], ACTIVITIES ARE CONTAINERS as in [6.3] or EVENTS ARE CONTAINERS as in [6.4], etc.

[6.2] We're *out of* trouble now.

[6.3] I put a lot of energy *into* dancing.

[6.4] Are you *in* the race on Sunday?

Other common metaphorical concepts which incorporate prepositions are so-called *orientational metaphors*, because they provide a spatial orientation to a non-spatial concept. The typical orientations are up-down, front-back, etc. (cf. Boers (1996) and Lakoff et al. (1980)) – e.g. HIGH (SOCIAL) STATUS IS UP and LOW (SOCIAL) STATUS IS DOWN as in [6.5] or FUTURE IS IN THE FRONT and PAST IS IN THE BACK as in [6.6].

[6.5] that would be *beneath* me.

[6.6] the weeks *ahead of* us.

Although locational uses of prepositions are often implicitly spatio-temporal, I only consider the spatial aspect in the current approach. Lakoff et al. (1980) notes that these major orientations seem to cut across all cultures, but the direction in which the concepts are oriented vary from culture to culture.

6.1.2 Syntactic Constraints

This subsection will explain the syntactic constraints on spatial prepositions with respect to the goals of the research in a practical manner. It is therefore not meant to provide a deep analysis of prepositional phrase structures in English and uses terminology typically found in computer linguistic publications and not in purely linguistically driven research.
Spatial relations can be encoded by other word categories such as adverbs (e.g. *here*, *downstairs*, *nearby*) and verbs (usually indicating a directed path, e.g., *to enter*, *to descend*, or implicitly describing a spatial arrangement, e.g., *to follow* or *to surround*), but the focus in this research is on the closed group of prepositions as indicators of spatial relations. The reasons for this are as follows:
(i) path-indicating verbs can in general be expressed by a simpler verb denoting movement (usually the manner) and a preposition providing the direction, such as to go *in(to)* instead of *to enter*; and (ii) adverbial terms lack an explicit relatum, and can usually not be decoded without discourse or context knowledge.

[6.7] I'm working *nearby*.

Example [6.7] is only fully understandable if a reference object is mentioned in the preceding discourse or is obvious in a specific situation – e.g., the listener knows the current location of the speaker and can assume the reference object equals the current location. However, discourse analysis and co-reference resolution spanning over the boundaries of a sentence has not yet been considered in this work.

In terms of syntax, an (optional) modifier (MOD), a preposition (P) and a complement (C) establish a prepositional phrase (PP). A PP denotes a single sentence constituent, which in general can not be separated. For an extensive analysis of syntactic and semantic cases of LEs that goes beyond the scope of this work, see Kracht (2002).

$$PP \Rightarrow (MOD) + P + C$$

It is important to distinguish between (transitive) prepositions and verb particles (i.e. intransitive prepositions). Verb particle constructions do not take a complement (e.g., *He blacked out*) or can be moved to the right of the following noun phrase (NP) as in *turn off the light* and *turn the light off*. Hence, they are not constituents, as the NP is a direct object of the verb and not of the preposition.
Verb particle constructions (VPCs) often form a semantic unit with the verb, where the particle does not carry its own semantic meaning and thus is not the head of a PP (Baldwin, Kordoni, et al., 2009), as in [6.8].

[6.8] We *looked up* the answer.

However, in [6.9] the word *up* is in fact the head of the following NP and therefore a (transitive) preposition.

[6.9] We looked *up* the street.

Prepositions can also take different types of complements such as participial verb phrases as in [6.10], sentences as in [6.11], NPs as in [6.12] or other PPs as in [6.13].

[6.10] John left *before eating dessert*.

[6.11] He was nervous before *the President called*.

[6.12] The book was placed on *the table*.

[6.13] She jumped *out from behind the tree*.

Only the latter two are of interest for this research because the former two can not describe spatial relations.
It follows that the present approach exclusively studies transitive prepositions, i.e., taking a NP as complement, and complex PPs, i.e., taking one or more PPs as complement where the last preposition has a NP complement. In these complex PPs, not every preposition will be recursively disambiguated, but rather assessed as one compound preposition that will get one class label. Thus, in [6.14], the compound preposition would be *from inside of*.

[6.14] The function was called *from inside of* itself.

6.1.3 Choice of Prepositions

In this research, an extensive list of English prepositions is studied that are typically considered to be potentially spatial. The list is compiled with the help of several English dictionaries (e.g.

Merriam-Webster, Inc., 2015; Oxford University Press, 2015) and the distinguished work done by Litkowski (2014) in the **P**attern **D**ictionary of **E**nglish **P**repositons (PDEP).
Prepositions that, to the author's best knowledge, do not (or not anymore) occur with a spatial sense in natural language, such as *as*, *because of*, *despite*, *during*, *for*, *in line with*, *in the face of*, *like*, *since*, *until*, *with*, and *without* are not investigated, i.e. they are directly excluded as non-spatial. I acknowledge that *until* might in some cases be interpreted as temporal and spatial, however, the temporal aspect is usually the more prominent one.
Additionally, there are a few English prepositions which can denote spatial relations but are archaic (e.g. *betwixt* and *nigh*) or domain specific (e.g. nautical terms such as *athwart* and *abeam*). They are also excluded because they are extremely rare in everyday language. The final list of prepositions considered in this study is presented in Table 6.1.

Table 6.1: Potentially spatial English prepositions considered in the scope of this work.

about	before	from	northwest of	southeast of
above	behind	in	of	southwest of
across	below	in (the) back of	off	through
after	beneath	in (the) front of	on	throughout
against	beside	in the middle of	on top of	to
ahead of	between	in the midst of	onto	toward
along(side)	beyond	inside (of)	opposite (of)	under
amid(st)	by	into	out (of)	underneath
among(st)	close to	left of	outside (of)	up
aside	down	near	over	upon
at	east of	next to	past	via
atop	far from	north of	right of	west of
back to	forth from	northeast of	south of	within

6.2 Manual Decision Schema

In this section the manual decision schema is presented covering the scope described in the preceding section. The schema comprises three rules which mainly help to identify non-spatial uses – i.e. based on these rules, utterances incorporating prepositions, which do not comply with the definition of *spatial* given in Section 6.1.1 are excluded. Each rule is explained with examples, drawn from the corpus of this study wherever possible.
Figure 6.1 shows a process diagram that is used as graphical representation of the schema. It depicts a very compact form of the schema and its three exclusion rules, and serves as quick reference for annotators (cf. Section 7.1).

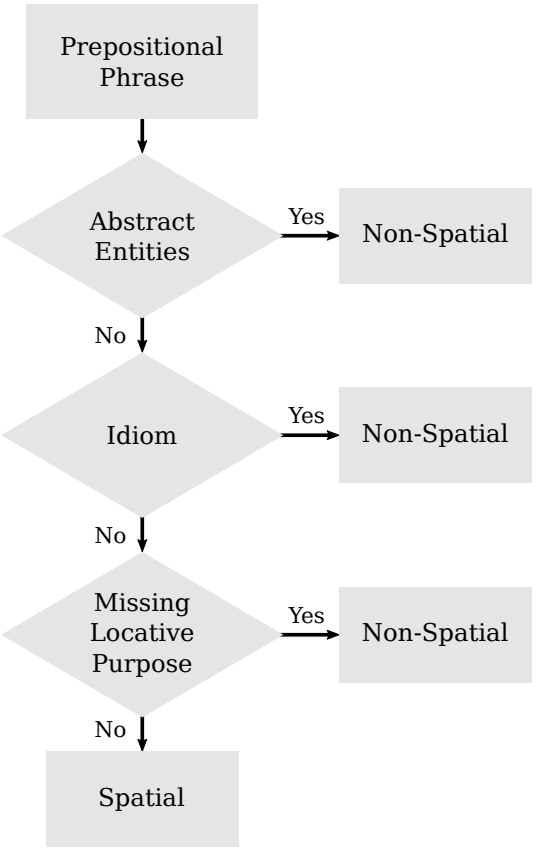

Figure 6.1: The manual decision schema depicted in a simplified form as workflow diagram.

6.2.1 Abstract Locatum or Abstract Relatum

If the potential LE has an abstract locatum **or** an abstract relatum it should be excluded as non-spatial, i.e. the locatum **and** relatum have to denote physical entities. Depending on the application domain, this rule can be relaxed to allow specific events as locatum, such as earthquake or fire.
Typical examples for abstract entities include emotions [6.15], ideologies [6.16], actions [6.17] and perceptual structures [6.18].

[6.15] [I']m already in *love* with someone else[...]

[6.16] [...]individuals can be drawn into world *capitalism* [...]

[6.17] I wish I was good at *singing*.

[6.18] We've settled into a *pattern*.

Terms that usually describe physical entities but are used in an abstract sense, count as abstract entities here and should be excluded as non-spatial as well.

[6.19] The author lures her reader into dark and dangerous *territory*.

The term *territory* in its most common sense can be described as a confined geographic area. In [6.19], however, it is used as an imaginary or abstract instance of its physical equivalent.

6.2.2 Idioms

If the potential locative expression itself is in fact a (frozen) idiom or if it is used idiomatically, the example should also be excluded as non-spatial. Idioms often contain potentially spatial prepositions, but utilize them in a non-spatial meaning.

[6.20] Peter *is over the hill*.

[6.21] She *felt under the weather.*

In [6.20], the preposition *over* does neither imply that the subject *Peter* is located nor that he is living over *the hill*, but rather that he is too old to accomplish something according to the *Cambridge Idioms Dictionary* (Walter et al., 2006). [6.21] also does not locate an entity *under the weather*, but describes a status of being ill, sick or intoxicated. The idiom is based on the parallelism of someone being un-healthy and the negative influence that weather change can have on someone's well-being.

6.2.3 Others Missing Locative Purpose

The last group of non-spatial expressions incorporating a preposition is a rather heterogeneous one. Here, all non-spatial examples are subsumed that are not "captured" by the previous two decision steps.

If the preposition used in the example does not locate the locatum relative to the relatum, it should be excluded as non-spatial. Examples for this rule include, but are not limited to, prepositions that denote a temporal relation, the material of an object [6.22], the agent of an action [6.23], or the topic of some means of communication [6.24].

[6.22] The paint is *made from* resin.

[6.23] He was paid *by* the customers.

[6.24] I read the paper *on* construction sites.

6.3 Social Media Corpus

In this study, a processed version of the mixed social media corpus is used that was originally compiled in Baldwin, Cook, et al. (2013) for the purpose of investigating and testing the common assumption of strong noisiness in textual data from social media sources. The original corpus comprised the following sources:

- **Twitter1 and Twitter2** — posts (tweets) from Twitter; 1M documents respectively
- **Comments** — user comments on Youtube; 874772 documents

- **Forums** — posts from the top-1000 valid vBulletin-based forums in the Big Boards forum ranking; 1M documents
- **Blogs** — blog posts from tier one of the ICWSM-2011 Spinn3r dataset (Burton et al., 2011); 1M documents
- **WIKIPEDIA** — text from an English WIKIPEDIA dump; 200K documents
- **British National Corpus** (BNC) — all documents from the written part of the BNC (cf. Burnard, 1995), a balanced corpus of British English used mainly as a point of comparison to the social media corpora; 3141 documents

The single entities in the corpus usually correspond to one sentence. In the following, however, they will be referred to as documents as they can comprise several sentences as well.
The authors further restricted the corpus to English documents by applying automatic language identification using LANGID.PY. For a deeper description of the original corpus and the processing tools used see Baldwin, Cook, et al. (2013). Liu et al. (2014) further sampled the corpus down to a selection of 100K random sentences from each source. Additionally, they extracted 500 sentences from each source for their hand-annotation. The raw string representation without the hand-annotation of the remaining 3500 sentences in total, eventually depict the base corpus investigated in this work for extracting and disambiguating spatial prepositions.

In the final corpus, all prepositions as well as their corresponding locatum and relatum were annotated. This was conducted by the author of this thesis in repeated discussion with two other experts, in particular Dr. Maria Vasardani (Department of Infrastructure Engineering, The University of Melbourne, Australia) and Prof. Timothy Baldwin (Department of Computing and Information Systems, The University of Melbourne, Australia).

6.3.1 Corpus Statistics

As described above, the investigated and annotated corpus consists of diverse sources that all have their own characteristics. Several statistics are provided in order to offer a rough idea of the respective similarities and differences.
Table 6.2 depicts the number of potential LEs – these are occurrences of potentially spatial prepositions – and the number of LEs annotated as spatial with the respective percentage, the number of different prepositions and the median word count per document for each source and in total. The sources with a higher median word count naturally also exhibit more potentially spatial prepositions, as well as a higher variety of prepositions. Nonetheless, the percentage of prepositions annotated as spatial is rather stable across the sources. Still, this illustrates the need for efficient filtering steps when processing high velocity data such as streaming data. Particularly when considering that this represents the percentage of LEs given that a potentially spatial preposition is provided.

Table 6.2: Different statistics for the corpus split into the different resources – the number of potential LEs (#LEs) and the number of LEs annotated as spatial (#spatial LEs) with the respective percentage (%spatial LEs), the number of different prepositions (#different prepositions) and the median word count per document (Median #Words).

	#LEs	#spatial LEs	%spatial LEs	#different prepositions	Median #Words
TWITTER 1	160	23	14.38%	24	9
TWITTER 2	153	27	17.65%	21	8
Comments	174	30	17.24%	23	9
Forums	273	51	18.68%	22	13
Blogs	423	87	20.57%	27	14
Wikipedia	681	150	22.03%	34	20
BNC	575	112	19.48%	34	17
Corpus	2439	480	19.68%	52	17

6.3.2 Preposition Statistics

Concerning the single prepositions that occurred in the whole corpus, interesting statistics can be observed. Figure 6.2 depicts the number of absolute occurrences for each preposition in the whole corpus, whereas Figure 6.3 shows the percentage of spatial uses of the respective prepositions.

The most frequent preposition *in* represents over 30% of all potentially spatial prepositions in the corpus. The 7 most common prepositions – *in*, *to*, *on*, *at*, *by*, *from*, *about* – account for more than 82%, and the first half of prepositions adds up to almost 98%. Interestingly, only 13 out of the 52 prepositions are more likely to occur in their spatial meaning than in a non-spatial one with respect to the corpus. Reversely follows that the others – that means 39 prepositions – are more likely to be used in non-spatial sense. Moreover, the really frequent prepositions exhibit low probabilities of being used spatially with respect to the corpus. The first preposition to exceed 50% is *near* with 7 spatial instances out of 10 in total.

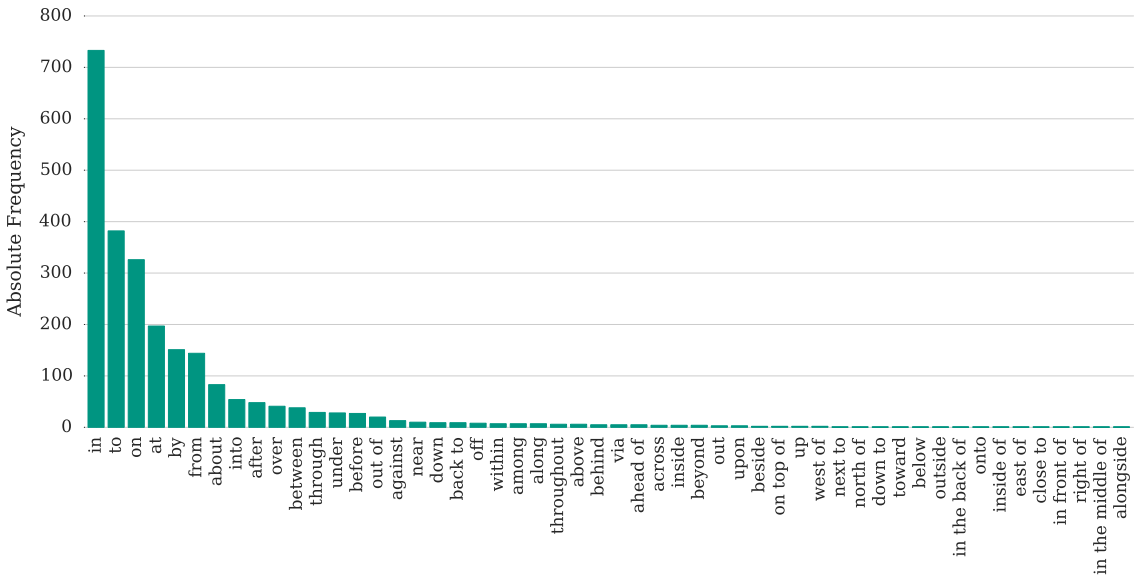

Figure 6.2: The absolute frequencies of all 52 prepositions that occur in the corpus.

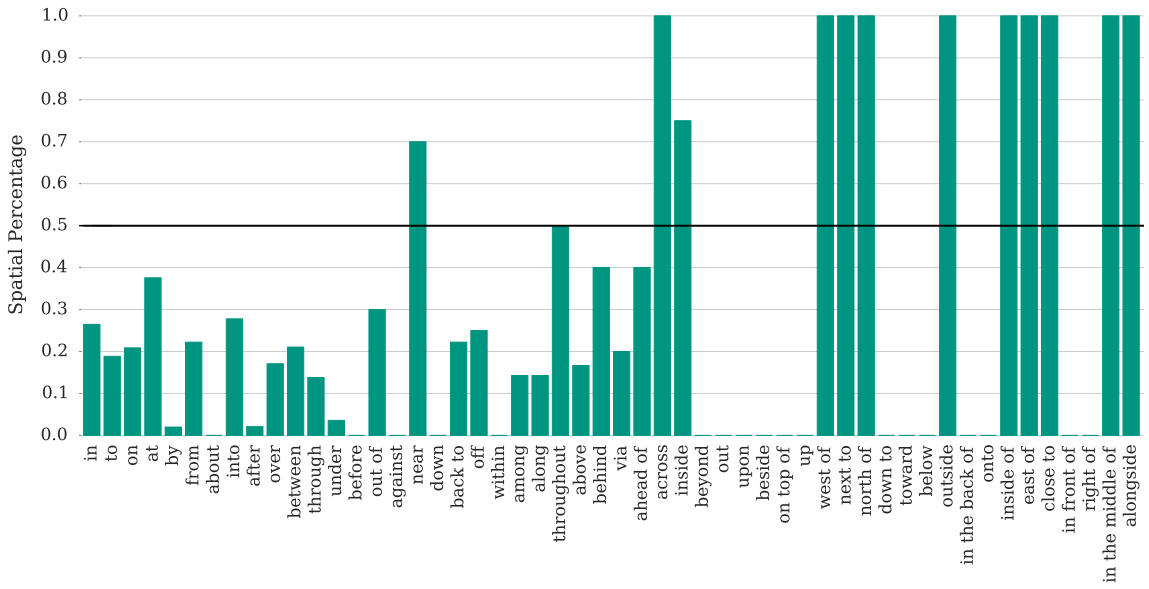

Figure 6.3: The percentage of instances annotated as spatial for each preposition in the corpus.

6.4 Prepositional Phrase Detection

The preceding sections established the conceptual foundation for an informed decision process to identify, extract and disambiguate LEs. On this basis, the first step of automating this process in a combined system using machine learning methods as well as fixed rules will now be explained.

The main purpose of this step is to identify possible candidate LEs based on the existence of at least one of the prepositions from Table 6.1 in a document. This is a valuable filter step to rapidly reduce the amount of data for the next, more costly, processing steps particularly when handling streaming data in real-time.

6.4.1 Approximate Regular Expressions

In many processing tasks of linguistic data, there is some basic pattern matching involved – that is finding a string or parts of it that equals some simple or complex pattern. The patterns are represented as so-called *regular expressions*, or often simply called *regexp* or *regex patterns* in software engineering.

With regex patterns (in the following highlighted in guillemets, i.e. «`pattern`») all textual substrings that start with an *i* and end in an *f* can be retrieved, e.g. the string *in front of*. More complex concepts can be expressed by combining the basic operators for boolean, grouping, and quantification matching. In the case of *in front of*, a simple pattern could just be the exact string, preceded and followed by a non-word character (e.g. whitespace) or underscore as in the pattern

«`[\W|_](in front of)[\W|_]`».

But what about the typical issues of the input source, e.g. the frequent misspellings? There might be documents where a simple typing error produces the substring *in frint of*, and possibly valuable information would already be missed. Now, the following pattern could account for that.

```
«[\W|_](in fr[i|o]nt of)[\W|_]»
```

It would match both spellings by incorporating the boolean *or* operator «|». But obviously such an approach is not the best way to handle arbitrary spelling mistakes, as every possible error would have to be modeled explicitly.
For this reason, the pattern constraints could be softened to match, for instance, three consecutive words with the starting letters *i*, *f* and *o*, followed by any alphabetical character and even constraining the respective lengths to ± 1 of the original word length, as in

```
«[\W|_](i[^\W\d]{0,2}[\W|_]f[^\W\d]{3,5}[\W|_]o[^\W\d]{0,2})[\W|_]».
```

However, not only would a lot of false positives be matched such as the substring *is fond of* in the sentence *He is fond of his job*, but also the pattern just accounts for in-word or end-of-word errors as the starting letters were fixed.

Thus, as powerful as regex patterns are, a possibility to handle arbitrary spelling mistakes is necessary, which still limits the results to close and plausible matches. I opted for an approach that combines regex patterns with the idea of spelling correction by edit distances that were already introduced in Section 3.4.2. The implemented method is based on the so-called BITAP algortihm by S. Wu et al. (1992)[2]. The algorithm employs bitmasks which represent each element of the pattern as one bit. The following search and approximate matching steps can then mostly be achieved by bitwise operations, thus making it the fastest solution for approximate string matching. The algorithm allows for an error-tolerant search based on the operations defined for the Levenshtein distance – that means deletion, insertion and substitution (Levenshtein, 1966). The transposition of two adjacent characters, as defined by the Damerau-Levenshtein distance is not yet implemented (Damerau, 1964).
Due to the shortness of the patterns – the potentially spatial prepositions – the results are limited to matches with a maximum of one error. Additionally, three exceptions have to be handled applying common sense heuristics.

1. The match is an existing English word.
   ⇒ match is not considered as preposition
2. The match has less than three letters.
   ⇒ match is not considered as preposition because of the high amount of other possibly correct words
3. The match is another potentially spatial preposition.
   ⇒ match is considered for the preposition that matches exactly

[2]The PYTHON bindings of the C library TRE from Ville Laurikari are employed, which are available at https://github.com/laurikari/tre.

So far the approach has not accounted for potential symbolic representations of prepositions such as @ for *at* or *2* for *to*, unless they are part of an expanded phrasal abbreviation.

6.4.2 Additional Pre-Processing

Following this filtering for documents including potentially spatial prepositions, some additional pre-processing is now applied to the extracted candidates by the approximate regex matching step.
Similar to the pre-processing steps employed in Section 3.4.2, the documents are tokenized to retrieve individual words for further analysis. In contrast to the IR approach taken for modeling the topic of an aggregated document (i.e. several messages), now more information than just frequencies of informative terms is needed. A small-scale view on single documents is taken and not a large-scale view on an aggregation of documents. As I now want to exploit the syntactic and semantic information encoded in the text, only really irrelevant terms should be deleted or ignored – i.e. no stop word removal is conducted. Although some words do not necessarily carry a lot semantic content in themselves they might still be important for acquiring insight into the structure of a sentence.
Mainly for the same reasons, neither any stemming algorithms nor any lemmatizers are applied to the documents before analyzing their structure. The inflections of words such as tenses or the voice provide valuable hints to the meaning of a sentence.
Nonetheless, the approach needs again, to some extent, consider the special characteristics of social media input. Accordingly, the same minor cleansing steps as described in Section 3.4.2 are applied, as well as the correction of obvious spelling mistakes. Moreover, also *phrasal abbreviations* are extended to their full form to optimize the input for the following POS-tagging and dependency parsing.

6.5 Triplet Extraction

The preceding sections showed how LE candidates are identified, and that some prepositions are more likely to occur in a spatial sense than others. However, before the spatial cases can be disambiguated from the non-spatial ones, two important features of the assumed LE need to be extracted. Those are the subject (potential locatum) and the object (potential relatum), which represent the involved entities and carry essential information for the disambiguation of the preposition.

The main input for the extraction is the dependency parser result, together with information on a per-token basis – that is the POS-tag and the potential NER output. Due to the already mentioned characteristics of social media input or complicated sentence structures, the output of the dependency parser can sometimes be wrong. In order to overcome this drawback I developed a rule-based approach to recover the correct relatum of the preposition, as well as an approach based on candidate selection and ranking to retrieve the correct locatum. Therefore, the graph structure of the parser output and the respective relation paths are exploited. As

described in Section 5.3.3, the inherent directions are ignored here, that means, the parser output is represented as undirected graph.

6.5.1 Identifying the Relatum

The easier of the two tasks is the identification of the object of the respective preposition, as it has an explicit representation in the dependency graph: the relation *case*. However, only if a parser was always yielding correct output. In the following, the *case* relation refers specifically to the one that (potentially) connects the respective preposition to another word, although there might be other *case* relations in the respective dependency graph.
Completely relying on the *case* relation of the preposition would mean that, if the *case* relation exists, the word connected to the preposition in this manner is taken as relatum, and if it does not exist the preposition is marked as having no relatum. Concerning the identified candidates (see Section 6.4.1) from the investigated corpus, this would generate 81.74% correct results in total, whereof 5.01% arise from cases with missing relatum. The remaining examples either have (i) a relatum but the *case* relation does not exist (0.64%) or have (ii) no relatum but the *case* relation exists (3.92%) or have (iii) a relatum and the *case* relation but the relatum is not the word directly connected to the preposition in this way (13.70%).
The first group exhibits a high heterogeneity and no patterns are perceptible. Fortunately, with respect to their minor number, the relatum extraction step can do without these cases. Concerning the second and third group, however, a few clear patterns can be observed. These patterns can be expressed as simple decision rules. The rather straightforward rule-based algorithm I developed, is given in 6.1. It is able to retrieve the correct relatum in 97.74% of the candidates over the whole corpus, hence increasing the performance quite significantly. Table 6.3 depicts the result split for the different resources represented in the corpus. The resources with generally longer sentences and a less noisy nature such as the WIKIPEDIA entries and the excerpts from the BNC performed slightly better than the rest.

Table 6.3: The accuracy of the relatum extraction for the different sources of the corpus.

TWITTER 1	TWITTER 2	Comments	Forums	Blogs	WIKIPEDIA	BNC
0.946	0.971	0.962	0.955	0.977	0.983	0.996

In the pseudo-code representation in Algorithm 6.1, the variable $case\_word$ refers to the word connected with the preposition via the *case* relation – that is case-marking. In contrast $nmod\_word$, $comp\_word$ and $conj\_word$ refer to the word connected with $case\_word$ via the relation *nmod*, *compound* or *conj*, respectively – these are denoting a nominal modifier, a compound word or a coordinating conjunction, respectively.

Algorithm 6.1. Extracting the relatum of a potentially spatial preposition.

```
1: s ← complete sentence containing the preposition
2: p ← preposition
3: pts ← possible POS-tags for relatum      ▷ these are NN, NNS, NNP, NNPS, PRP, WDT, WP, WRB
4: ners ← possible NER-tags for relatum ▷ these are PERSON, LOCATION, ORGANIZATION and TIME/DATE/DURATION
5:
6: function GETRELATUM(s, p, pts, ners)
7:     relations ← all dependency relations of the preposition
8:     if relations = {} then
9:         return {}                                                    ▷ no relatum
10:    end if
11:    if 'case' in relations then
12:        if CHECKPOS_NER (case_word, pts, ners) then
13:            return case_word
14:        else
15:            cw_rels ← all dependency relations of the case_word
16:            if 'compound' in cw_rels and CHECKPOS_NER (comp_word, pts, ners) then
17:                return comp_word
18:            else if 'conj' in cw_rels and CHECKPOS_NER (conj_word, pts, ners) then
19:                return conj_word
20:            else if 'nmod' in cw_rels and CHECKPOS_NER (nmod_word, pts, ners) then
21:                return nmod_word
22:            else
23:                return {}                                            ▷ no relatum
24:            end if
25:        end if
26:    else
27:        return {}                                                    ▷ no relatum
28:    end if
29: end function
30:
31: function CHECKPOS_NER(word, pts, ners)
32:    if word.pos_tag in pts or word.ner_tag in ners then
33:        return true
34:    else
35:        return false
36:    end if
37: end function
```

6.5.2 Identifying the Locatum

The more complex task is now the identification of the locatum of the preposition. The relations path that connects preposition and locatum in the dependency graph is generally longer and shows significantly more variation than for the relatum. Accordingly, an approach based on simple *if-then-else* decisions is not feasible. The solution I designed consists of three steps, which are detailed below.

1. Candidate selection
2. Candidate ranking
3. Threshold-based decision

Candidate Selection In order to be a potential candidate for the locatum, the words in the current sentence have to met at least one of the following necessary conditions.

- The POS-tag is one of NN, NNS, NNP, NNPS, PRP, WDT, WRB or WP.
- The NER-tag is one of PERSON, LOCATION, ORGANIZATION or TIME/DATE/DURATION

This corresponds to the possible POS-tags and NER-tags for relata in Algorithm 6.1. In addition, the relatum is excluded from the candidates. If the sentence does not contain any possible candidates, the locatum is classified as non-existent, i.e. as an degenerated triplet.

Candidate Ranking The identification of the most suitable candidate is conducted via a classifier that predicts the probability for each candidate to be the locatum. Only four features are sufficient to inform the learning of the classifier.

- The relative position of the candidate in the sentence with respect to the preposition as one of -1 or 1.
  $\rightarrow$ In general, the locatum appears before the preposition in the sentence, that means a relative position of -1.
- The absolute difference of the position indexes of the preposition and the candidate in the sentence, i.e. the positional distance.
  $\rightarrow$ In general, the locatum appears close to the preposition in the sentence.
- The combined absolute frequencies of the relation transitions of the relations path from preposition to candidate.
  $\rightarrow$ The single transition frequencies are derived for all occurring relation paths from preposition to locatum in the corpus. Here, a single transition refers to two specific relations directly following each other in a relation path, for example *case* — *nmod*. Eventually, the combined frequencies are calculated by simply summing the individual transition frequencies of the current relations path.
- The combined absolute frequencies of the relation transitions of the relations path from preposition to candidate, normalized by the relations path length.
  $\rightarrow$ The same as the absolute frequencies but length normalized. Accordingly, this feature indicates if a certain relation path is likely from a global perspective.

The classifier is trained with a randomized $80\,\%$ to $20\,\%$ training set to test set split. In addition, optimized parameters are derived based on an exhaustive grid search on the training set. Here, a random forest classifier is used as it is able to directly output class probabilities for the candidate ranking and allows for an inherent feature importance analysis. A random forest is a so-called meta estimator, because it fits several single decision tree classifiers on sub-samples

of the input. Then it uses averaging to optimize the accuracy of the prediction and to avoid over-fitting. The differences in feature importances do not necessitate a dedicated feature selection in this case (cf. Figure 6.4).

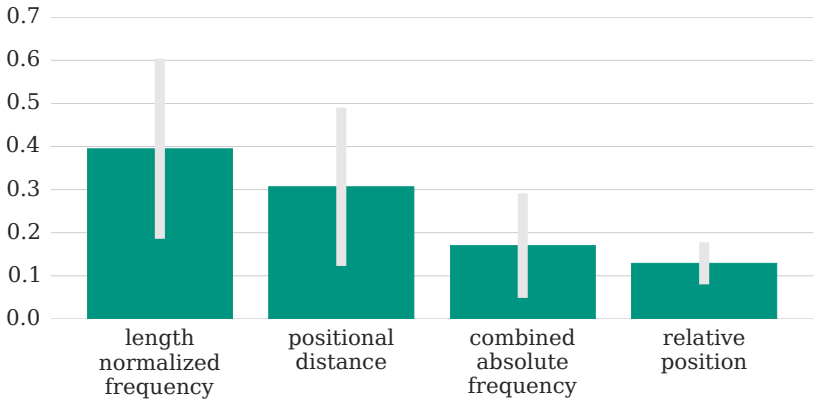

Figure 6.4: The feature importances for the locatum extraction. The so-called mean decrease impurity is used as importances as defined by Breiman et al. (1984). It is the total decrease in node impurity averaged over all trees of the forest. The impurity describes how many times an arbitrary selected item from the set is incorrectly labeled if it was randomly classified according to the subset class distribution. The importances are plotted with their inter-trees standard deviation.

Threshold-Based Decision Finally, the candidate with the highest probability as predicted by the classifier is labeled as the locatum. Yet, the example could be a degenerated triplet even if candidates exist. In order to take these cases into account, the probability of the best candidate has to met a certain threshold to be accepted. Based on a search optimization with respect to recalling an existing locatum, the threshold t was set to 0.21. The result of the search is shown in Figure 6.5 and the respective optimal accuracy of 89.98 for the whole corpus.

Table 6.4 depicts the result split for the different resources represented in the corpus. Here, the results for Comments and Forums fall a little behind the rest, but still exhibit good performance. Again, the sources WIKIPEDIA and BNC yield the best results.

Table 6.4: The accuracy of the locatum extraction for the different sources of the corpus.

TWITTER 1	TWITTER 2	Comments	Forums	Blogs	WIKIPEDIA	BNC
0.904	0.906	0.840	0.862	0.899	0.924	0.908

6.6 Semantic Disambiguation

In order to automatically disambiguate spatial from non-spatial triplets, discriminating features need to be identified. Mainly inspired by the corpus annotation and the three rules of the manual schema, different types of features are considered promising. Among them are general linguistic features and advanced linguistic features using an external knowledgebase or word lists. Together with simple graph-based features, these build the input for a binary classifier that predicts the category spatial or non-spatial for each extracted triplet.

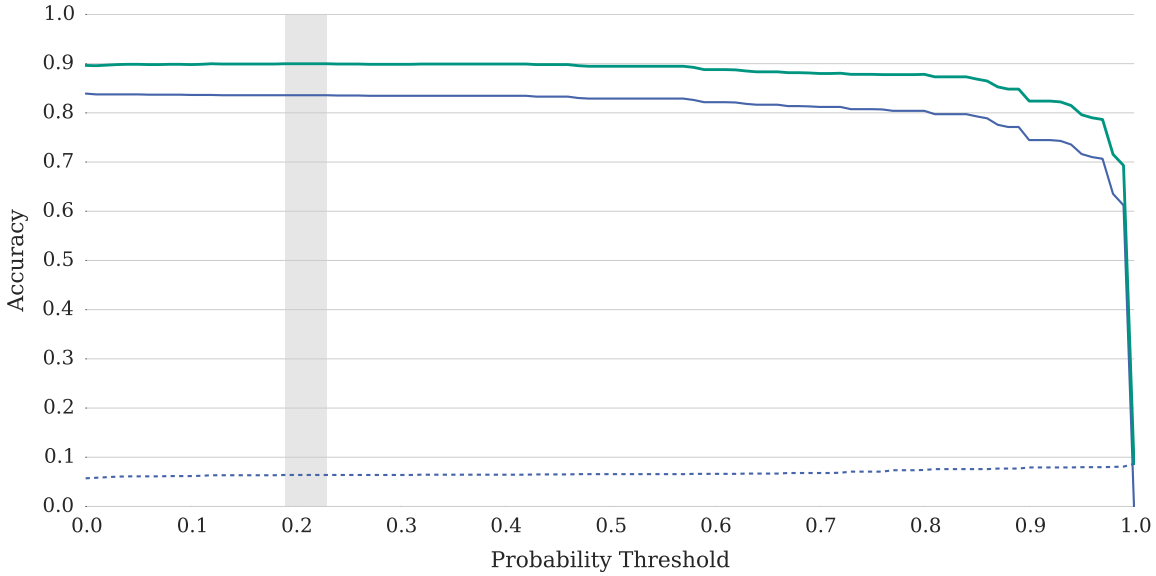

Figure 6.5: The graph shows the overall accuracies (green) of the extraction approach on the test set. The maximum accuracy as well as the maximum recall was reached throughout the range of $0.19 \leq t \leq 0.23$ (highlighted in light gray). In addition, the graph depicts the number of correctly identified locata, for the cases that have a locatum, with respect to the total size of the test set (blue solid line), and the equally normalized number of correctly identified degenerated triplets (blue dashed line). These values are of course inversely arranged and sum up to the total accuracy.

Still, some characteristics are so pervading for non-spatial examples that they can be applied as fixed exclusion rules before training the classifier and in some cases even before or in between the triplet extraction.

6.6.1 General Linguistic Features

The general linguistic features can again be subdivided in three main groups which are related to a certain word category. The following three paragraphs list the identified features in these groups and a short explanation were deemed necessary. These features can all be extracted based on the *StanfordCoreNLP* suite.

Verb-Related Features The features are extracted by the help of the output of the dependency parser, the POS-tagger, and the lemmatizer.

- the verb modifying the preposition
  → The verb that modifies the preposition is given by the dependency parsing as governor of the *nmod* relation.
- the verb's position in the sentence
- the verb lemma
- the used tense or inflection of the verb (encoded in the POS-tag)
- the used voice of the verb (as one of the labels *active* or *passive*)

Noun-Related Features The noun-related features are only referring to characteristics of the extracted locatum and relatum and not any other nouns present in the example. They are extracted based on the output of the POS-tagger, the NER and the dependency parser. The term noun in the following list refers to the locatum and relatum, respectively. That means the following features are extracted for both.

- the lemma of the noun
- the noun's position in the sentence
- if the noun is singular or plural (encoded in the POS-tag)
- if the noun can be identified as a location and consequently as a physical entity (encoded in the NER-tag)
  → The NER can generally identify locations down to city scale in all settings, and even very rural towns in rather well-formed sentences.
- if the noun can be identified as a person and consequently as a physical entity (encoded in the NER-tag)
  → The NER has a high precision and recall concerning the recognition of references to persons in text (0.93 and 0.95 respectively according to Atdag et al. (2013)). The performance slightly decreases when names are not capitalized.
- the accompanying adjective(s) of the noun, if existent
- the determiner of the noun, if existent
  → The possible values are e.g. articles (definite or indefinite), demonstrative and possessive pronouns (*this*, *those*, etc. and *my*, *his*, etc.), quantifiers (*many*, *some*, *a lot*, *most*, etc.)

Preposition-Related Features The preposition-related features refer only to the preposition of the extracted triplet and not any other preposition in the sentence. These are analyzed individually unless they directly follow each other and thus build a compound preposition such as *inside of*.

- the preposition itself
- the preposition's position in the sentence
- the modifier of the preposition, if existent
  → A modifier of a preposition is for example the term *right* in the utterance *I'm right in front of the house*.

6.6.2 Advanced Linguistic Features

Three advanced features were engineered in order to optimize the input information for the disambiguation process. The first two are motivated by the first rule of the manual decision schema (see Section 6.2). Accordingly, the features should indicate if the locatum and relatum

are likely to refer to physical entities. The last is based on the second rule, which excludes non-spatial idioms.

Personal Pronouns First, locatum and relatum are checked if they are masculine or feminine personal pronouns – no matter if used as subject or as object of the respective verb. These are *I*, *you*, *he*, *she*, *we*, *they*, and *me*, *him*, *her*, *us*, *them*. Consequently, they are marked as being a person and thus a physical entity.

Physical Entities Secondly, a WSD is applied to the identified locatum and relatum of the preposition. Here the adapted lesk algorithm was used as described in Banerjee et al. (2002)[3]. The output is the most likely sense of the term in question with a distinct identifier from the WordNet database. The identifier is used to access the hypernym hierarchy and follow its path up to the concept *entity*. If the concept *physical entity* is on the path of the investigated term – here the locatum or relatum – it is marked as such and used as input feature for the classifier.

Potential Idioms The identification of specific indicators for idioms is difficult, if not impossible to generalize into patterns. Therefore, a list of English idioms containing at least one of the prepositions in Table 6.1 was manually compiled from a specialized dictionary for English idioms by Götz et al. (2002) and the dedicated idiom website (IdiomSite, 2015). The different combinations of the lemmas of the preposition (P), the locatum (L), the relatum (R) and the verb (V) modifying the preposition are tested against this list and are marked as potential idiom when indicated.
One example for each of the possible combinations of V, L, P and R in the list is given in lemmatized form below.

- *feel under weather* ⇒ V – P – R
- *hit nail on head* ⇒ V – L – P – R
- *ace in hole* ⇒ L – P – R
- *out of blue* ⇒ P – R

6.6.3 Graph-Based Features

In addition to the different linguistic features, I also incorporated some simple graph based features to potentially improve the performance of the disambiguation. Such features are often not used in linguistic approaches as they do not provide much linguistic insights.

[3]The implementation by Liling Tan in the Python library pywsd was used available at https://github.com/alvations/pywsd.

- the number of nodes in the graph
  → That is not the same as the number of tokens as the dependency parser excludes punctuation.
- the number of edges in the graph
- the eccentricity of important nodes
  → The eccentricity of a node is its maximum distance to all other nodes in the graph. It is calculated for the locatum, the relatum, the preposition, and the verb modifying the preposition, if applicable respectively.
- the normalized eccentricity of important nodes
  → The normalized eccentricity of a node is its eccentricity divided by the number of edges in the graph.

6.6.4 Fixed Exclusion Rules

The manual investigation of the corpus and other linguistic data quickly revealed certain characteristics, which invariably occurred with non-spatial examples. These characteristics can be exploited to exclude a subgroup of non-spatial examples early on in the whole process.
The following three groups of rules can be applied at different steps of the complete extraction and disambiguation process. This will be mentioned at the end of the respective explanations.

No Relatum or No Locatum In the case that the extraction process yielded no relatum for the preposition, the example is immediately classified as non-spatial. This rule is applied after the relatum extraction. Analogously, examples identified as degenerated triplet are also excluded. This is applied after the locatum extraction.

Agent of Action The preposition *by* is frequently used in English to denote the agent of a passive verb. In terms of the dependency parser this relation is named *nmod:agent* and is assigned with high precision. The relation links the relatum as governor with the passive verb modifying the preposition as dependent.
A second possibility to detect agency constructions is the identification of the relatum as a person combined with the preposition *by* and the verb in passive voice. The relatum is marked as person either by the NER-tag PERSON (see Section 6.6.1) or if it is a specific feminine or masculine pronoun as explained in Section 6.6.2.
Both ways of detection can only be applied after the relatum extraction but before the locatum extraction.

Non-Spatial Word Collocation Another large group that can directly be excluded as non-spatial are typical word collocations including a preposition, which do not express any spatial relation. This comprises verb-preposition (V-P) collocations, adjective-preposition (A-P) collocations and preposition-noun (P-N) collocations. Lists were compiled manually for each group based on McIntosh et al. (2002) and English corpora analysis. Here, each group can be applied at different steps of the overall process depending on the respectively involved word categories.

Non-spatial V-P collocations build by far the largest group among the three with more than 180 items. Some non-spatial V-P collocations in the analyzed corpus are *cheat on*, *expect from*, *think about* and *yell at*. Communication verbs such as *say*, *talk*, *listen* and *write* are particularly common in non-spatial examples. This exclusion rule can already be applied before the triplet extraction.
The number of unambiguously non-spatial A-P and P-N collocations is far smaller but still beneficial for the disambiguation approach. Examples for A-P collocations in the corpus are *contrary to* and *inferior to*. These can also be applied before the triplet extraction. P-N collocations occurred in the form of *in fact*, *at least*, *in regard* and *at all*, and can be applied after the relatum identification.

6.7 Summary

In this chapter, the used methods for extracting spatial information from text in the form of LEs have been explained.
First, the scope of the developed approach and the resulting syntactic constraints and investigated prepositions have been detailed, and three rules for a manual decision schema have been derived. The investigated mixed social media corpus together with important statistics has been introduced. The main part of this chapter has detailed the developed automatic extraction and disambiguation approach with a particular focus on social media data. The process has been described along the three core steps of prepositional phrase detection, triplet extraction and the semantic disambiguation. The triplet extraction has been separated in the two steps of first extracting the relatum and subsequently extracting the locatum. The semantic disambiguation has focused on the engineering of relevant features for a classifier, as well as on the identification of fixed exclusion rules for non-spatial uses of prepositions.
In the next chapter, the methods will be evaluated based on an annotator agreement study and the performance of different machine learning algorithms.

7 Experimental Results

In this final chapter of Part II the manual decision schema as well as the implementation of the extraction and disambiguation process is evaluated. First, I detail the conducted annotator agreement study to show the feasibility and comprehensibility of the manual decision schema as well as the consistency of the annotation – what I want to demonstrate is that the definition of *spatial* given in Section 6.1.1 and the derived rules are plausible, also to non-experts. Then the numerical results will be presented and common mis-classifications will be discussed.
The topic of Section 7.2 is the evaluation of the automation of the complete extraction and disambiguation process, as well as in particular the disambiguation. Finally, a short discussion and summary conclude the chapter and also this second part of the thesis.

7.1 Annotator Agreement Study

In contrast to domains with numerical data, corpus linguistics usually lacks a classical mathematical definition with necessary and sufficient conditions for the categories of interest. Even some very basic categories such as the POS-tags (see Section 5.3.1) are not entirely unambiguously defined. The same applies for the definition of the category spatial introduced in Section 6.1.1. Still I want to generally quantify how well the categories can be delineated as well as how trustworthy the annotations are. To put it more precisely, I want to show that the definition of spatial given in Section 6.1.1 is comprehensible and plausible in the context of the goals of this thesis.
Thus, an inter-annotator agreement study was conducted on a subset of the mixed social media corpus with three annotators.
This section details the setup of the annotator agreement study, followed by the presentation of the numerical results and a discussion of common wrongly classified examples.

7.1.1 Study Setup

For the purpose of annotator agreement testing, the documents were filtered with the approximate regular expression method described in Section 6.4.1 – that means all examples for the human annotators included at least one preposition from Table 6.1. Neither the corresponding locatum nor relatum was highlighted in any way.
A random subset of 500 documents was generated for the annotator agreement test. The annotators were provided with the three decision steps (see Section 6.2) plus compact explanations and examples for each rule.
For the classification, a spreadsheet was provided that contained one document per row. The

prepositions that needed to be classified were completely in upper case letters. The successive columns were reserved for the annotation. The class labels were **1** for spatial and **0** for non-spatial. A document could contain several prepositions. In these cases, the annotators were advised to use one column for each preposition for the classification according to the order of appearance. These columns were labeled from P_1 to P_n with n being the number of prepositions to disambiguate in the respective example (see Table 7.1).

Table 7.1: Example document with multiple prepositions highlighted for the classification.

ID	Example	P_1	P_2	P_3
1	I'm standing IN FRONT OF the house NEXT TO my car. . . I'm ON time.	1	1	0

Additionally, the annotators were advised to always classify prepositions that are in upper case and directly following each other, as one compound preposition, as in Table 7.2. Due to the possibility of multiple preposition candidates in one document, the 500 documents for the study contained 804 prepositions for disambiguation.

Table 7.2: Example document with a compound preposition highlighted for the classification.

ID	Example	P_1
1	The rabbit came FROM INSIDE OF the hole.	1

7.1.2 Results

The manual classification was conducted independently by three annotators. One of the classifications was done by the author in repeated discussion with two other experts as described in Section 6.3. This will be referred to as the **Reference Annotation (RA)** which is sometimes called the gold standard in NLP and computer linguistic research. The **RA** yielded a prevalence of spatial instances of 22.76% for the random sample, which is slightly higher than the prevalence for the whole corpus (19.68%).
The remaining classifications will subsequently be called annotation A and B and were conducted by non-experts. These two annotators were given 50 simple examples before the actual annotation as a dry run. Thus, I could have responded to any general misunderstandings within the annotation procedure. The dry run yielded no evidence that changes to the procedure were necessary.

Evaluation Measures In order to comprehensively evaluate the outcome of the annotations, they were compared with the **RA** annotation. The evaluation is based on the number of conformities and nonconformities, which are typically distinguished as

- True Positives (TP)
  ⇒ the annotator makes a positive prediction (i.e. spatial), and the example has a positive result under **RA**
- False Positives (FP)
  ⇒ the annotator makes a positive prediction, and the example has a negative (i.e. non-spatial) result under the **RA**
- True Negatives (TN)
  ⇒ the annotator makes a negative prediction, and the example has a negative result under the **RA**
- False Negatives (FN)
  ⇒ the annotator makes a negative prediction, and the example has a positive result under the **RA**

From these basic values more advanced and meaningful measures can be derived. These are often denoted as

- *accuracy*
  ⇒ the fraction of all correct predictions of an annotator, i.e. that have the same annotation result under the **RA**, i.e.

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{7.1}$$

  In the given setting, this measure has to be interpreted with care as the class distribution is imbalanced.
- *precision*
  ⇒ the fraction of all positive predictions of an annotator that actually have a positive result under the **RA**, i.e.

$$precision = \frac{TP}{TP + FP} \tag{7.2}$$

- *recall*
  ⇒ the fraction of all positive results under the **RA** that have a positive prediction from an annotator, i.e.

$$recall = \frac{TP}{TP + FN} \tag{7.3}$$

- F_1*-score*
  ⇒ the harmonic mean of precision and recall, i.e.

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{7.4}$$

In addition to these typical "positive" evaluation measures for classification systems, the following correspondent measures are considered as well:

- *Negative Predictive Value (NPV)*
  ⇒ the fraction of all negative predictions of an annotator that actually have a negative result under the **RA**, i.e.
$$NPV = \frac{TN}{TN + FN} \tag{7.5}$$
- *specificity*
  ⇒ the fraction of all negative results under the **RA** that have a negative prediction from an annotator, i.e.
$$specificity = \frac{TN}{TN + FP} \tag{7.6}$$
- *negative agreement*
  ⇒ the harmonic mean of the negative predictive value and specificity, i.e.
$$negative\ agreement = 2 \cdot \frac{NPV \cdot specificity}{NPV + specificity} \tag{7.7}$$

Thus, I can account for the correct handling of negative examples as the corpus exhibits unbalanced class distributions. In order to summarize the results of the confusion matrix in one comprehensive measure, **M**atthews **C**orrelation **C**oefficient (MCC) of Matthews (1975) is used. The MCC commonly provides a very balanced evaluation of the prediction compared to other comprehensive measures (cf. Baldi et al., 2000). It is limited from -1 to 1 and formally given as
$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)\,(TP + FN)\,(TN + FP)\,(TN + FN)}} \tag{7.8}$$
Finally, the value of Fleiss's Kappa is calculated to measure the agreement between all three classifications without considering the special status of the **RA**. It is often called **I**nter-**A**nnotator **A**greement (IAA).

Agreement between Annotators and the RA The results of Annotation **A** and **B** are displayed in comparison to the **RA** as confusion matrices in Table 7.3.

Table 7.3: Outcome of Annotations **A** and **B** as confusion matrices vs. the **RA**.

		RA	
		spatial	non-spatial
A	spatial	159	25
	non-spatial	24	596

(a) **A** vs. **RA**

		RA	
		spatial	non-spatial
B	spatial	149	17
	non-spatial	34	604

(b) **B** vs. **RA**

From the confusion matrices the basic evaluation values for annotator **A** and **B**, respectively, can directly be extracted and are depicted in Table 7.4.

Based on these values, the advanced statistical measures can be calculated. These measures are summarized in Table 7.5 for both annotators.

Table 7.4: Outcome of Annotations **A** and **B** as list.

	A	**B**
TP	159	149
FP	25	17
FN	24	34
TN	596	604

Table 7.5: The statistical measures for the evaluation of the agreement of Annotation **A** and **B** vs. the **RA**.

Measure	**A**	**B**
Accuracy	0.939	0.937
Precision	0.864	0.898
Recall	0.869	0.814
F_1-Score	0.866	0.854
NPV	0.961	0.947
Specificity	0.960	0.973
Negative Agreement	0.961	0.959
MCC	0.827	0.815

Inter-Annotator Agreement For the IAA, all three annotations were taken into account without any weighting or preferential treatment. Fleiss' Kappa (κ_π) was used, as it allows the calculations of the agreement between more than two annotators in case of nominal data, taking into account the probability of agreement occurring by chance (cf. Fleiss, 1971). This leads to a conservative estimation of the IAA. The value range is dependent on the number of annotators ($r = 3$) and the number of classes ($q = 2$) but not on the number of examples to annotate ($n = 804$), i.e. the possible value range in this setup is -0.5 to 1. The value of κ_π was computed to 0.81 according to the formulae in Equation (7.9) from Gwet (2008).

$$\kappa_\pi = \frac{p_a - p_e}{1 - p_e}, \tag{7.9}$$

where

$$p_a = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{q} \frac{r_{ik}\,(r_{ik} - 1)}{r\,(r-1)} \tag{7.10}$$

and

$$p_e = \sum_{k=1}^{q} \pi_k^2 \quad with \quad \pi_k^2 = \frac{1}{n} \sum_{i=1}^{n} \frac{r_{ik}}{r} \tag{7.11}$$

In Equations (7.11) and (7.10), the term r_{ik} refers to the number of times example i is labeled as class k. The term p_a then denotes the overall agreement probability and p_e the probability of agreement due to chance.

7.1.3 Discussion

The results show that the delineation of the two categories of spatial and non-spatial is possible for the vast majority of linguistic examples. Especially, the "negative" measures are extremely

high, but also the "positive" measures are very good for linguistic data. Annotator **B** had more problems identifying all spatial examples, but annotated them with a higher precision than annotator **A**.
Concerning the inter-annotator agreement, the presented result is at the lower limit of an "almost perfect agreement" which ranges from 0.81 to 1 according to the guidelines provided in Landis et al. (1977).

Despite the considerable agreement among all annotations shown by the IAA and the agreement of Annotations **A** and **B** with the **RA**, there were still some cases where the annotators analyzed the utterances differently. These cases were in fact quite often ambiguous concerning the actual triplet, especially concerning the context of the utterance. This section first identifies systematic or recurring cases within the FP category followed by the FN category.

False Positives Two types of FP classifications occurred in the annotation experiment. The most common ones were the misinterpretations of actual abstract entities as physical ones (e.g. *project*, *demand*, *capital allocations*, *voice*, *cost*, *word*, etc.). Annotator **A** produced 21 wrong classifications of this type and Annotator **B** 12. However, often these examples included two entities that could arguably be taken as locatum. In [7.1] the physical entity (person) *you* and the abstract entity *difference* can be seen as being the subject of the preposition *on*. In a further study, this error source could be reduced by highlighting the complete potential LE (i.e. locatum + preposition + relatum) for the annotators instead of only the preposition.

[7.1] [...], I don't know if you'd notice a huge difference on the street.

False Negatives The FN classifications showed only one smaller group of common misinterpretation sources, but the majority of cases was very heterogeneous. The group consisted of cases where the annotators rejected a spatial example with a place as relatum. The rejected places often were either very large [7.2] or they were just not very common toponyms (or unfamiliar to the annotators [7.3]), and as such hard to classify as actual places. In general, Annotator **A** produced 4 wrong classifications of this type and Annotator **B** 16.

[7.2] it moves inside Mercury's orbit and [...]

[7.3] Chornovil , [...] in Lvov oblast [...]

7.2 Evaluation of the Automatic Extraction and Disambiguation

This section details the results of the developed automatic extraction and disambiguation approach based on the annotated corpus. A main focus is laid upon the results of the machine learning classifier based on the engineered features explained in Section 6.6.1 and Section 6.6.2.

The accuracy of the two triplet extraction steps for relatum and locatum were already presented in the respective sections (see Section 6.5.1 and Section 6.5.2).

7.2.1 Disambiguation Results

For the disambiguation process of the potential spatial triplets, five effective machine learning classifiers were trained with the hand-annotated documents of the **RA**. That means the training was conducted based on the correct locatum and relatum.
For each of the classifiers an extensive grid search for the best parameters based on the recall score for spatial instances was conducted in a 5-fold cross-validation for each parameter combination. The choice to optimize based on recall is motivated by one of the main ideas of this thesis – the identification of potentially valuable information. Consequently, some false positive results are favored, that means lower precision, over possibly missing information, that means lower recall.

The classifiers that were used are shortly introduced below:

- Support Vector Machine with linear kernel (Linear SVC)
  → Support Vector Machines are a so-called maximum-margin classifier that tries to separate the data points with a hyperplane in a way that the distance between this hyperplane and the closest data point is maximized.
- Random Forest Classifier (Random Forest)
  → see Section 6.5.2
- AdaBoost Classifier
  → AdaBoost is a meta-estimator that starts by fitting a classifier (here a simple Decision Tree) on the original dataset. Then, additional instances of the classifier are fitted on the same dataset but using adjusted weights for incorrectly classified instances. Accordingly, subsequent classifiers adapt better to difficult cases.
- Logistic Regression Classifier (Logit)
  → Logistic regression is in fact a linear model for classification. The probabilities describing the possible labels of one trial are modeled by a logistic function. It is also known as Logit model.
- Ridge Classifier (Ridge)
  → As the name implies, the Ridge Classifier uses ridge regression to address some of the issues of ordinary least squares by putting a penalty on the size of coefficients. It belongs to the group of generalized linear models.

The results shown in Table 7.6 are based on a randomized 5-fold cross-validation using the optimized parameters respectively on the dataset with all 1563 triplets identified by the **RA**.

In order to overcome the respective weaknesses of the single classifiers, I incorporated a voting classifier. It takes into account the output of all classifiers and applies a hard majority voting.

Table 7.6: Statistical measures for the evaluation of the automatic disambiguation with all features. The standard deviations are derived from the 5-fold cross-validation.

Measure	Linear SVC	Random Forest	AdaBoost	Logit	Ridge	Voting
Accuracy	0.91 ± 0.02	0.90 ± 0.01	0.90 ± 0.02	0.89 ± 0.01	$\mathbf{0.91 \pm 0.01}$	$\mathbf{0.91 \pm 0.01}$
Precision	0.86 ± 0.03	$\mathbf{0.86 \pm 0.02}$	0.83 ± 0.03	0.79 ± 0.01	0.86 ± 0.03	0.81 ± 0.04
Recall	0.85 ± 0.05	0.79 ± 0.02	0.85 ± 0.05	0.86 ± 0.03	0.85 ± 0.04	$\mathbf{0.89 \pm 0.03}$
F_1-Score	0.85 ± 0.03	0.83 ± 0.02	0.84 ± 0.04	0.83 ± 0.01	$\mathbf{0.85 \pm 0.02}$	$\mathbf{0.85 \pm 0.02}$
NPV	0.93 ± 0.02	0.92 ± 0.01	0.93 ± 0.02	0.94 ± 0.01	0.93 ± 0.02	$\mathbf{0.95 \pm 0.02}$
Specificity	$\mathbf{0.94 \pm 0.01}$	$\mathbf{0.94 \pm 0.01}$	0.92 ± 0.02	0.90 ± 0.01	$\mathbf{0.94 \pm 0.01}$	0.92 ± 0.02
Negative Agreement	$\mathbf{0.94 \pm 0.01}$	0.93 ± 0.01	0.92 ± 0.01	0.92 ± 0.00	$\mathbf{0.94 \pm 0.01}$	0.93 ± 0.01
MCC	0.78 ± 0.05	0.77 ± 0.01	0.77 ± 0.04	0.75 ± 0.02	$\mathbf{0.79 \pm 0.03}$	0.78 ± 0.03

Feature Selection A lot of features explained in Section 6.6 are of type string, that means categorical data. Except for decision trees and its derivatives such as random forests or extra trees, most classifiers can not handle categorical features directly. But the simple approach of encoding them as integer values introduces misleading information as it implies that the features have an inherent order, which is usually not the case.
Consequently, they need to be vectorized using the so-called one-hot-encoding or one-of-k encoding, where each feature with k different values is mapped to k different features. It can be thought of as a projection into k-dimensional space where all data points have the same distance from the origin, and are all equidistant. In practice this means that for example the original feature *preposition*, which has 52 different possible string values in the corpus, is encoded in 52 *preposition="preposition"* features. In case the preposition of a triplet is *at*, the new feature *preposition=at* has a value of 1 and all other preposition features have a value of 0 for this triplet.

In consequence this leads to a large but sparse feature space (here 8140 features), which is again sometimes negatively affecting the classifier performance in terms of speed but also concerning accuracy and other measures. Therefore, I performed a **R**ecursive **F**eature **E**limination (RFE) combined with a grid search to estimate the best number of features.
The RFE is trained on the initial set of features and assigns weights to each one. Then, features whose absolute weights are the smallest get removed from the current feature set. This is recursively repeated on the diminished set until the best number of features is eventually reached. This procedure yielded 3340 features as the best choice for the disambiguation problem.
The results shown in Table 7.7 are also based on a randomized 5-fold cross-validation using the optimized parameters respectively and the optimally reduced feature set.

A ranking of the 20 most important of the reduced features according to χ^2 scores is presented in Figure 7.1. In general, χ^2 scores express how likely it is that a feature is independent from the class labels, which means the features with a lower score are of minor importance for the classification. However, the absolute scores are less relevant for the interpretation but should rather be taken as relative importances. Thus, features are ranked with respect to their usefulness, and not to make strict assumptions about their statistical dependence or independence.

Table 7.7: Statistical measures for the evaluation of the automatic disambiguation on the optimally reduced feature set yielded by the RFE. The standard deviations are derived from the 5-fold cross-validation.

Measure	Linear SVC	Random Forest	AdaBoost	Logit	Ridge	Voting
Accuracy	0.91 ± 0.02	0.91 ± 0.01	0.90 ± 0.02	0.89 ± 0.01	0.91 ± 0.02	$\mathbf{0.92 \pm 0.01}$
Precision	0.86 ± 0.03	0.85 ± 0.02	0.83 ± 0.03	0.80 ± 0.01	$\mathbf{0.87 \pm 0.03}$	0.85 ± 0.03
Recall	0.86 ± 0.04	0.85 ± 0.04	0.85 ± 0.04	0.87 ± 0.03	0.84 ± 0.04	$\mathbf{0.88 \pm 0.02}$
F_1-Score	0.86 ± 0.03	0.86 ± 0.02	0.84 ± 0.02	0.83 ± 0.01	0.85 ± 0.03	$\mathbf{0.87 \pm 0.02}$
NPV	0.94 ± 0.02	0.93 ± 0.02	0.93 ± 0.02	0.94 ± 0.01	0.93 ± 0.02	$\mathbf{0.95 \pm 0.00}$
Specificity	$\mathbf{0.94 \pm 0.01}$	$\mathbf{0.94 \pm 0.01}$	0.92 ± 0.02	0.90 ± 0.01	0.94 ± 0.02	0.93 ± 0.02
Negative Agreement	$\mathbf{0.94 \pm 0.01}$	$\mathbf{0.94 \pm 0.01}$	0.92 ± 0.01	0.92 ± 0.00	$\mathbf{0.94 \pm 0.01}$	$\mathbf{0.94 \pm 0.01}$
MCC	0.79 ± 0.04	0.78 ± 0.04	0.76 ± 0.04	0.76 ± 0.02	0.79 ± 0.04	$\mathbf{0.81 \pm 0.02}$

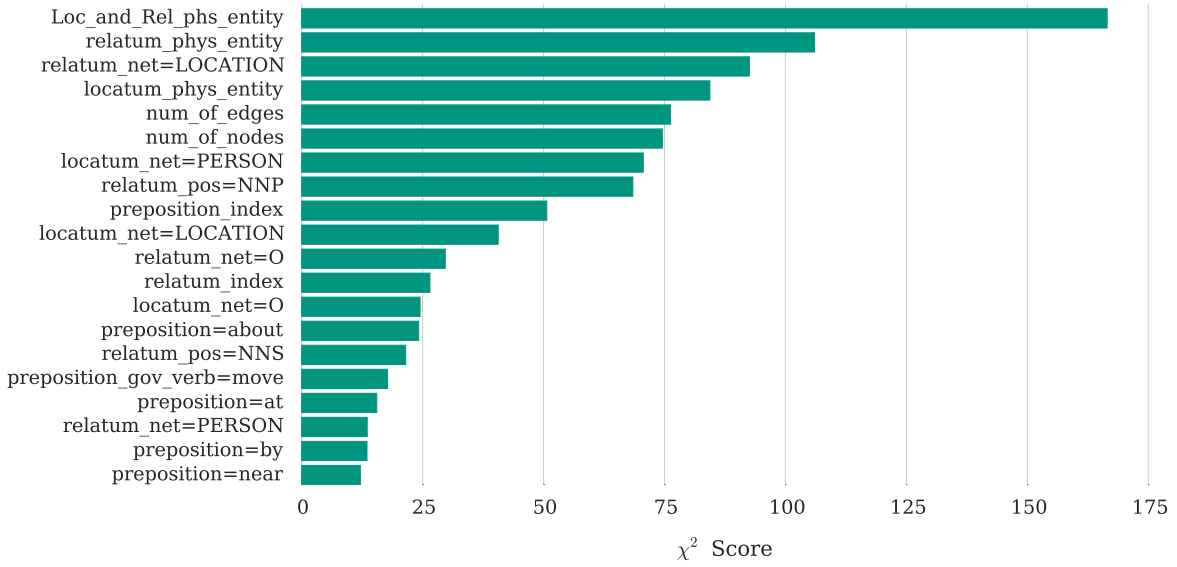

Figure 7.1: The ranking of the 20 most important features for the disambiguation process according to χ^2 scores.

As expected, the features relating to the locatum and relatum are of major importance for the classification with 11 out of the 20 most relevant features. Notable exceptions are the number of nodes and edges in the dependency graph, the specific verb *to move*, the index of the locatum and the relatum in the tokenized sentence, and the prepositions *at*, *about*, *by* and *near*.

7.2.2 Complete Workflow Results

In order to evaluate the extraction and disambiguation approach as a whole, the complete workflow shown in Figure 7.2 was applied to the 3500 short documents in the corpus.
This workflow depiction also includes the number of documents or triplets at each step, respectively. In addition, the typical time that the step needs to process one document or triplet, respectively, is stated.

Temporal Aspects The times depicted on the right side of each step in Figure 7.2 are measured on a laptop with an Intel Core i7-3517U processor (1.9 GHz, Dual Core) and 8GB of RAM running a 64Bit Ubuntu 14.04 LTS (Trusty Tahr).

Approximately 85% of the time is spent on the WSD for the locatum and relatum. Consequently, the whole process takes approximately 3.0 s to 3.2 s, which seems not to be fast enough to be applied directly to incoming messages with respect to TWITTER's velocity characteristics. But in a real-time system similar to the prototype introduced in Section 3.6, the workflow could be split into two parts where the latter is executed in a just-in-time fashion. That means that only the approximate regex matching is applied to incoming messages, which are then marked accordingly in the database. The cleansing steps and parts of the NLP pipeline are already performed for the event analysis. Thus, the rest would only be triggered in case of an actual event detection, that means at a point when the amount of messages is already reduced significantly. Concerning the prototype implementation, this task can even be parallelized for the remaining tweets as MONGODB enables document-level concurrency.

Overall and Source-Dependent Performance The final numbers for the class spatial in Figure 7.2 shows that, when applying the complete workflow to the whole corpus, it extracts and predicts 443 triplets as spatial. However, 480 spatial triplets actually exist in the corpus.
Out of the 443 predicted spatial triplets another 339 in fact are perfect predictions – that means correct locatum and relatum and correctly classified as spatial. When taking a strict, holistic view on the whole corpus of 3500 documents, these values can be interpreted as depicted in Table 7.8.

Table 7.8: Outcome for the complete workflow applied to all 3500 documents in the corpus.

	Complete Workflow
TP	339
FP	146
FN	104
TN	2916

Here, TP is the number of completely correctly identified triplets (339). The cases where the triplet was identified as spatial but had an error either concerning the locatum, the relatum or the disambiguation result, are here denoted as FP ($443 - 339 = 104$). These cases even might be correctly classified as spatial, but would produce false information concerning the involved entities in the spatial relation. The FN refers to the number of triplets that were wrongly identified as non-spatial ($480 - 339 = 141$). Eventually, the TN cases are all triplets that are correctly classified as non-spatial, no matter if it was out of the right reason. For example, a non-spatial but full triplet could be falsely identified as degenerated triplet by the locatum extraction, and thus is still correctly classified as non-spatial.
Eventually, this leads to the performance measures in Table 7.9.

The second and third exclusion steps in the complete workflow are subject to errors introduced by the relatum and locatum extraction respectively. The quantitative impact of cases where (i) the locatum or relatum could not be extracted or (ii) a locatum or relatum was misleadingly extracted, are incorporated in Figure 7.2 for the whole corpus. In contrast, Table 7.10 again splits the exclusion results and according disambiguation for the different sources.

Table 7.9: The statistical performance measures for the evaluation of the complete workflow applied to the whole corpus of 3500 documents.

Measure	Complete Workflow
Accuracy	0.930
Precision	0.765
Recall	0.706
F_1-Score	0.735
NPV	0.954
Specificity	0.966
Negative Agreement	0.960
MCC	0.695

Table 7.10: The quantitative results yielded by the complete workflow for the 2nd and 3rd exclusion step, as well as the disambiguation for the individual sources, compared to the respective results of the **RA**.

Source		2. Exclusion	3. Exclusion	Disambiguation	Correct Triplets
TWITTER 1	**RA**	125	108	23	
	Workflow	120	103	25	18
TWITTER 2	**RA**	117	100	27	
	Workflow	114	95	24	17
Comments	**RA**	131	114	30	
	Workflow	128	117	32	22
Forums	**RA**	194	189	51	
	Workflow	192	174	45	30
Blogs	**RA**	305	275	87	
	Workflow	299	246	78	56
WIKIPEDIA	**RA**	438	390	150	
	Workflow	438	380	137	107
BNC	**RA**	436	387	112	
	Workflow	437	368	102	89
Corpus	**RA**	1746	1563	480	
	Workflow	1728	1483	443	399

Eventually, the statistical measures can be calculated for the individual sources (see Table 7.11).

7.3 Discussion and Summary

7.3.1 Discussion

The evaluation of the results showed that the disambiguation process can achieve high quality results in terms of the typical statistical measures when trained with correctly labeled locatum and relatum. As expected however, none of the classifiers could completely reach the perfor-

Table 7.11: The statistical performance measures for the evaluation of the complete workflow applied to the individual sources of 500 documents each.

Measure	TWITTER 1	TWITTER 2	Comments	Forums	Blogs	WIKIPEDIA	BNC
Accuracy	0.98	0.97	0.96	0.93	0.89	0.85	0.93
Precision	0.72	0.71	0.69	0.67	0.72	0.78	0.87
Recall	0.78	0.63	0.73	0.59	0.64	0.71	0.79
F_1-Score	0.75	0.67	0.71	0.62	0.68	0.75	0.83
NPV	0.99	0.98	0.98	0.95	0.93	0.88	0.94
Specificity	0.99	0.99	0.98	0.97	0.95	0.91	0.97
Negative Agreement	0.99	0.98	0.98	0.96	0.94	0.90	0.95
MCC	0.74	0.65	0.69	0.59	0.62	0.64	0.79

mance of the human annotators. This is not surprising as the average annotator agreement is usually marking an upper bound in terms of plausibility for the automatic classification of linguistic data.

The idea of exploiting the output of all classifiers in a voting wrapper proved beneficial for the recall of spatial instances in particular, and for the NPV. For the other measures, except for the precision, the voting result reaches similar performance as the single classifiers for the non-constraint feature input (cf. Table 7.6).

The performance was pushed closer to the annotators' performance by an informed selection of the best features from the large and sparse feature space via an RFE approach. Again a final voting wrapper generally improved the results. It outperformed the single classifiers with respect to the overall accuracy as well as concerning the recall, the F_1-Score, the NPV and the MCC.

The feature ranking showed that the locatum and relatum related features are the most important group for the disambiguation.

A closer inspection of misclassified results, in particular false negatives, yields that a better word sense disambiguation would significantly improve the disambiguation result. However, WSD is still a largely open field of research and the problem is considered rather far from solved. Nonetheless, the WSD as an indicator for the disambiguation step is essential to the informed prediction of the trained classifier.

The evaluation of the complete workflow revealed that the performance suffers from the drawbacks of the sequential setup of the workflow – that means errors are propagated to the next step and negatively affect its performance. In particular, the difference between the "positive" measures (precison, recall and F_1-score) and the "negative" measures (NPV, specificity and negative agreement) is significant. This is not only due to the imbalance of the classes spatial and non-spatial in the corpus, but the definition I chose of what constitutes a result as TP, FP, FN or TN, respectively, for this holistic view of the complete workflow.

A TP needs to have all parts of the triplet correct – locatum, relatum and the classification as spatial. Whereas a TN only needs the correct assignment as non-spatial, whether it was excluded because no relatum could be identified (no matter if correctly or falsely) or it was identified as degenerated triplet (no matter if correctly or falsely), or it was simply identified as non-spatial by the final classifier based on its extracted features.

The performance evaluation for the individual sources concerning the complete workflow, reveal the same gap between positive and negative measures. Taking the MCC as a comprehensive measure of the overall performance, the BNC as "normal" text corpus exhibits the best results as expected. Concerning the other social media sources, the source TWITTER 1 yielded the best results in terms of recall and MCC, whereas the source Forums stays below 0.6 for both measures.

7.3.2 Summary

This chapter has evaluated the manual decision schema and the automatic extraction and disambiguation process. The conducted annotator agreement study has been described and the results confirmed the feasibility and comprehensibility of the manual decision schema, and consequently that the definition of *spatial* and the derived rules in this work are generally intelligible. The numerical results have been stated and common false classifications were discussed.
The automatic disambiguation process, as well as the complete workflow have been numerically evaluated and revealed very good and satisfying results, respectively. In addition the most relevant disambiguation features for the machine learning classifiers have been presented, as well as the results for the different individual sources in the corpus.

The chapter concludes the second part of this thesis "Spatial Information Extraction From Text" and thus also the methodological explanations of this thesis.

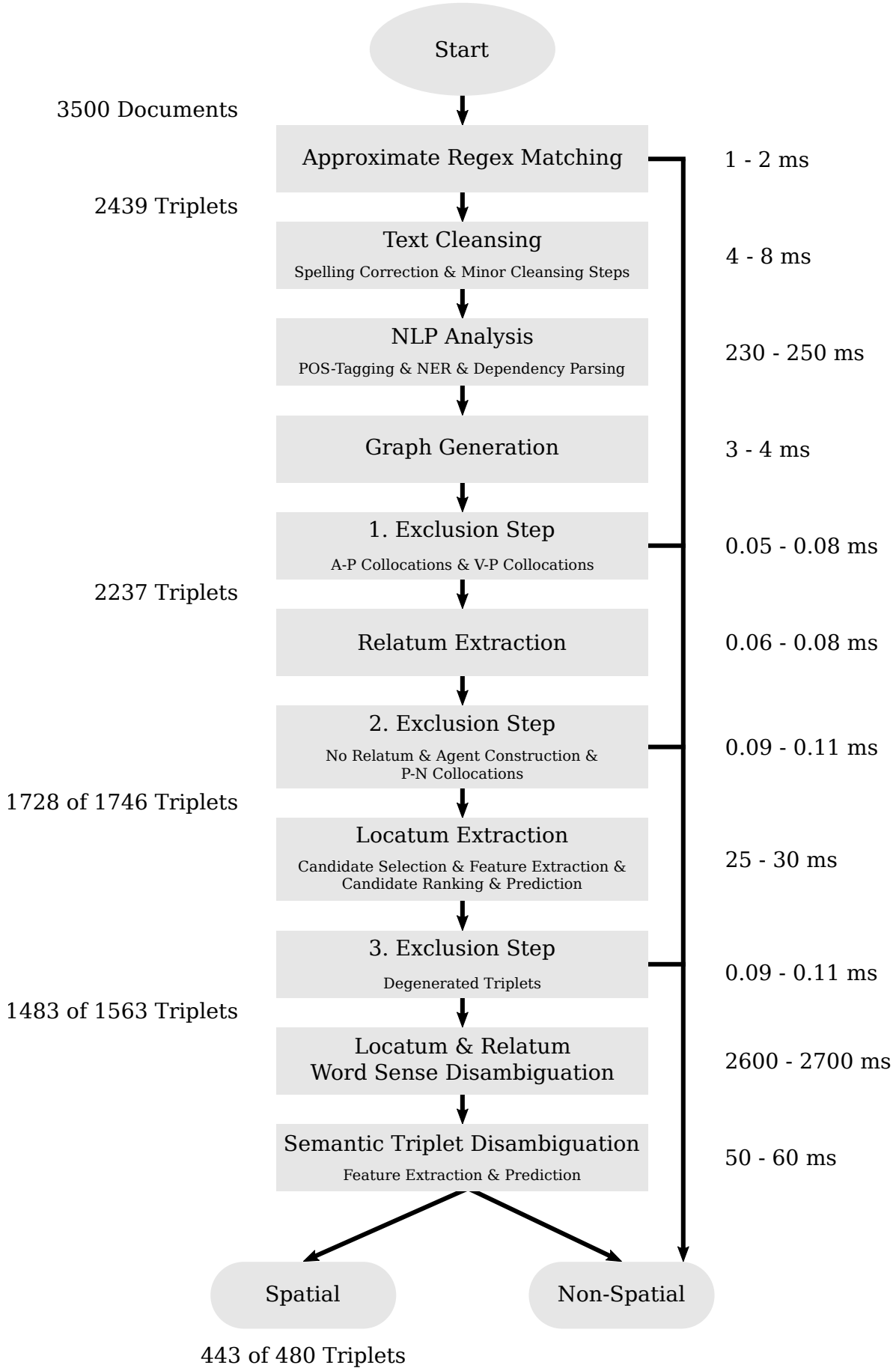

Figure 7.2: All steps of the extraction and disambiguation workflow together with the time each steps commonly needs for processing one document or triplet, respectively, and the number of triplets after the exclusion steps and the disambiguation.

Part III

SYNOPSIS

8 Conclusions and Outlook

In this third and last part, I recapitulate the goals of the research carried out, and summarize the respective approaches developed to achieve these goals. I will present the accomplishments as well as reveal current shortcomings and trade-offs. Moreover, I will point out possible optimizations and future research directions mainly based on the identified limitations of the current approaches.

8.1 Concluding Summary

Based on the potential of exploiting real-time data streams from social media platforms, two research goals were established for this thesis. The development of a fully automatic and operational framework for global-scale, real-time event analysis using social media data (Part I), and the development of methods to identify, disambiguate and extract spatial information, encoded as LEs from English social media text (Part II).
Several times in this work I used examples from the potential joint use case scenario natural disasters. Not only is the domain of natural disasters an adequate link between the two generally distinct research goals, but it also offers the possibility to highlight the respective strengths and also the potential benefit of the framework beyond the scientific community.

8.1.1 Conclusions and Outlook for Part I

As a basis the essential terms *event* and *real-time* have been defined in the context of this work. With respect to the input source of social media, the work has focused on events that affect a critical mass of people – here the users of a specific platform. This is why I have acknowledged the limits of the used data source in terms of spatial and temporal coverage, as well as availability, right from the beginning.

Important methods from the field of IR have been introduced. These enable to model the topic of textual documents and to determine the similarity between documents and topic models.
I have elaborated on current state-of-the-art approaches and have shown that these do not address the full range of real-time event analysis – that means they usually miss one or several of the following features: an operational prototype, multi-lingual coverage, proven real-time capabilities and/or a concrete classification of the event in the form of a class label.

A global-scale, real-time event analysis framework using social media data has been detailed that, despite the inherent drawbacks of the input data source, is feasible to overcome the

shortcomings of current state-of-the-art approaches.
A detailed investigation of TWITTER real-time data has yielded valuable characteristics especially concerning the spatial and temporal distribution of message volumes. The approach based on these findings has introduced a highly efficient grid-based spatial discretization method, capable of (i) capturing global-scale tweet volume statistics and (ii) processing the data in accordance with the specific real-time requirements.In order to achieve this, I have accepted the minor drawbacks of a equidistant grid to represent the surface of the earth, in particular the increasingly elongated shape and smaller areas towards the poles. Fortunately, a relevant core latitude range for the input source has been identified that limits these effects to a tolerable size.
Three more advanced discretization alternatives have been introduced and their respective pros and cons have been discussed. None of them is capable to provide the same storage and retrieval efficiency as the implemented grid, although the quadtree implementation comes close. In contrast, all of them increase the complexity concerning the implementation, the maintenance and possible future adaptations. Moreover, it is dubious if one of the alternatives would actually have improved the pure statistical detection capabilities. But this has still to be shown. However, such an extensive but specific comparison was out of scope for this thesis. In the future, extended investigations of the usage of a quadtree implementation to discretize space could overcome the manual decision for a certain grid resolution, as well as the reliance on a limited core area of data input.

A new way to optimize the temporal efficiency of the detection based on temporally overlapping windows instead of disjoint intervals has been illustrated. However, the absolute gain in speed as well as the applicability in general depends on the used input source and the performance of the implementation. For the prototype using TWITTER data, so far, time steps of just ten seconds have been realized with overlapping time intervals of one minute. It is very likely that for other social media platforms, these time constraints would need to be relaxed significantly. However, an approach exploiting several platforms in parallel could probably achieve similar temporal efficiency. Moreover, the data on other social media sources contains generally more detailed information and also of higher quality concerning relevant content. Leveraging the high velocity of TWITTER in combination with a subsequent querying of additional sources could substantially improve the information coverage for all sorts of events. However, such a system would need to overcome the complexity of fusing the information in a reasonable fashion.

A hierarchical domain model – a domain taxonomy – for the thematic classification of events has also been introduced. The taxonomy has been implemented as a module that is independent from the rest of the framework, and can be exchanged in a plug-and-play fashion to keep the classification approach generic. Within the taxonomy the different event types are modeled as BoWs and enable the calculation of similarity scores at adaptive granularities.
The classification process has been illustrated with examples from the use case scenario. I have demonstrated the approach of aggregating several short messages to one compound document based on the grid cells and time intervals. Additionally, the idea of dynamically compiled document collections has been introduced to represent a current baseline of typical topics in a

certain area. Thus, the collections are always up-to-date and the whole thematic classification approach stays quite generic. In terms of a final class label, I have favored the notion of a fuzzy membership in the form of similarity rankings, to account for the rather unspecific style in which social media messages are commonly written.
In order to meet the requirements of a global approach, several aspects to handle 64 different languages have been illustrated, e.g. the usage of specialized tokenizers for unsegmented languages, the integration of different dictionaries and stop word lists, and the implementation of a custom domain-restricted translation engine for keywords. In spite of all these efforts, the approach is optimized for the English language, due to the availability of suitable tools and resources, as well as of course the missing language proficiency of the author for most of the 64 languages (except for German, English, and to a lesser extent French).

In order to account for spatially widespread events, I have shown that *region growing*, a technique borrowed from the field of digital image processing, is capable of efficiently clustering event cells that have been yielded by the same real world event, based on the results of the preceding localization and thematic classification. The clustering also considers the ranking of different class similarities and not only the final class label.
The monitoring of temporally extended events has been accomplished by a rather simple database query that constrains the results by the impact area, a certain historic time range, and a criterion for the final class label. However, for the very rare case of ambiguous results, I have introduced a decision algorithm to retrieve the best match. So far this algorithm is based on common sense heuristics and subjective preferences, and could in the future be evaluated by empirical tests.

The necessary technical resources to implement an operational prototype that exhibits the explained capabilities have been stated and shortly introduced, together with its alert mechanism and ad hoc visualization features. As the prototype has been developed in parallel to the research for this thesis right from the start, it suffers from the typical issues of a scientific prototypical software. At the time of writing the framework comprises almost 8000 lines of code in a total of four different programming languages – JAVA (7000 lines of code), PYTHON (600 lines of code), MATLAB (150 lines of code) and JAVASCRIPT (200 lines of code). Although the main code in JAVA is highly modular, a complete redesign in one programming language (e.g. PYTHON) would not only reduce the verbose code base, but also the maintenance workload and at the same time improve the clarity and readability of the code. Moreover, except for ELASTICSEARCH, so far no dedicated tools for *Big Data* or streaming data handling is utilized. Instead, the framework is implemented from scratch in particular concerning the real-time complexities. In the course of a complete redesign, these very powerful and by now also quite flexible tools, such as HADOOP[1] or SPARK[2], should be incorporated to get the most out of the employed data sources and hardware.

[1] The Apache Hadoop library is an open-source framework that enables distributed processing of larg data sets on a single machine to thousands of clusters (cf. The Apache Software Foundation, 2016b).
[2] The Apache Spark library is a fast and general processing engine, which is capable of performing both batch processing and tasks like streaming, interactive queries, and machine learning (cf. The Apache Software Foundation, 2016a).

The conducted evaluation of the real-time event analysis framework has been based on earthquake ground truth data from the ANSS ComCat database. The system has been proven to work outstandingly well in regions with high TWITTER penetration and for earthquakes with potentially damaging impact, i.e. with a magnitude ≥ 5.0. The system detects earthquakes usually in less than ten minutes within reasonable and plausible distance from the epicenter locations. The median over all detections even is only slightly above four minutes. However, some shortcomings with respect to the data source also became apparent. In countries with a very low amount of daily georeferenced tweets with respect to the population, the detection rates drop significantly. Especially, China, Afghanistan, Papua New Guinea, Nepal, as well as parts of Russia and India exhibit detection rates below 50%.
Possible extensions here are the investigation of other domains than natural disasters that offer ground truth data for more than just one subtype. An interesting domain could be traffic accidents in urban areas. The ground truth information could be made available through cooperations with local police stations and traffic control units.

8.1.2 Conclusions and Outlook for Part II

In order to familiarize the reader with the idea of advanced processing and analyzing of linguistic data in general, and for the purpose of extracting and disambiguating textual spatial information in particular, I have started by introducing LEs as the most common form how people describe spatial scenes – i.e. how they express their spatial knowledge.
Subsequently, the research gap has been identified based on current approaches for prepositional sense disambiguation and spatial relation extraction. These currently disregard the need for a more fine-grained differentiation of the prepositional uses labeled as spatial. In particular, the physically spatial uses need to be disambiguated from semantically transformed uses such as in metaphors or metonymies.
Although a considerable amount of research has been conducted on spatial prepositions in the field of linguistics and cognitive sciences, these studies are rather concerned with all possible interpretations of prepositions and not the extraction and disambiguation. Moreover, in contrast to more recent and practical approaches, these fundamental studies often rely on introspective examples rather than on data from large real corpora. Thus, extensive papers analyzing only a single or a handful of prepositions are not uncommon (see Carlson-Radvansky et al., 1993; Coventry et al., 2004; Garrod, Ferrier, et al., 1999; Vasardani, Winter, et al., 2012). Hence, I acknowledge that their goals have been different from the ones I aimed for in the scope of this thesis.
Subsequently, important methods from the field of NLP have been introduced. Especially, dependency parsing and WSD are essential to my approach of automatically retrieving the relatum and locatum based on the identified preposition, and for the following disambiguation process.
As stated before, the approach is limited to English utterances, as the structure of spatial relations between entities is rather heterogeneous across different languages. A rather bold but nonetheless resourceful approach would be to overcome the language constraint by leveraging a general purpose machine translation tool. For example by translating incoming tweets,

written in other languages, into English and then perform the developed extraction and disambiguation workflow, the amount of valuable spatial information in case of events could possibly be expanded.

Based on the introductory chapter the developed extraction and disambiguation process, which is capable of handling the typical noisiness of social media text to a large extent, has been detailed.
The definition of what makes a preposition spatial in the context of this research has been elaborated on with respect to the core notion of locating in physical space, mainly in contrast to semantically transformed uses. Thereupon I derived syntactic constraints as well as a certain choice of prepositions for the investigations.
Hence, the approach has been restricted to the word category of prepositions, although textual spatial information in general can be encoded in other ways as well (e.g. adverbs, verbs and nouns). However, several authors have emphasized the predominant use of prepositions to denote spatial relations in English. Nonetheless, the identification of place names or toponyms (i.e. geoparsing) as a subcategory of NER, has also been a very active field in IR and GIScience research for quite some years already. Thus, I have included the output of a high-quality, general purpose named entity recognizer to enrich the input features for the learning approach with the possible tags of PERSON, TIME/DATE/DURATION, ORGANIZATION and LOCATION.

A compact manual decision schema has been introduced that allows human operators to annotate examples according to the definition in a straightforward fashion. The three rules of the schema mainly aim for a complete exclusion of all non-spatial use cases of a preposition. However, they are not entirely distinct in their targets but overlap, that means there are examples that would be excluded based on more than just one rule. The conducted annotator agreement study showed that the schema – and consequently the definition of spatial in this work – is generally understandable also for non-experts and allows to delineate the two classes. Common misinterpretations arose for example from ambiguous locata or relata, or unfamiliar place names.

The three main steps for extracting and disambiguating LEs in natural language input in general, and social media in particular, have been detailed and evaluated.
The first step is the detection of prepositional phrases of interest, which contain at least one of the investigated prepositions. The shortcomings of simple regular expressions matching particularly concerning the social media context have been explained. Hence, approximate regex patterns, which combine regular expressions with the idea of an edit distance between two strings, were introduced to handle the common misspellings in the input source with respect to prepositions. Nonetheless, minor decisions have had to be made based on common sense heuristics in unclear cases.
The identification of the locatum and relatum of the respective preposition was described as a necessary requirement for the disambiguation step. The extraction of the relatum has been possible with high accuracy by applying a straightforward rule-based algorithm based on the dependency parser output. In contrast, the extraction of the locatum has required a more complex approach. A classifier was trained to assign probabilities to identified candidates and

an optimized threshold was used to also account for degenerated triplets. Again the most valuable input has been shown to come from the output of the dependency parser, in particular the length normalized transition frequency of dependency relations. However, the achieved results could not reach the accuracy of the relatum extraction.

The disambiguation approach almost reached the performance of a human operator given the correct locatum and relatum. It combines the output of five effective machine learning classifiers in a majority voting wrapper. Additionally, an informed feature selection approach not only identified the most relevant features, but also improved the performance of the single classifiers as well as for the voting wrapper. As expected, the features related to the locatum and relatum have been proven to be essential to the disambiguation. However, the performance of the whole extraction and disambiguation process suffers from the drawbacks and the consequential "error propagation" of the sequential setup of the workflow. But latest developments in dependency parsing[3] show significantly improved accuracy for relation assignment between terms in a sentence. Due to the time frame of this work only single tests could be conducted, these however show very promising impacts on the performance of the locatum and relatum extraction.

In general, I have been aiming for a practical solution to the problem of automatically disambiguating spatial from non-spatial uses of prepositions. Hence, I am not claiming that a holistic linguistic analysis of prepositional senses has been conducted. Consequently, I recognize that in this approach, different aspects are disregarded such as certain meta-operators. These meta-operators include, for example, negations and aspects of existentiality, thus the utterances [8.1] and [8.2] would be classified as spatial. One could of course argue that in [8.1] the actual relation that should be extracted is *pencil outside of box* rather than *pencil in box*. However, taking the "inverse" preposition to handle negations can not account for the many different spatial configurations a single preposition can describe and is left to works dedicated to deep linguistic research.

[8.1] The pencil is *not* in the box.

[8.2] The pencil *was/might be/will be* in the box.

8.1.3 Synthesis Potentials

Finally, the joint implementation of both workflows – event analysis and spatial information extraction – has a lot of potential for future developments, in particular for the natural disaster use case.
In case of large-scale events in urban areas, the compiled context knowledge of the real-time event analysis could trigger subsequent search queries for relevant information in other social media sources. In addition, the identified impact area allows for disambiguating the specific location of extracted LEs by significantly reducing the search space in case of local street names and points of interest. Moreover, the involved increased amount of spatial relation descriptions

[3] The StanfordCoreNLP suite is now available in version 3.6.0 at http://stanfordnlp.github.io/CoreNLP/.

could lead to several messages reporting the same situation from different viewpoints. If it was possible to assign these complementing messages to one another, they could ultimately be used to construct spatial scenes and could consequently be tested for plausibility against each other, as well as to increase their reliability for real-word applicability.

Eventually, the probably most exciting and challenging possibility to extend the combined workflow is the usage of an additional data layer – the visual information provided in the form of pictures from an event. Social media messages are frequently accompanied by on-site pictures, in particular during disaster events. These provide a visual impression of the situation and with that are able to convey a level of insight, which can not be provided by text alone. A combined real-time analysis of the textual description and visual representation of an event site, where both data types mutually benefit from each other's information, would mean a major step forward for situational awareness during disaster events.

Abbreviations

RA Reference Annotation 122–124, 126, 127, 131

idf inverse document frequency 17

tf term frequency 16, 17, 62

A-GPS Assisted GPS 33

AI Artificial Intelligence 92

ANSS ComCat Advanced National Seismic System Comprehensive Catalog . . 77, 81, 140

API Application Programming Interface 9, 28, 30, 61, 77

BNC British National Corpus 106, 107, 111, 114, 131–133

BoW Bag of Words 16, 17, 20, 22, 55, 56, 61–65, 138

CDF cumulative distribution function 48, 51

CEDIM CEnter for DIsaster Management and Risk Reduction Technology 12, 73, 151

DARPA Defense Advanced Research Projects Agency 14

DBSCAN Density-Based Spatial Clustering of Applications with Noise 22

ENS Earthquake Notification Service 83, 85, 86, 166

FN False Negatives 123, 125, 126, 130, 132

FP False Positives 123, 125, 126, 130, 132

GeoJSON Geographical JSON 32, 162

GIF Graphics Interchange Format 2

GIScience Geographic Information Science 91, 92, 94, 141

GNSS Global Navigation Satellite System 33, 53

GPS Global Positioning System 33

GUM Generalized Upper Model 93

HMM Hidden Markov Model 95

HTM Hierarchical Triangular Mesh 23, 24, 52, 54

HTML HyperText Markup Language 59

IAA Inter-Annotator Agreement 124–126

IANA Internet Assigned Numbers Authority 47, 78

IE Information Extraction 30

IMD Indian Meteorological Service 87

IQR interquartile range 83, 84

IR Information Retrieval 5, 9, 14, 15, 17, 30, 43, 62, 63, 110, 137, 141

JSON Java Script Object Notation 28, 29, 32, 56, 57, 73, 162

LE Locative Expression . 5, 91, 92, 94, 96, 99, 100, 102, 104, 106–108, 110, 119, 126, 137, 140–142

MCC Matthews Correlation Coefficient 124, 125, 128, 129, 131–133

MMI Modified Mercalli Intensity 78, 79, 83

NER Named Entity Recognition 30, 96, 97, 110, 112, 113, 116, 118, 141, 166

NLP Natural Language Processing 5, 18, 29, 30, 57, 91, 92, 94–97, 122, 130, 140

NPV Negative Predictive Value 124, 125, 128, 129, 131, 132

OSM Open Street Map 74, 75, 78

PDEP Pattern Dictionary of English Prepositons 103

PGA Peak Ground Acceleration 79

POS Part-Of-Speech 95–97, 110, 112, 113, 115, 116, 121, 166

PST Pacific Standard Time 18, 37–40

RFE Recursive Feature Elimination 128, 129, 132

SQL Structured Query Language 73

TDT Topic Detection and Tracking 15

TENAS Twitter Event Notification and Analysis Service 71–74, 85

TN True Negatives 123, 125, 130, 132

TP True Positives 123, 125, 130, 132

UGC user-generated content 1, 34

UNISDR United Nations International Strategy for Disaster Reduction 11, 159

URL Uniform Resource Locator 2, 18, 32, 59, 74, 78

USGS United States Geological Survey 69, 74, 83, 85, 86

UTC Coordinated Universal Time 18, 29, 30, 37, 46, 47, 53, 69, 74, 79

VGI Volunteered Geographic Information 3

VSM **V**ector **S**pace **M**odel 15–17, 20, 21, 62, 63

WGS84 **W**orld **G**eodetic **S**ystem 19**84** 29, 32, 54

WSD **W**ord **S**ense **D**isambiguation 98, 117, 130, 132, 140

ZIP **z**ero-**i**nflated **P**oisson 50, 51

References

Aggarwal, Charu C. and Karthik Subbian (2012). “Event Detection in Social Streams”. In: *Proceedings of the 2012 SIAM International Conference on Data Mining*, pp. 624–635 (cit. on p. 21).

Aiello, Luca Maria, Georgios Petkos, Carlos Martin, David Corney, Symeon Papadopoulos, Ryan Skraba, Ayse Göker, Ioannis Kompatsiaris, and Alejandro Jaimes (2013). “Sensing Trending Topics in Twitter”. In: *IEEE Transactions on Multimedia* 15.6, pp. 1268–1282 (cit. on p. 9).

Allan, James, Jaime Carbonell, George Doddington, Jonathan Yamron, Yiming Yang, Brian Archibald, Doug Beeferman, Adam Berger, Ralf Brown, Ira Carp, Alex Hauptmann, John Lafferty, Victor Lavrenko, Xin Liu, Steve Lowe, Paul Van Mulbregt, Ron Papka, Thomas Pierce, Jay Ponte, and Mike Scudder (1998). “Topic Detection and Tracking Pilot Study Final Report”. In: *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, pp. 194–218 (cit. on p. 15).

Ambraseys, N. (1985). “Intensity-attenuation and magnitude-intensity relationships for northwest european earthquakes”. In: *Earthquake Engineering & Structural Dynamics* 13.6, pp. 733–778 (cit. on p. 78).

Ambraseys, N. and J. Douglas (2000). “Reappraisal of surface wave magnitudes in the Eastern Mediterranean region and the Middle East”. In: *Geophysical Journal International* 141.2, pp. 357–373 (cit. on p. 78).

Atdag, Samet and Vincent Labatut (2013). “A Comparison of Named Entity Recognition Tools Applied to Biographical Texts”. In: *CoRR* abs/1308.0661 (cit. on p. 116).

Atefeh, Farzindar and Wael Khreich (2015). “A Survey of Techniques for Event Detection in Twitter”. In: *Computational Intelligence* 31.1, pp. 132–164 (cit. on p. 21).

Backfried, Gerhard, Johannes Göllner, Gerald Quirchmayr, Karin Rainer, Gert Kienast, Georg Thallinger, Christian Schmidt, and Andreas Peer (2013). “Integration of Media Sources for Situation Analysis in the Different Phases of Disaster Management: The QuOIMA Project”. In: *Intelligence and Security Informatics Conference (EISIC), 2013 European*, pp. 143–146 (cit. on p. 3).

Baldi, Pierre, Søren Brunak, Yves Chauvin, Claus A. F. Andersen, and Henrik Nielsen (2000). “Assessing the accuracy of prediction algorithms for classification: an overview”. In: *Bioinformatics* 16.5, pp. 412–424 (cit. on p. 124).

Baldwin, Timothy, Paul Cook, Marco Lui, Andrew MacKinlay, and Li Wang (2013). “How Noisy Social Media Text, How Diffrnt Social Media Sources?” In: *Proceedings of the Sixth International Joint Conference on Natural Language Processing*. Nagoya, Japan: Asian Federation of Natural Language Processing, pp. 356–364 (cit. on pp. 2, 57, 105, 106).

Baldwin, Timothy, Valia Kordoni, and Aline Villavicencio (2009). “Prepositions in Applications: A Survey and Introduction to the Special Issue”. In: *Computational Linguistics* 35.2, pp. 119–149 (cit. on p. 102).

Banerjee, Satanjeev and Ted Pedersen (2002). "An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet". In: *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing*. CICLing '02. London, UK, UK: Springer-Verlag, pp. 136–145 (cit. on pp. 98, 117).

Bateman, John A., Joana Hois, Robert Ross, and Thora Tenbrink (2010). "A linguistic ontology of space for natural language processing". In: *Artificial Intelligence* 174.14, pp. 1027–1071 (cit. on p. 93).

Bauer, Christine (2013). "On the (In-)Accuracy of GPS Measures of Smartphones: A Study of Running Tracking Applications". In: *Proceedings of International Conference on Advances in Mobile Computing & Multimedia*. MoMM '13. Vienna, Austria: ACM, pp. 335–341 (cit. on p. 33).

Becker, Hila, Mor Naaman, and Luis Gravano (2011). "Beyond Trending Topics: Real-World Event Identification on Twitter." In: *Proceedings of the 5th International AAAI Conference on Web and Social Media*. Vol. 11. ICWSM'11. Barcelona, Spain, pp. 438–441 (cit. on pp. 10, 21–23).

Below, Regina, Angelika Wirtz, and Debarati Guha-Sapir (2009). *Disaster Category Classification and peril Terminology for Operational Purposes*. Tech. rep. Centre for Research on the Epidemiology of Disasters (CRED) andăMunich Reinsurance Company (Munich RE) (cit. on pp. 12, 13, 55).

Benson, Edward, Aria Haghighi, and Regina Barzilay (2011). "Event Discovery in Social Media Feeds". In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*. HLT '11. Portland, Oregon: Association for Computational Linguistics, pp. 389–398 (cit. on p. 21).

Berners-Lee, Tim and Mark Fischetti (1999). *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by Its Inventor*. 1st. Harper San Francisco (cit. on p. 1).

Bird, Steven, Ewan Klein, and Edward Loper (2009). *Natural Language Processing with Python*. O'Reilly Media (cit. on pp. 18, 94–96).

Boers, Frank (1996). *Spatial Prepositions and Metaphor. A cognitive Semantic Journey along the UP-DOWN and the FRONT-BACK Dimensions*. .12. Language in Performance. Tübingen: Gunter Narr Verlag (cit. on pp. 92, 101).

Boonthum, Chutima, Shunichi Toida, and Irwin Levinstein (2006). "Preposition Senses: Generalized Disambiguation Model". In: *Computational Linguistics and Intelligent Text Processing*. Ed. by Alexander Gelbukh. Vol. 3878. Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 196–207 (cit. on p. 93).

Breiman, Leo, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone (1984). "Classification and regression trees". In: *CA: Wadsworth International Group*. Statistics/Probability Series (cit. on p. 114).

Brody, Samuel and Nicholas Diakopoulos (2011). "Cooooooooooooooollllllllllllll!!!!!!!!!!!!!!: Using Word Lengthening to Detect Sentiment in Microblogs". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. EMNLP '11. Edinburgh, United Kingdom: Association for Computational Linguistics, pp. 562–570 (cit. on p. 57).

Bullinaria, John A. and Joseph P. Levy (2007). "Extracting semantic representations from word co-occurrence statistics: A computational study". English. In: *Behavior Research Methods* 39.3, pp. 510–526 (cit. on p. 20).

Burger, Wilhelm and Mark J. Burge (2008). *Digital Image Processing. An Algorithmic Introduction Using Java*. 1st ed. Texts in Computer Science. Springer-Verlag London, p. 99 (cit. on p. 44).

Burnard, Lou (1995). *User Guide for the British National Corpus* (cit. on p. 106).

Burton, K., N. Kasch, and I. Soboroff (2011). "The ICWSM 2011 Spinn3r Dataset". In: *Proceedings of the 5th Annual Conference on Weblogs and Social Media*. ICWSM'11. Barcelona, Spain (cit. on p. 106).

Cameron, A. Colin and Pravin K. Trivedi (2013). *Regression Analysis of Count Data*. Ed. by Rosa L. Matzkin and George J. Mailath. 2nd ed. Vol. 53. Econometric Society Monographs. Camebridge University Press (cit. on p. 48).

Carlson-Radvansky, Laura A. and David E. Irwin (1993). "Frames of reference in vision and language: Where is above?" In: *Cognition* 46.3, pp. 223–244 (cit. on p. 140).

CEDIM (2005). *GLOSSARY Terms and definitions of risk sciences*. Tech. rep. Center for Disaster Management and Risk Reduction Technology (cit. on p. 12).

Chae, Junghoon, Dennis Thom, Harald Bosch, Yun Jang, Ross Maciejewski, David S. Ebert, and Thomas Ertl (2012). "Spatiotemporal social media analytics for abnormal event detection and examination using seasonal-trend decomposition". In: *Proceedings of the IEEE Conference on Visual Analytics Science and Technology (VAST), 2012*, pp. 143–152 (cit. on p. 21).

Chen, Ying, Yilu Zhou, Sencun Zhu, and Heng Xu (2012). "Detecting Offensive Language in Social Media to Protect Adolescent Online Safety". In: *International Conference on Privacy, Security, Risk and Trust (PASSAT), 2012, and International Confernece on Social Computing (SocialCom), 2012*. Amsterdam, Netherlands, pp. 71–80 (cit. on p. 3).

Chu, Zi, Steven Gianvecchio, Haining Wang, and Sushil Jajodia (2010). "Who is Tweeting on Twitter: Human, Bot, or Cyborg?" In: *Proceedings of the 26th Annual Computer Security Applications Conference*. ACSAC '10. Austin, Texas, USA: ACM, pp. 21–30 (cit. on pp. 9, 48).

Cielen, Davy and Arno D. B. Meysman (2015). *Introducing Data Science*. Ed. by Mohamed Ali. Manning (cit. on p. 2).

Cieri, Christopher, Stephanie Strassel, David Graff, Nii Martey, Kara Rennert, and Mark Liberman (2002). "Corpora for Topic Detection and Tracking". English. In: *Topic Detection and Tracking*. Ed. by James Allan. Vol. 12. The Information Retrieval Series. Springer US, pp. 33–66 (cit. on p. 10).

Cohen, Kevin Bretonnel and Dina Demner-Fushman (2014). *Biomedical Natural Language Processing*. Ed. by Ruslan Mitkov. Natural Language Processing 11. John Benjamins Publishing Company (cit. on p. 96).

Coventry, Kenny R. and Simon C. Garrod (2004). *Saying, Seeing and Acting: The Psychological Semantics of Spatial Prepositions*. Ed. by Alan Baddeley, Vicki Bruce; Henry L. Roediger, and James R. Pomerantz. Essays in Cognitive Psychology. 27 Church Road, Hove, East Sussex, BN3 2FA: Taylor & Francis (cit. on pp. 94, 140).

Damerau, Fred J. (1964). "A Technique for Computer Detection and Correction of Spelling Errors". In: *Communications of the ACM* 7.3, pp. 171–176 (cit. on pp. 58, 109).

De Longueville, Bertrand, Robin S. Smith, and Gianluca Luraschi (2009). ""OMG, from Here, I Can See the Flames!": A Use Case of Mining Location Based Social Networks to Acquire Spatio-temporal Data on Forest Fires". In: *Proceedings of the 2009 International Workshop on Location Based Social Networks*. LBSN '09. Seattle, Washington: ACM, pp. 73–80 (cit. on p. 12).

Derczynski, Leon, Alan Ritter, Sam Clark, and Kalina Bontcheva (2013). "Twitter Part-of-Speech Tagging for All: Overcoming Sparse and Noisy Data". In: *Recent Advances in Natural Language Processing*. RANLP'13. Hissar, Bulgaria, pp. 198–206 (cit. on p. 96).

Dittrich, André and Christian Lucas (2013). "A step towards real-time analysis of major disaster events based on tweets". In: *Proceedings of the 10th International ISCRAM Conference*. Ed. by T. Comes, F. Fiedrich, S. Fortier, J. Geldermann, and T. Müller. Baden-Baden, Germany, pp. 868–874 (cit. on p. 9).

Doan, Son, Bao-Khanh Ho Vo, and Nigel Collier (2012). "An Analysis of Twitter Messages in the 2011 Tohoku Earthquake". English. In: *Electronic Healthcare*. Ed. by Patty Kostkova, Martin Szomszor, and David Fowler. Vol. 91. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering. Springer Berlin Heidelberg, pp. 58–66 (cit. on p. 12).

Dong, Xiaowen, Dimitrios Mavroeidis, Francesco Calabrese, and Pascal Frossard (2015). "Multiscale event detection in social media". English. In: *Data Mining and Knowledge Discovery* 29.5, pp. 1374–1405 (cit. on p. 21).

Eisenstein, Jacob (2013). "What to do about bad language on the internet". In: *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Atlanta, Georgia: Association for Computational Linguistics, pp. 359–369 (cit. on pp. 56, 57).

Finkel, Raphael A. and Jon Louis Bentley (1974). "Quad Trees: A Data Structure for Retrieval on Composite Keys." In: *Acta Informatica* 4, pp. 1–9 (cit. on p. 52).

Fleiss, Joseph L. (1971). "Measuring nominal scale agreement among many raters". In: *Psychological Bulletin* 76.5, pp. 378–382 (cit. on p. 125).

Fuhrmann, Thomas, Xiaoguang Luo, Andreas Knöpfler, and Michael Mayer (2014). "Generating statistically robust multipath stacking maps using congruent cells". In: *GPS Solutions* 19.1, pp. 83–92 (cit. on pp. 53, 54).

Furnas, George W., Thomas K. Landauer, Louis M. Gomez, and Susan T. Dumais (1983). "Statistical semantics: Analysis of the potential performance of keyword information access systems". In: *Bell System Technical Journal, (Special Issue on Human Factors in Computer Systems)* 62.6. Ed. by John C. Thomas and Michael L. Schneider, pp. 1753–1806 (cit. on p. 15).

Gärdenfors, Peter (2014). *The Geometry of Meaning: Semantics Based on Conceptual Spaces*. The MIT Press (cit. on pp. 93, 94).

Garrod, Simon C., Gillian Ferrier, and Siobhan Campbell (1999). "In and on: investigating the functional geometry of spatial prepositions". In: *Cognition* 72, pp. 167–189 (cit. on p. 140).

Garrod, Simon C. and Anthony J. Sanford (1989). "Discourse Models as Interfaces between Language and the Spatial World". In: *Journal of Semantics* 6, pp. 147–160 (cit. on p. 94).

Gelernter, Judith and Shilpa Balaji (2013). "An algorithm for local geoparsing of microtext". In: *GeoInformatica* 17.4, pp. 635–667 (cit. on p. 93).

Giardini, Domenico, Gottfried Gruenthal, Kaye Shedlock, and Peizhen Zhang (2003). "The GSHAP global seismic hazard map". In: *International Handbook of Earthquake and Engineering Seismology*. Ed. by William H.K. Lee, Hiroo Kanamori, Paul C. Jennings, and Carl Kisslinger. Vol. 81.B. International Geophysics. Academic Press. Chap. 74, pp. 1233–1239 (cit. on pp. 79, 81).

Goodchild, Michael F. (2007). "Citizens as sensors: the world of volunteered geography". English. In: *GeoJournal* 69.4, pp. 211–221 (cit. on p. 2).

Gorelick, Micha and Ian Ozsvald (2014). *High Performance Python. Practical Performant Programming for Humans*. 1st ed. OReilly Media, Inc. (cit. on p. 73).

Gosling, James, Bill Joy, Guy Steele, Gilad Bracha, and Alex Buckley (2015). *The Java Language Specification. Java SE 8 Edition*. Tech. rep. Oracle America, Inc. (cit. on p. 72).

Götz, Dieter and Gunter Lorenz (2002). *Englische Idioms von A - Z. Mit Erklärungen, Beispielen aus dem Sprachgebrauch und Übersetzungen*. 1st ed. The COBUILD series from the Bank of English. Ismaning: Hueber (cit. on p. 117).

Guzman, Jheser and Barbara Poblete (2013). "On-line Relevant Anomaly Detection in the Twitter Stream: An Efficient Bursty Keyword Detection Model". In: *Proceedings of the ACM SIGKDD Workshop on Outlier Detection and Description*. ODD '13. Chicago, Illinois: ACM, pp. 31–39 (cit. on p. 9).

Gwet, Kilem Li (2008). "Computing inter-rater reliability and its variance in the presence of high agreement". In: *British Journal of Mathematical and Statistical Psychology* 61.1, pp. 29–48 (cit. on p. 125).

Hahmann, Stefan, Ross Purves, and Dirk Burghardt (2014). "Twitter location (sometimes) matters: Exploring the relationship between georeferenced tweet content and nearby feature classes". In: *Journal of Spatial Information Science* 9.1, pp. 1–36 (cit. on p. 9).

Hall, Daniel B. (2000). "Zero-Inflated Poisson and Binomial Regression with Random Effects: A Case Study". In: *Biometrics* 56, pp. 1030–1039 (cit. on p. 50).

Hall, Mark and Christopher Jones (2012). "Cultural and Language Influences on the Interpretation of Spatial Prepositions". In: *GI Forum 2012* (cit. on p. 92).

Hall, Mark, Philip D. Smart, and Christopher B. Jones (2010). "Interpreting spatial language in image captions". In: *Cognitive Processing* 12.1, pp. 67–94 (cit. on p. 92).

Han, Bo and Timothy Baldwin (2011). "Lexical Normalisation of Short Text Messages: Makn Sens a #Twitter". In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*. HLT '11. Portland, Oregon: Association for Computational Linguistics, pp. 368–378 (cit. on p. 2).

Han, Bo, Paul Cook, and Timothy Baldwin (2012). "Automatically Constructing a Normalisation Dictionary for Microblogs". In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. EMNLP-CoNLL '12. Jeju Island, Korea: Association for Computational Linguistics, pp. 421–432 (cit. on p. 2).

Haralick, Robert M. and Linda G. Shapiro (1992). *Computer and Robot Vision*. 1st. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc. (cit. on p. 66).

Hecht, Brent, Lichan Hong, Bongwon Suh, and Ed H. Chi (2011). "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. Vancouver, BC, Canada: ACM, pp. 237–246 (cit. on pp. 9, 31, 32).

Herskovits, Anette (1985). "Semantics and Pragmatics of Locative Expressions". In: *Cognitive Science* 9, pp. 341–378 (cit. on p. 91).

Herskovits, Anette (1986). *Language and Spatial Cognition: An interdisciplinary study of the prepositions in English*. Cambridge, UK: Cambridge University Press (cit. on pp. 91, 92, 94).

Hill, Linda L. (2006). *Georeferencing: The Geographic Associations of Information*. Digital Libraries and Electronic Publishing. The MIT Press (cit. on p. 32).

Huck, Jonny J., J. Duncan Whyatt, and Paul Coulton (2015). "Visualizing patterns in spatially ambiguous point data". In: *Journal of Spatial Information Science* 10.1, pp. 47–66 (cit. on p. 9).

Hughes, Amanda L. and Leysia Palen (2009). "Twitter adoption and use in mass convergence and emergency events". In: *International Journal of Emergency Management* 6.3, pp. 248–260 (cit. on p. 12).

IFRC (2013). *India: Cyclone Phailin*. 1. International Federation of Red Cross and Red Crescent Societies (IFRC) Disaster Relief Emergency Fund (DREF) (cit. on p. 88).

Imran, Muhammad, Shady Elbassuoni, Carlos Castillo, Fernando Diaz, and Patrick Meier (2013). "Extracting Information Nuggets from Disaster-Related Messages in Social Media". In: *Proceedings of the 10th International Conference on Information Systems for Crisis Response and Management*. Ed. by Tina Comes, Frank Fiedrich, Stephen C. Fortier, Jutta Geldermann, and Tim Müller. ISCRAM. Baden-Baden (cit. on p. 3).

International Organization for Standardization (2015). *Information technology – Vocabulary*. Norm ISO/IEC 2382:2015(en). Geneva, Switzerland: International Organization for Standardization (cit. on p. 14).

Jackoway, Alan, Hanan Samet, and Jagan Sankaranarayanan (2011). "Identification of Live News Events Using Twitter". In: *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks*. LBSN '11. Chicago, Illinois: ACM, pp. 25–32 (cit. on p. 9).

Jagadish, Hosagrahar V., Johannes Gehrke, Alexandros Labrinidis, Yannis Papakonstantinou, Jignesh M. Patel, Raghu Ramakrishnan, and Cyrus Shahabi (2014). "Big Data and Its Technical Challenges". In: *Communications of the ACM* 57.7, pp. 86–94 (cit. on p. 2).

Java, Akshay, Xiaodan Song, Tim Finin, and Belle Tseng (2007). "Why We Twitter: Understanding Microblogging Usage and Communities". In: *Procedings of the Joint 9th WEBKDD and 1st SNA-KDD Workshop 2007*. Springer, pp. 56–65 (cit. on pp. 3, 9, 28).

Jurafsky, Daniel and James H. Martin (2009). *Speech and Language Processing (2nd Edition)*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc. (cit. on pp. 17, 20, 95–97).

Keranen, Kathleen M., Geoffrey A. Abers, Matthew Weingarten, Barbara A. Bekins, and Shemin Ge (2014). "Sharp increase in central Oklahoma seismicity since 2008 induced by massive wastewater injection". In: 345.6195, pp. 448–451 (cit. on p. 69).

Khan, Arbaz, Maria Vasardani, and Stephan Winter (2013). "Extracting Spatial Information From Place Descriptions". In: *Proceedings of The First ACM SIGSPATIAL International Workshop on Computational Models of Place*. COMP '13. Orlando FL, USA: ACM, 62:62–62:69 (cit. on pp. 93, 99).

Kireyev, Kirill, Leysia Palen, and Kenneth Anderson (2009). "Applications of Topics Models to Analysis of Disaster-Related Twitter Data". In: *NIPS Workshop on Applications for Topic Models: Text and Beyond*. Whistler, Canada (cit. on pp. 9, 12).

Kordjamshidi, Parisa, Paolo Frasconi, Martijn Van Otterlo, Marie-Francine Moens, and Luc De Raedt (2012). "Relational Learning for Spatial Relation Extraction from Natural Language". In: *Inductive Logic Programming*. Ed. by StephenH. Muggleton, Alireza Tamaddoni-Nezhad, and FrancescaA. Lisi. Vol. 7207. Lecture Notes in Computer Science. Berlin/Heidelberg, Germany, pp. 204–220 (cit. on p. 93).

Kordjamshidi, Parisa, Martijn Van Otterlo, and Marie-Francine Moens (2010). "Spatial Role Labeling: Task Definition and Annotation Scheme". In: *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10)*. Ed. by Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias. Valletta, Malta: European Language Resources Association (ELRA) (cit. on p. 93).

Kordjamshidi, Parisa, Martijn Van Otterlo, and Marie-Francine Moens (2011). "Spatial Role Labeling: Towards Extraction of Spatial Relations from Natural Language". In: *ACM Transactions on Speech and Language Processing* 8.3, 4:1–4:36 (cit. on p. 93).

Kracht, Marcus (2002). "On the Semantics of Locatives". English. In: *Linguistics and Philosophy* 25.2, pp. 157–232 (cit. on p. 102).

Krumm, John and Eric Horvitz (2015). "Eyewitness: Identifying Local Events via Space-Time Signals in Twitter Feeds". In: *Proceedings of the 23rd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. SIGSPATIAL '15. ACM (cit. on pp. 9, 12, 21, 23–25, 32, 41, 52).

Kummu, Matti and Olli Varis (2011). "The world by latitudes: A global analysis of human population, development level and environment across the northsouth axis over the past half century". In: *Applied Geography* 31.2, pp. 495–507 (cit. on p. 36).

Lakoff, George and Mark Johnson (1980). *Metaphors we live by*. Chicago: The University of Chicago Press (cit. on pp. 92, 101).

Landau, Barbara and Ray Jackendoff (1993). ""What" and "where" in spatial language and spatial cognition". In: *Behavioral and Brain Sciences* 16, pp. 217–265 (cit. on pp. 91, 94).

Landis, J. Richard and Gary G. Koch (1977). “The Measurement of Observer Agreement for Categorical Data”. In: *Biometrics* 33.1, pp. 159–174 (cit. on p. 126).

Lee, Chung-Hong, Chih-Hong Wu, and Tzan-Feng Chien (2011). “BursT: A Dynamic Term Weighting Scheme for Mining Microblogging Messages”. English. In: *Advances in Neural Networks ISNN 2011*. Ed. by Derong Liu, Huaguang Zhang, Marios Polycarpou, Cesare Alippi, and Haibo He. Vol. 6677. Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 548–557 (cit. on pp. 9, 22).

Lee, Chung-Hong, Hsin-Chang Yang, Tzan-Feng Chien, and Wei-Shiang Wen (2011). “A Novel Approach for Event Detection by Mining Spatio-temporal Information on Microblogs”. In: *International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2011*, pp. 254–259 (cit. on pp. 10, 21–23, 25).

Lee, Kyumin, Brian D. Eoff, and James Caverlee (2011). “Seven Months with the Devils: A Long-Term Study of Content Polluters on Twitter”. In: *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*. ICWSM’11 (cit. on p. 9).

Lee, Ryong and Kazutoshi Sumiya (2010). “Measuring Geographical Regularities of Crowd Behaviors for Twitter-based Geo-social Event Detection”. In: *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks*. LBSN ’10. San Jose, California: ACM, pp. 1–10 (cit. on p. 9).

Lee, Ryong, Shoko Wakamiya, and Kazutoshi Sumiya (2011). “Discovery of Unusual Regional Social Activities Using Geo-tagged Microblogs”. In: *World Wide Web* 14.4, pp. 321–349 (cit. on p. 9).

Leetaru, Kalev, Shaowen Wang, Guofeng Cao, Anand Padmanabhan, and Eric Shook (2013). “Mapping the global Twitter heartbeat: The geography of Twitter”. In: *First Monday* 18.5 (cit. on pp. 9, 31, 32, 59, 60).

Leibniz, Gottfried Wilhelm (1765). *Nouveaux essais sur l’entendement humain*. Amsterdam, Netherlands: v. R. E. Raspe (cit. on p. 94).

Lesk, Michael (1986). “Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone”. In: *Proceedings of the 5th Annual International Conference on Systems Documentation*. SIGDOC ’86. Toronto, Ontario, Canada: ACM, pp. 24–26 (cit. on p. 98).

Levenshtein, Vladimir I. (1966). “Binary codes capable of correcting deletions, insertions, and reversals”. In: *Cybernetics and Control Theory* 10 (8), pp. 707–710 (cit. on pp. 58, 109).

Levinson, Stephen C. (1996). “LANGUAGE AND SPACE”. In: *Annual Review of Anthropology* 25.1, pp. 353–382 (cit. on p. 92).

Li, Chenliang, Aixin Sun, and Anwitaman Datta (2012). “Twevent: Segment-based Event Detection from Tweets”. In: *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*. CIKM ’12. Maui, Hawaii, USA: ACM, pp. 155–164 (cit. on pp. 9, 21).

Li, Hanjing, Tiejun Zhao, Sheng Li, and Yanhai Han (2006). “The Extraction of Spatial Relationships from Text Based on Hybrid Method”. In: *Information Acquisition, 2006 IEEE International Conference on*, pp. 284–289 (cit. on p. 93).

Li, Rui, Kin Hou Lei, Ravi Khadiwala, and Kevin Chen-Chuan Chang (2012). “TEDAS: A Twitter-based Event Detection and Analysis System”. In: *Proceedings of the 2012 IEEE 28th International Conference on Data Engineering*. ICDE ’12. Washington, DC, USA: IEEE Computer Society, pp. 1273–1276 (cit. on p. 21).

Litkowski, Ken (2014). “Pattern Dictionary of English Prepositions”. In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Baltimore, Maryland: Association for Computational Linguistics, pp. 1274–1283 (cit. on p. 103).

Litkowski, Ken and Orin Hargraves (2007). "SemEval-2007 Task 06: Word-sense Disambiguation of Prepositions". In: *Proceedings of the 4th International Workshop on Semantic Evaluations*. SemEval '07. Prague, Czech Republic: Association for Computational Linguistics, pp. 24–29 (cit. on p. 93).

Liu, Fei, Maria Vasardani, and Timothy Baldwin (2014). "Automatic Identification of Locative Expressions from Social Media Text: A Comparative Analysis". In: *Proceedings of the 4th International Workshop on Location and the Web*. LocWeb '14. Shanghai, China: ACM, pp. 9–16 (cit. on p. 106).

Lotan, Gilad (2011). "Mapping Information Flows on Twitter". In: *The Future of the Social Web*. Vol. WS-11-03. AAAI Workshops. AAAI (cit. on p. 9).

Lucas, Christian (2012). "Multi-criteria modelling and clustering of spatial information". In: *International Journal of Geographical Information Science* 26.10, pp. 1897–1915 (cit. on p. 92).

Lui, Marco and Timothy Baldwin (2012). "langid.py: An Off-the-shelf Language Identification Tool". In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL)*. Jeju, Republic of Korea, pp. 25–30 (cit. on p. 61).

Lui, Marco and Timothy Baldwin (2014). "Accurate Language Identification of Twitter Messages". In: *Proceedings of the 5th Workshop on Language Analysis for Social Media (LASM)*. Gothenburg, Sweden: Association for Computational Linguistics, pp. 17–25 (cit. on p. 61).

Malik, Momin, Hemank Lamba, Constantine Nakos, and Jürgen Pfeffer (2015). "Population Bias in Geotagged Tweets". In: *Proceedings of the 9th International AAAI Conference on Web and Social Media*. ICWSM'15. Barcelona, Spain (cit. on p. 35).

Mani, Inderjeet, Janet Hitzeman, Justin Richer, Dave Harris, Rob Quimby, and Ben Wellner (2008). *SpatialML: Annotation Scheme, Corpora, and Tools*. Tech. rep. The MITRE Corporation (cit. on p. 93).

Manning, Christopher D. (2011). "Part-of-Speech Tagging from 97% to 100%: Is It Time for Some Linguistics?" In: *Proceedings of the 12th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing'11)*. Ed. by Alexander F. Gelbukh. .1. Tokyo, Japan: Springer Berlin Heidelberg, pp. 171–189 (cit. on p. 95).

Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze (2008). *Introduction to Information Retrieval*. Cambridge University Press (cit. on pp. 14, 16–20).

Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky (2014). "The Stanford CoreNLP Natural Language Processing Toolkit". In: *Association for Computational Linguistics (ACL) System Demonstrations*, pp. 55–60 (cit. on p. 97).

Marneffe, Marie-Catherine de, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D. Manning (2014). "Universal Stanford dependencies: A cross-linguistic typology". In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation*. LREC'14. Reykjavik, Iceland, pp. 4585–4592 (cit. on p. 97).

Mathioudakis, Michael and Nick Koudas (2010). "TwitterMonitor: Trend Detection over the Twitter Stream". In: *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data*. SIGMOD '10. Indianapolis, Indiana, USA: ACM, pp. 1155–1158 (cit. on p. 9).

Matthews, B. W. (1975). "Comparison of the predicted and observed secondary structure of T4 phage lysozyme". In: *Biochimica et Biophysica Acta* 405, pp. 442–451 (cit. on p. 124).

McIntosh, Colin, Ben Francis, and Richard Poole, eds. (2002). *Oxford Collocations Dictionary for Students of English*. Oxford University Press, USA (cit. on p. 118).

Miller, George A. (1995). "WordNet: A Lexical Database for English". In: *Communications of the ACM* 38.11, pp. 39–41 (cit. on p. 56).

Miller, George A. and Philip Johnson-Laird (1976). *Language and Perception*. Harvard, USA: Harvard University Press (cit. on pp. 91, 94).

Mitkov, Ruslan (2003). *The Oxford Handbook of Computational Linguistics (Oxford Handbooks in Linguistics S.)* Oxford University Press (cit. on p. 18).

Mühr, Bernhard, Daniel Köbele, Tina Bessel, Joachim Fohringer, and Christian Lucas (2013). *Super Cyclonic Storm 02B Phailin*. Report 1. Center for Disaster Management and Risk Reduction Technology (cit. on pp. 87, 88).

Naaman, Mor, Hila Becker, and Luis Gravano (2011). "Hip and trendy: Characterizing emerging trends on Twitter". In: *Journal of the American Society for Information Science and Technology* 62.5, pp. 902–918 (cit. on p. 9).

Nurwidyantoro, Arif and Edi Winarko (2013). "Event detection in social media: A survey". In: *International Conference on ICT for Smart Society (ICISS)*. IEEE, pp. 1–5 (cit. on p. 21).

O'Hara, Tom and Janyce Wiebe (2003). "Preposition Semantic Classification via Penn Treebank and FrameNet". In: *Proceedings of the Seventh Conference on Natural Language Learning*. CONLL '03. Edmonton, Canada: Association for Computational Linguistics, pp. 79–86 (cit. on p. 93).

O'Keefe, John (1996). "The Spatial Prepositions in English, Vector Grammar, and the Cognitive Map Theory". In: *Language and Space*. Ed. by Paul Bloom, Mary A. Peterson, Lynn Nadel, and Merrill F. Garrett. Cambridge, MA, USA: MIT Press. Chap. 7, pp. 277–316 (cit. on p. 94).

Oliveira, Maxwell Guimarães, Cláudio E. C. Campelo, Cláudio Souza Baptista, and Michela Bertolotto (2015). "Leveraging VGI for Gazetteer Enrichment: A Case Study for Geoparsing Twitter Messages". In: *Proceedings of the 14th International Symposium for Web and Wireless Geographical Information Systems (W2GIS)*. Ed. by Jérôme Gensel and Martin Tomko. 9080 vols. Lecture Notes in Computer Science. Grenoble, France: Springer International Publishing, pp. 20–36 (cit. on p. 93).

Paek, Jeongyeup, Kyu-Han Kim, Jatinder P. Singh, and Ramesh Govindan (2011). "Energy-efficient Positioning for Smartphones Using Cell-ID Sequence Matching". In: *Proceedings of the 9th International Conference on Mobile Systems, Applications, and Services*. MobiSys '11. Bethesda, Maryland, USA: ACM, pp. 293–306 (cit. on p. 33).

Pesyna Jr., Kenneth M., Robert W. Heath Jr., and Todd E. Humphreys (2014). "Centimeter Positioning with a Smartphone-Quality GNSS Antenna". In: *Proceedings of the 27th International Technical Meeting of The Satellite Division of the Institute of Navigation*. Tampa, Florida, pp. 1568–1577 (cit. on p. 33).

Petrov, Slav, Dipanjan Das, and Ryan McDonald (2012). "A Universal Part-of-Speech Tagset". In: *Proceedings of the Eight International Conference on Language Resources and Evaluation*. Ed. by Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uur Doan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis. LREC'12. Istanbul, Turkey: European Language Resources Association (ELRA) (cit. on p. 97).

Petrovi, Saa, Miles Osborne, and Victor Lavrenko (2012). "Using Paraphrases for Improving First Story Detection in News and Twitter". In: *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. NAACL HLT '12. Montreal, Canada: Association for Computational Linguistics, pp. 338–346 (cit. on pp. 2, 9).

Pustejovsky, James, Jessica Moszkowicz, and Marc Verhagen (2011). "ISO-Space: The annotation of spatial information in language". In: *Proceedings of the Sixth Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation*. Oxford, UK, pp. 1–9 (cit. on p. 93).

Pustejovsky, James, Jessica Moszkowicz, and Marc Verhagen (2013). "A Linguistically Grounded Annotation Language for Spatial Information". In: *Revue TAL-Traitement Automatique des Langues* 53.2, pp. 87–113 (cit. on p. 93).

Ratcliff, John W. and David E. Metzener (1988). "Pattern Matching: the Gestalt Approach". In: *Dr.Dobb's Journal* 13.7, pp. 46–51 (cit. on p. 58).

Retz-Schmidt, Gudula (1988). "Various Views On Spatial Prepositions". In: *AI Magazine* 9.2, pp. 95–105 (cit. on p. 92).

Rijsbergen, C. J. Van (1979). *Information Retrieval*. 2nd. Newton, MA, USA: Butterworth-Heinemann (cit. on p. 20).

Ritter, Alan, Sam Clark, Mausam, and Oren Etzioni (2011). "Named Entity Recognition in Tweets: An Experimental Study". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. EMNLP '11. Edinburgh, United Kingdom: Association for Computational Linguistics, pp. 1524–1534 (cit. on p. 93).

Ritter, Alan, Mausam, Oren Etzioni, and Sam Clark (2012). "Open Domain Event Extraction from Twitter". In: *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '12. Beijing, China: ACM, pp. 1104–1112 (cit. on p. 21).

Sakaki, Takeshi, Makoto Okazaki, and Yutaka Matsuo (2010). "Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors". In: *Proceedings of the 19th International Conference on World Wide Web*. WWW '10. Raleigh, North Carolina, USA: ACM, pp. 851–860 (cit. on pp. 12, 21).

Saleem, Haji Mohammad, Yishi Xu, and Derek Ruths (2014). "Effects of Disaster Characteristics on Twitter Event Signature". In: *Procedia Engineering* 78. Humanitarian Technology: Science, Systems and Global Impact 2014, HumTech2014, pp. 165–172 (cit. on p. 12).

Salton, Gerard M., Andrew K.C. Wong, and Chung-shu Yang (1975). "A Vector Space Model for Automatic Indexing". In: *Communications of the ACM* 18.11, pp. 613–620 (cit. on p. 15).

Samet, Hanan (1984). "The Quadtree and Related Hierarchical Data Structures". In: *ACM Computing Surveys* 16.2, pp. 187–260 (cit. on p. 52).

Sanderson, Mark and W. Bruce Croft (2012). "The History of Information Retrieval Research." In: *Proceedings of the IEEE* 100.Centennial-Issue, pp. 1444–1451 (cit. on p. 14).

Sankaranarayanan, Jagan, Hanan Samet, Benjamin E. Teitler, Michael D. Lieberman, and Jon Sperling (2009). "TwitterStand: News in Tweets". In: *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. GIS '09. Seattle, Washington: ACM, pp. 42–51 (cit. on pp. 9, 23).

Santorini, Beatrice (1990). *Part-of-speech tagging guidelines for the Penn Treebank Project*. Technical report MS-CIS-90-47. Department of Computer and Information Science, University of Pennsylvania (cit. on p. 95).

Sayyadi, Hassan, Matthew Hurst, and Alexey Maykov (2009). "Event detection and tracking in social streams". In: *Proceedings of the 3rd International AAAI Conference on Weblogs and Social Media*. ICWSM'09 (cit. on p. 21).

Shebalin, Nikolai. V., Günter Leydecker, N.G. Mokrushina, R.E. Tatevossian, O.O. Erteleva, and V.YU. Vassiliev (1997). *Earthquake Catalogue for Central and Southeastern Europe 342 BC - 1990 AD*. ETNU CT 93 - 0087. Brussels: European Commission (cit. on p. 78).

Shen, Qijun, Xueying Zhang, and Wenming Jiang (2009). "Annotation of Spatial Relations in Natural Language". In: *Proceedings of the International Conference on Environmental Science and Information Application Technology, 2009*. Vol. 3. Wuhan, China, pp. 418–421 (cit. on p. 93).

Snyder, John P. (1993). "Flattening the Earth: Two Thousand Years of Map Projections". In: Chicago: The University of Chicago Press. Chap. 3, pp. 112–113 (cit. on p. 41).

Srikumar, Vivek and Dan Roth (2013). "An Inventory of Preposition Relations". In: *CoRR* abs/1305.5785 (cit. on p. 93).

Starbird, Kate and Leysia Palen (2010). “Pass It On?: Retweeting in Mass Emergencies”. In: *Proceedings of the 7th International Conference on Information Systems for Crisis Response and Management*. ISCRAM. Seattle, USA (cit. on p. 12).

Stefanidis, Anthony, Andrew Crooks, and Jacek Radzikowski (2013). “Harvesting ambient geospatial information from social media feeds”. English. In: *GeoJournal* 78.2, pp. 319–338 (cit. on p. 9).

Stenetorp, Pontus, Sampo Pyysalo, Goran Topi, Tomoko Ohta, Sophia Ananiadou, and Jun’ichi Tsujii (2012). “BRAT: A Web-based Tool for NLP-assisted Text Annotation”. In: *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*. EACL ’12. Avignon, France: Association for Computational Linguistics, pp. 102–107 (cit. on p. 96).

Stollberg, Beate and Tom de Groeve (2012). “The Use of Social Media Within the Global Disaster Alert and Coordination System (GDACS)”. In: *Proceedings of the 21st International Conference on World Wide Web*. WWW ’12 Companion. Lyon, France: ACM, pp. 703–706 (cit. on p. 3).

Sugitani, Takuya, Masumi Shirakawa, Takahiro Hara, and Shojiro Nishio (2013). “Detecting Local Events by Analyzing Spatiotemporal Locality of Tweets”. In: *27th International Conference on Advanced Information Networking and Applications Workshops (WAINA), 2013*, pp. 191–196 (cit. on p. 21).

Szalay, Alex, Jim Gray, Gyorgy Fekete, Peter Kunszt, Peter Kukol, and Ani Thakar (2005). *Indexing the Sphere with the Hierarchical Triangular Mesh*. Tech. rep. MSR-TR-2005-123. Microsoft Research (cit. on p. 23).

Tapia, Andrea H., Kathleen A. Moore, and Nichloas J. Johnson (2013). “Beyond the Trustworthy Tweet: A Deeper Understanding of Microblogged Data Use by Disaster Response and Humanitarian Relief Organizations”. In: *Proceedings of the 10th International Conference on Information Systems for Crisis Response and Management*. Ed. by Tina Comes, Frank Fiedrich, Stephen C. Fortier, Jutta Geldermann, and Tim Müller. ISCRAM. Baden-Baden (cit. on p. 3).

Terpstra, Teun, Richard Stronkman, Arnout De Vries, and Geerte L. Paradies (2012). “Towards a realtime Twitter analysis during crises for operational crisis management”. In: *Proceedings of the 9th International Conference on Information Systems for Crisis Response and Management*. ISCRAM (cit. on pp. 3, 12).

Tobler, Waldo R. (1970). “A Computer Movie Simulating Urban Growth in the Detroit Region”. In: *Economic Geography* 46, pp. 234–240 (cit. on p. 62).

Toutanova, Kristina, Dan Klein, Christopher D. Manning, and Yoram Singer (2003). “Feature-rich Part-of-speech Tagging with a Cyclic Dependency Network”. In: *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*. Vol. 1. NAACL ’03. Edmonton, Canada: Association for Computational Linguistics, pp. 173–180 (cit. on p. 95).

Turney, Peter D. and Patrick Pantel (2010). “From Frequency to Meaning: Vector Space Models of Semantics”. In: *J. Artif. Int. Res.* 37.1, pp. 141–188 (cit. on pp. 15, 17, 20).

UNISDR (2009). *Terminology on Disaster Risk Reduction*. Tech. rep. Geneva, Switzerland: United Nations International Strategy for Disaster Reduction (cit. on p. 11).

U.S. Geological Survey (2000). *The severity of an earthquake*. USGS Unnumbered Series. U.S. Geological Survey. eprint: `http://pubs.usgs.gov/gip/earthq4/severitygip.html` (cit. on p. 78).

Valkanas, George and Dimitrios Gunopulos (2013). “Event Detection from Social Media Data”. In: *IEEE Data Engineering Bulletin* 36.3, pp. 51–58 (cit. on p. 21).

Vanderwende, Lucy, Hisami Suzuki, Chris Brockett, and Ani Nenkova (2007). “Beyond SumBasic: Task-focused Summarization with Sentence Simplification and Lexical Expansion”. In: *Information Processing & Management* 43.6, pp. 1606–1618 (cit. on p. 24).

Vasardani, Maria, Sabine Timpf, Stephan Winter, and Martin Tomko (2013). "From Descriptions to Depictions: A Conceptual Framework". In: *Spatial Information Theory*. Ed. by Thora Tenbrink, John Stell, Antony Galton, and Zena Wood. Vol. 8116. Lecture Notes in Computer Science. Springer, pp. 299–319 (cit. on p. 92).

Vasardani, Maria, Stephan Winter, Kai-Florian Richter, Lesley Stirling, and Daniela Richter (2012). "Spatial Interpretations of Preposition "at"". In: *Proceedings of the 1st ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information*. GEOCROWD '12. Redondo Beach, California: ACM, pp. 46–53 (cit. on p. 140).

Vieweg, Sarah, Amanda L. Hughes, Kate Starbird, and Leysia Palen (2010). "Microblogging During Two Natural Hazards Events: What Twitter May Contribute to Situational Awareness". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '10. Atlanta, Georgia, USA: ACM, pp. 1079–1088 (cit. on p. 12).

Villanueva, Diana, Alma-Delia Cuevas-Rasgado, Omar Juárez, and Adolfo Guzmán-Arenas (2013). "Using frames to disambiguate prepositions". In: *Expert Systems with Applications* 40.2, pp. 598–610 (cit. on pp. 93, 97).

Wald, David J., Vincent Quitoriano, Thomas H. Heaton, and Hiroo Kanamori (1999). "Relationships between Peak Ground Acceleration, Peak Ground Velocity, and Modified Mercalli Intensity in California". In: *Earthquake Spectra* 15.3, pp. 557–564. eprint: http://dx.doi.org/10.1193/1.1586058 (cit. on p. 79).

Walter, Elizabeth, Virginia Klein, Mairi MacDonald, and Melissa Good, eds. (2006). *Cambridge Idioms Dictionary*. 2nd ed. Cambridge University Press (cit. on p. 105).

Walther, Maximilian and Michael Kaisser (2013). "Geo-spatial Event Detection in the Twitter Stream". In: *Proceedings of the 35th European Conference on Advances in Information Retrieval*. ECIR'13. Moscow, Russia: Springer, pp. 356–367 (cit. on pp. 9, 21).

Wang, Xun, Feida Zhu, Jing Jiang, and Sujian Li (2013). "Real Time Event Detection in Twitter". English. In: *Web-Age Information Management*. Ed. by Jianyong Wang, Hui Xiong, Yoshiharu Ishikawa, Jianliang Xu, and Junfeng Zhou. Vol. 7923. Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 502–513 (cit. on p. 21).

Watanabe, Kazufumi, Masanao Ochi, Makoto Okabe, and Rikio Onai (2011). "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*. CIKM '11. Glasgow, Scotland, UK: ACM, pp. 2541–2544 (cit. on pp. 21, 32).

Weaver, Warren (1955). "Translation". In: *Machine translation of languages: fourteen essays*. Ed. by William Nash Locke and Andrew Donald Booth. The Technology Press of The Massachusetts Institute of Technology and John Wiley & Sons, Inc., New York Chapman & Hall, Ltd., London. Chap. 1, pp. 15–23 (cit. on p. 15).

Weng, Jianshu and Bu-Sung Lee (2011). "Event Detection in Twitter". In: *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*. ICWSM '11. Barcelona, Spain: AAAI Press (cit. on p. 21).

Wertheimer, Max (1923). "Untersuchungen zur Lehre von der Gestalt". In: *Psychologische Forschung: Zeitschrift für Psychologie und ihre Grenzwissenschaften* 4, pp. 301–350 (cit. on p. 58).

Wu, Sun and Udi Manber (1992). "Fast Text Searching: Allowing Errors". In: *Communications of the ACM* 35.10, pp. 83–91 (cit. on p. 109).

Xiang, Guang, Bin Fan, Ling Wang, Jason Hong, and Carolyn Rose (2012). "Detecting Offensive Tweets via Topical Feature Discovery over a Large Scale Twitter Corpus". In: *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*. CIKM '12. Maui, Hawaii, USA: ACM, pp. 1980–1984 (cit. on p. 3).

Yang, Yiming, Tom Pierce, and Jaime Carbonell (1998). "A Study of Retrospective and On-line Event Detection". In: *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '98. Melbourne, Australia: ACM, pp. 28–36 (cit. on p. 10).

Zandbergen, Paul A. (2009). "Accuracy of iPhone Locations: A Comparison of Assisted GPS, WiFi and Cellular Positioning". In: *Transactions in GIS* 13, pp. 5–25 (cit. on p. 33).

Zandbergen, Paul A. and Sean J. Barbeau (2011). "Positional Accuracy of Assisted GPS Data from High-Sensitivity GPS-enabled Mobile Phones". In: *Journal of Navigation* 64 (03), pp. 381–399 (cit. on p. 33).

Zeman, Daniel (2008). "Reusable Tagset Conversion Using Tagset Drivers". In: *Proceedings of the Sixth International Conference on Language Resources and Evaluation*. Ed. by Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, and Daniel Tapias. LREC'08. http://www.lrec-conf.org/proceedings/lrec2008/. Marrakech, Morocco: European Language Resources Association (ELRA) (cit. on p. 97).

Zhang, Chunju, Xueying Zhang, Wenming Jiang, Qijun Shen, and Shanqi Zhang (2009). "Rule-Based Extraction of Spatial Relations in Natural Language Text". In: *Proceedings of the International Conference on Computational Intelligence and Software Engineering, 2009*, pp. 1–4 (cit. on p. 93).

Zhang, Wei and Judith Gelernter (2014). "Geocoding location expressions in Twitter messages: A preference learning method". In: *Journal of Spatial Information Science* 9.1, pp. 37–70 (cit. on p. 31).

Zhang, Xueying, Chunju Zhang, Chaoli Du, and Shaonan Zhu (2011). "SVM based extraction of spatial relations in text". In: *Proceedings of the 2011 IEEE International Conference on Spatial Data Mining and Geographical Knowledge Services (ICSDM 2011)*, pp. 529–533 (cit. on p. 93).

Zhao, Qiankun, Prasenjit Mitra, and Bi Chen (2007). "Temporal and Information Flow Based Event Detection from Social Text Streams". In: *Proceedings of the 22nd National Conference on Artificial Intelligence - Volume 2*. AAAI '07. Vancouver, British Columbia, Canada: AAAI Press, pp. 1501–1506 (cit. on pp. 10, 21).

Zhao, Wayne Xin, Baihan Shu, Jing Jiang, Yang Song, Hongfei Yan, and Xiaoming Li (2012). "Identifying Event-related Bursts via Social Media Activities". In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. EMNLP-CoNLL '12. Jeju Island, Korea: Association for Computational Linguistics, pp. 1466–1477 (cit. on p. 21).

Zin, Thi Thi, Pyke Tin, Hiromitsu Hama, and Takashi Toriu (2013). "Knowledge based Social Network Applications to Disaster Event Analysis". In: *Proceedings of the International MultiConference of Engineers and Computer Scientists*. IMECS'13. Hong Kong (cit. on p. 3).

Online Resources

Bender, Jeremy (2015). *The six countries that block social media*. Business Insider. URL: http://www.businessinsider.com/the-six-countries-that-block-social-media-2015-4?IR=T (visited on Jan. 4, 2016) (cit. on p. 28).

Cambridge University Press (2015). *event Meaning in the Cambridge English Dictionary*. URL: http://dictionary.cambridge.org/dictionary/english/event (visited on Oct. 3, 2015) (cit. on p. 10).

Cavers, Ian (1998). *An Introductory Guide to MATLAB*. Ed. by University of British Columbia. URL: http://www.cs.ubc.ca/~ascher/542-403/MatlabGuide.pdf (visited on Feb. 23, 2016) (cit. on p. 72).

Doyle, Damian (2016). *Stopwords*. Ranks NL. URL: http://www.ranks.nl/stopwords (visited on Jan. 28, 2016) (cit. on p. 56).

Geographical JSON Working Group (2008). *GeoJSON Specification*. URL: http://geojson.org/geojson-spec.html (visited on Nov. 17, 2015) (cit. on p. 32).

Guardian News and Media Ltd. (2009). *China blocks Twitter, Flickr, YouTube and Hotmail ahead of Tiananmen anniversary | Technology | The Guardian*. URL: http://www.theguardian.com/technology/2009/jun/02/twitter-china (visited on Oct. 30, 2015) (cit. on p. 28).

IdiomSite (2015). *IdiomSite.com - Find out the meanings of common sayings*. URL: http://www.idiomsite.com/ (visited on Apr. 19, 2016) (cit. on p. 117).

Internet Live Stats (2016). *Internet Live Stats - Internet Usage & Social Media Statistics*. URL: http://www.internetlivestats.com/ (visited on Jan. 2, 2016) (cit. on pp. 2, 28).

Merriam-Webster, Inc. (2015). *Event|Definition of event by Merriam-Webster*. URL: http://www.merriam-webster.com/dictionary/event (visited on Oct. 3, 2015) (cit. on pp. 10, 103).

MySQL (2016). *MySQL :: MySQL 5.5 Reference Manual :: 12.9.4 Full-Text Stopwords*. URL: http://dev.mysql.com/doc/refman/5.5/en/fulltext-stopwords.html (visited on Jan. 28, 2016) (cit. on p. 56).

Oxford University Press (2015). *event - definition of event in English from the Oxford dictionary*. URL: http://www.oxforddictionaries.com/definition/english/event (visited on Oct. 3, 2015) (cit. on pp. 10, 103).

PostgreSQL (2008). *PostgreSQL CVS Repository (archive)*. URL: http://anoncvs.postgresql.org/cvsweb.cgi/pgsql/src/backend/snowball/stopwords/ (visited on Jan. 28, 2016) (cit. on p. 56).

Privax Ltd. (2015). *Which countries block Twitter & which no longer ban Twitter*. URL: http://blog.hidemyass.com/2015/04/02/countries-block-twitter-no-longer-ban-twitter/ (visited on Jan. 4, 2016) (cit. on p. 28).

Rochford, Austin (2015). *Maximum Likelihood Estimation of Custom Models in Python with StatsModels*. URL: http://austinrochford.com/posts/2015-03-03-mle-python-statsmodels.html (visited on Jan. 25, 2016) (cit. on p. 50).

Smith, Craig (2015). *How Many People Use Facebook, Twitter and 1000 of the Top Social Media, Apps and Tools? (Year-End 2015)*. DMR. URL: http://expandedramblings.com/index.php/resource-how-many-people-use-the-top-social-media (visited on Jan. 2, 2016) (cit. on p. 2).

The Apache Software Foundation (2016a). *Apache Spark – Lightning-Fast Cluster Computing*. URL: http://spark.apache.org/ (visited on Apr. 24, 2016) (cit. on p. 139).

The Apache Software Foundation (2016b). *Welcome to Apache Hadoop!* URL: http://hadoop.apache.org/ (visited on Apr. 24, 2016) (cit. on p. 139).

Trim, Craig (2013). *The Art of Tokenization*. IBM. URL: https://www.ibm.com/developerworks/community/blogs/nlp/entry/tokenization?lang=en (visited on Dec. 18, 2015) (cit. on p. 18).

Twitter Inc. (2015a). *Removal requests. Transperancy Report*. URL: https://transparency.twitter.com/removal-requests/2015/jan-jun (visited on Jan. 4, 2016) (cit. on p. 28).

Twitter Inc. (2015b). *Twitter | About. Twitter Usage/Company Facts*. URL: https://about.twitter.com/company (visited on Sept. 30, 2015) (cit. on p. 28).

U.S. Geological Survey (2015). *Magnitude / Intensity Comparison*. URL: http://earthquake.usgs.gov/learn/topics/mag_vs_int.php (visited on Mar. 16, 2016) (cit. on p. 79).

Willaredt, Jonas (2011). *WiFi and Cell-ID based positioning - Protocols, Standards and Solutions*. SNET Project. URL: https://www.snet.tu-berlin.de/fileadmin/fg220/courses/WS1011/snet-project/wifi-cellid-positioning_willaredt.pdf (visited on Jan. 13, 2016) (cit. on p. 33).

WordNet (2015). *About WordNet*. URL: https://wordnet.princeton.edu/ (visited on Apr. 18, 2016) (cit. on p. 98).

List of Figures

List of Tables

Acknowledgments

The work presented in the first part of this thesis was funded by the Center for Disaster Management and Risk Reduction Technology (CEDIM).
The research visit at the University of Melbourne in the Department of Infrastructure Engineering was funded by the Karlsruhe House of Young Scientists (KHYS). During this visit, the main idea and basic work for the second part of this thesis was developed.
Several trips to important conferences where parts of this thesis were presented, was funded by the Graduate School of Climate and Environment (GRACE).

Colophon

This thesis was typeset with LaTeX 2_ε. It uses the *Clean Thesis* as base style developed by Ricardo Langner. The design of the *Clean Thesis* style is inspired by user guide documents from Apple Inc.

Download the *Clean Thesis* style at http://cleanthesis.der-ric.de/.