Masterthesis im Rahmen des

Universitätslehrganges „Geographical Information Science & Systems“

(UNIGIS MSc) am Zentrum für GeoInformatik (Z_GIS)

der Paris Lodron Universität Salzburg

zum Thema

VGI in Disaster Management – Fusing Remote Sensing Data with User-Generated Data for Improved Flood Management

Vorgelegt von:

Mag. Johannes Reiter

GIS 103486, UNIGIS MSc Jahrgang 2014

Zur Erlangung des Grades
„Master of Science (Geographical Information Science & Systems) – MSc(GIS)“

Gutachter:
Ass.-Prof. Dr. Bernd Resch

Erklärung der eigenständigen Abfassung der Arbeit

Ich versichere, diese Masterthesis ohne fremde Hilfe und ohne Verwendung anderer als der angeführten Quellen angefertigt zu haben. Die Arbeit hat in gleicher oder ähnlicher Form noch keiner anderen Prüfungsbehörde vorgelegen. Alle Ausführungen der Arbeit, die wörtlich oder sinngemäß übernommen wurden, sind gekennzeichnet.

Wien, 31. März 2017

Johannes

Acknowledgements

I would like to thank my supervisor Bernd Resch for inspiring me on this topic and his constant valuable hints in developing the methodology of this thesis.

Furthermore, I would like to thank the Center for Satellite Based Crisis Information (German Aerospace Center), especially Alexander Mager and Elisabeth Schöpfer, not only for the provided datasets and for their constant support, but also for the chance to present my results at the DLR-Oberpfaffenhofen. Special thanks go to Florian Usländer, whose hints on the technical implementation were a welcome help.

I would also like to thank Thomas Winklehner, for the weekly discussions on the thesis topic itself, his great advices and the very welcome distractions. Therefore, I would like to thank Marie Röder for the proof reading and her constant motivation and affection.

Finally, I would like to thank my parents Walter and Anna Maria for their great patience and their caring support. Last but not least special thanks to all my friends for supporting me, folks you're the best!

Johannes

Abstract

Various disasters like the severe German flood in 2003 keep demonstrating the vulnerability of human civilisation and infrastructure as well as the critical need of following up the information gap to offer spatial analyses to local decision-makers and international humanitarian mission. In the last few years Social Media services like Facebook and Twitter and their millions of followers have received high attention by research groups in relation to their situational awareness of occurring disasters. Local individuals not only consume but also produce valuable informations about disasters. However, the so-called Big Data, generated via these networks, is very challenging to analyse and to compute in a fair amount of time. While other research studies in the field of Disaster Management mainly concentrated on time-consuming keyword-searches, finding the relevant information, this thesis focused on similarity assessments in the form of a semantic probability-based topic model called Latent Dirichlet Allocation (LDA) (Blei et al., 2003). This unsupervised machine learning model identifies latent topics by clustering co-occurring words from a collection of Tweets. Furthermore, an analysis framework is presented providing the methods of extracting, organising, filtering and analysing the Tweets in near real-time within a developed application. Together with spatiotemporal filtering via the remote sensing data from the Center for Satellite Based Crisis Information (DLR Oberpfaffenhofen) the attempt is made to classify the flood related Tweets with the LDA algorithm. All calculations were assessed by a confusion matrix and further statistical analysis methods. The results of this thesis show that Social Media messages not only could be used for additional information sources on the crisis event itself, but also that LDA provides a stable overview in a fraction of time compared to manual or keyword-based filtering methods.

Keywords: VGI, Latent Dirichlet Allocation (LDA), Disaster Management, Twitter, Collective Sensing, Cascading LDA

Content

Abbreviations

AGI	Ambient Geographic Information
API	Application Programming Interface
DB	Database
DFD	German Remotes Sensing Data Center
DLR	Deutsches Zentrum für Luft- und Raumfahrt
GPS	Global Positioning System
GSM	Global System for Mobile Communications
ICT	Information and Communications Technology
IGO	Inter-Governmental Organisations
LBS	Location-based Services
LDA	Latent Dirichlet Allocation
NGO	Non-Governmental Organisation
NLP	Natural Language Processing
NLTK	Natural Language Toolkit
NPV	Negative Predicted Value
OSM	Open Street Map
SNS	Social Network Sites
VGI	Volunteered Geographic Information
ZKI	Zentrum für Satellitengestützte Kriseninformationen

1 Introduction

Various disasters such as earthquakes, tsunamis and floods – arising in the past and present alike – keep demonstrating the vulnerability of human civilisation and infrastructure as well as the critical need of following up closely the information gap in order to offer spatial analysis to local decision-makers and international humanitarian mission. (Goodchild and Glennon, 2010) These disasters also displayed the massive amount of data generated by local individuals spreading in near real-time over different Social Media platforms (e.g. Twitter, Flickr, Facebook, ...). Many projects already progressed in terms of (Un-) Volunteered Geographic Information (VGI) (Goodchild, 2007) and Location-based Services (LBS) sensor networks (Sagl et al., 2012) subsequently demonstrating the importance of collecting, complementing and analysing these data sources.

Recent studies have proposed that in addition to the pure Social Media contents (text, image, video, URL) their provided metadata is just as important (Albuquerque et al., 2015; Gesualdo et al., 2013; MacEachren et al., 2011). Beside timestamps, it is the geographical reference, which enriches the ordinary messages with additional information about when and where something happens. In the world of Social Media Twitter, Facebook or Flickr are not only the most popular platforms, but also the most researched.

On the one hand, the enormous quantity of data (millions of messages per day) is a big advantage compared to Emergency-Apps like MydisaterDroid , that holds only a few users (Jovilyn et al., 2010). On the other hand, the most challenging part is to compute and analyse these amounts of information (Miller and Goodchild, 2015). The central question then becomes: which messages contain useful information, how is it possible to classify their diverse content and how to speed up the process of extracting useful information? While previous works concentrated mainly on Social Media itself as a stand-alone collection of information, recent studies tried to combine traditional authoritative data (sensor data, remote sensing, hydrological data, etc.), in order to unfold additional benefits (Albuquerque et al., 2015; Fuchs et al., 2013; Peters and Albuquerque, 2015). Following this approach, this thesis tries to investigate the usage of Twitter messages in combination with established remote sensing methods, during the severe floods in Germany of 2013. This thesis is therefore supported by the Center for Satellite Based Crisis Information[1] (ZKI), which is part of the national aeronautics and space research centre of the Federal Republic of Germany (DLR) and provides the RS data and additional knowledge on the flood phenomena of 2013 in Central Europe. Unlike other studies the Twitter messages will not be filtered by keywords but by similarity assessments in the form of a semantic probability-based topic model called **Latent Dirichlet Allocation (LDA)** (Blei et al., 2003). This unsupervised machine learning model identifies latent topics by clustering co-occurring words from a collection of Tweets.

While the RS data displays flood-levels/masks as well as broken roads or buildings, the collected Tweets could provide confirmation, not only by text messages but also by photos, linked videos and URLs. In urban regions the big number of geolocated Tweets, which can be mapped and visualised as point features, may additionally allow for geo-statistical analyses like density-, cluster- or overlay- methods. Moreover, the frequency of Tweets or retweeted messages could be indicators of a valid problem at the ground. A classification of the relevant on-topic Tweets is therefore necessary as well as the spatiotemporal fusion with the water masks (RS data). This paper mainly concentrates on the technical implementations, statistics and workflow rather than the semantic content of the Tweets itself, which were already investigated by several other studies (Albuquerque et al., 2015; Fuchs et al., 2013; Peters and Albuquerque, 2015).

[1] https://www.zki.dlr.de/de (latest access: 02.12.2016)

Floods are highly dynamic disaster events, which can last for months but change their active zone in a few hours. While RS data delivers snap-shots in certain intervals, VGI data could help closing those gaps with near real-time information. The aim of this master's thesis is to explore the benefits of combining established remote sensing data with the concepts of '(A-)VGI' and 'Collective sensing' at devastated regions, gaining therefore more accurate and finer-grained results. While local experts impersonating 'citizen sensors' generating semi-professional observation-records, 'Collective sensing' could improve data analysis in terms of holistic events or information- and communications-technology-Networks (ICT-Networks) (Hawelka et al., 2014; Resch, 2013a; Stefanidis et al., 2013). Furthermore manual classifications of Twitter topics is very time consuming, so that probabilistic topic modelling could speed up this process tremendously without, however, gaining too much topic drift and losing thus even more critical information.

This paper proposes a suitable methodological approach comprising the process of harvesting messages from Twitter (applying a self-made application) as well as fusing the tweeted information with the water masks, in order to finally employ an unsupervised machine learning based classification. In the field of Disaster Management time is a critical factor and so together with high-resolution space- and airborne sensors it is in all probability to gain additional benefits.

This leads to the following research questions:

1. What additional benefits can be obtained from the intersection of 'Collective sensing' with the proven remote sensing data?

2. How can spatiotemporal sequences of events be evaluated to gain quicker information for near real-time operational strategies?

3. How can probabilistic topic models like LDA be used for automated topic classifications in the field of Twitter messages and how reliable are the results?

4. To what degree can the results of the LDA algorithm be improved, if only the flood-related topic is calculated a second time? Keyword: Cascading LDA

Furthermore, this thesis is structured as follows:

Beside this introduction, a brief overview of VGI and its three main concepts is given, as well as the basic concepts of Disaster Management and its interaction with Social Media platforms like Facebook, Flickr or Twitter (**Chapter 1-3**). The author then provides a description of the related studies, followed by the explanation of the case study and the used datasets (**Chapter 5**). In **Chapter 6** the methodology is presented. A Twitter app was registered accessing the company own public Streaming API and collecting a one week sample of georeferenced Tweets in Germany. Furthermore, this data is evaluated against the potential for the usage in Disaster Management, compared to the full access of Twitter Firehose data from 2013.

Chapter 6 also describes the extracting, pre-process and topic modelling (LDA) process of the Tweets, as well as the fusion with remote sensing data. The **Chapters 7** discusses the classification process of the on- and off-topic Tweets, within in the concept of the binary classifier of a confusion matrix.

The last **Chapter 8-9** presents a critical evaluation of the combination of the presented methods and also discusses further research avenues.

2 Volunteered Geographic Information (VGI)

In modern Emergency/Disaster Management, geographic tools and data are essential in all aspects and public authorities or relief organisations spare no effort to gain even quicker and more accurate information. Multiple actors include stakeholders such as inter-governmental organisations (IGO), non-governmental organisations (NGO) and ordinary citizens are in need of critical information. Knowing what, where, when something is happening and who is involved is one of the key aspects of Disaster Management. Until the 1990s creating accurate geographic data and map-production was in the hand of highly trained specialists and companies and their technology were only affordable for few people or organisations. With the development of the Global Position System (GPS) in the 1970s and the opening for civil usage in the 1990s, it became possible for an average citizen to determine their position not only accurately but cost-limited to a single device like an ordinary smartphone or car GPS navigators. In combination with the development of the so-called Web 2.0 people were given the ability to participate their experiences to the whole World Wide Web. At the turn of the millennium web-protocols and technology allowed a much more sophisticated usage of the Internet. Sites, blogs and wikis were constructed and allowed people to populate them with their content, without too much moderation or restrictions of the site owner (Goodchild, 2007).

In cases like Wikipedia people were even allowed to edit the content of others making it one of the leading knowledge platforms worldwide, outclassing even traditional reference works like Brockhaus[2] or the Encyclopaedia Britannica[3]. Furthermore, tools like Google Maps or Open Street Map (OSM) not only provide the availability of getting accurate geographic information, but people are

[2] http://www.sueddeutsche.de/digital/wikipedia-besser-als-der-brockhaus-1.324954 (latest access: 08.05.2016)

[3] http://www.forbes.com/sites/hbsworkingknowledge/2015/01/20/wikipedia-or-encyclopaedia-britannica-which-has-more-bias/#dc7c97d1ccf8 (latest access: 08.05.2016)

encouraged to participate on these platforms, with their local knowledge (Goodchild and Glennon, 2010).[4] With applications like OSM, even non-specialists can make a usable map and adapt it with their own features and information. These applications changed the way people think and used maps and made geographic information affordable for nearly everyone. In combination with cheap GPS-devices, people are given the ability to reference local features with coordinates, enrich them with additional information (text messages, photos, videos, URL, etc.) and link these to online platforms like OSM, Flickr, Twitter and much more. VGI-platforms have become very popular since most of the global social networks have the ability to georeference their content. From microblogs like Twitter or mapping platforms like OSM, there are many different types of contributing VGI. Humans are becoming intelligent sensors in a global aspect (Goodchild, 2007; Resch et al., 2010).

2.1 People/Humans/Citizens as Sensors

As the title implies, different terms have been introduced which more the less mean the same topic and were established by several authors. 'Citizens as Sensors' (Goodchild, 2007) , 'Humans as Sensors' (Forrest, 2010) or 'People as sensors' (Resch, 2013b) (as it is further called in this thesis) is an integral part of VGI as it describes a measurement model '*in which measurements are not only taken by calibrated hardware sensors but in which also humans can contribute their subjective 'measurements' such as their individual sensations, current perceptions or personal observations. These human sensors can thus complement—or in some cases even replace—specialised and expensive sensor networks*' (Resch, 2013b).

Resch separated the idea of people as sensors into three concepts 'People as sensors', 'Citizen Science' and 'Collective sensing' (**see Fig. 1.**)

[4] OpenStreetMap is an international effort to create a free source of map data through volunteer effort. https://www.openstreetmap.org/ (latest access: 08.05.2016)

People as Sensors

Within this concept, people record their personal and subjective observations (like weather, security, traffic, …) and submit that information through mobile or web applications. Popular platforms which use that concept are Ushahidi[5], MyDisasterDroid (Jovilyn et al., 2010), OpenTraffic[6] or SensePlace 2[7] and contributed high amounts of specialised data for different purposes. Subjective measurements of peoples, of course, cannot replace technical sensor networks in cases where precise measurements are required. On the other side, these observations can complement existing technical sensor networks or can provide additional information. While RS data delivers snap-shots in certain satellite intervals, VGI data could help closing those gaps with near real-time information. Also in cases of foggy or cloudy skies, people can submit on the ground information.

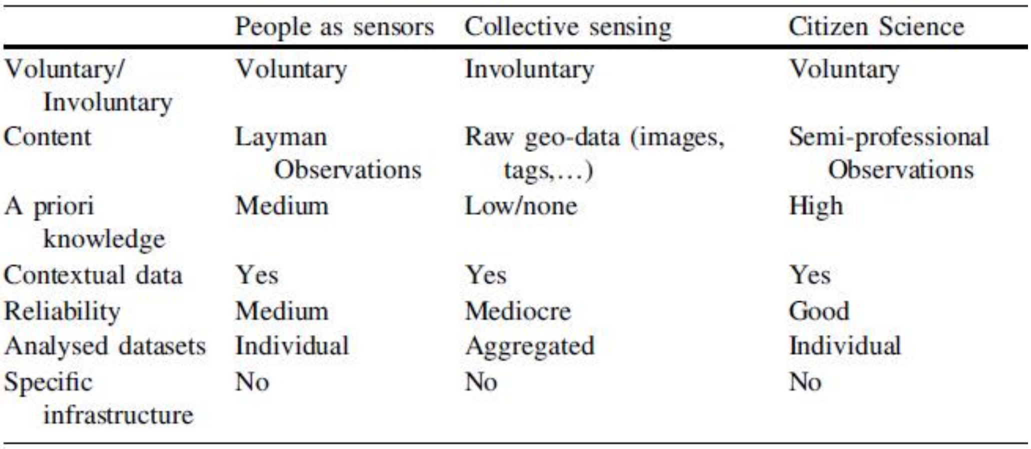
Citizen Science

This concept is similar to the ‘People as sensors’ model and can be seen as subpart. Individuals share their local knowledge and expertise for a certain aspect and therefore become citizen scientists. They combine traditional sensor networks with mobile or web applications, and so people are encouraged to take part in collecting and sharing measurements of their environment. People can help to evaluate their environments like air and water pollutions and share their local knowledge with researchers who therefore gain a much broader database. There are also positive synergetic effects because scientists get cheap near real-time datasets and the submitters are encouraged to not only participate to improve their environment but also to sharpen their perception of the surrounding area. In contrast to the ‘People as sensors’ the a priori knowledge has to be much higher than in the original concept but the reliability, the quality of the data and the contributor trustiness, is also positive correlated (Resch, 2013b).

[5] https://www.ushahidi.com/ (latest access: 02.01.2017)
[6] http://opentraffic.io/ (latest access: 03.01.2017)
[7] https://www.geovista.psu.edu/SensePlace2/ (latest access: 12.12.2016)

Collective Sensing

This concept is the fastest growing of the presented as it analyses aggregated data from collective networks, such as Twitter, Flickr, Facebook or anonymised data from mobile phone networks. In contrast to ‘People as sensors’ or ‘Citizen Science’, ‘Collective sensing’ uses existing ICT networks to generate contextual information. One of the main advantages is that no specialised smartphone apps or web applications are required to generate data (Resch, 2013b). For instance, if many messages on Twitter talk about a road accident or natural hazards this information can provide additional information for relief funds or Emergency Management. Other examples for ‘Collective sensing’ are earthquake detectors based on Twitter data, where recent studies showed that even expansive and complex sensor networks would not get faster results (Crooks et al., 2013). Another difference to the previous concepts is the involuntary character of the data acquisition. People mostly do not know that their data from Twitter or Flickr is used for different research purposes. One could argue that the Twitter is per se predetermined for public, but users are often not aware. Furthermore, it is important to treat privacy concerns with respect. Beside the ‘Collective sensing’ approach Stefanidis (et al., 2013) coined the term Ambient Geographic Information (AGI), which has many synergies to the former concept (**see Chapter 3**).



	People as sensors	Collective sensing	Citizen Science
Voluntary/ Involuntary	Voluntary	Involuntary	Voluntary
Content	Layman Observations	Raw geo-data (images, tags,…)	Semi-professional Observations
A priori knowledge	Medium	Low/none	High
Contextual data	Yes	Yes	Yes
Reliability	Medium	Mediocre	Good
Analysed datasets	Individual	Aggregated	Individual
Specific infrastructure	No	No	No

Figure 1. Comparison of the three concepts according to Resch (2012)

2.2 Data Quality and Trustworthiness of (A)-VGI

One of the main issues of VGI and its concepts like People as sensors is the data quality and the trustworthiness of their contributors. As mentioned earlier people do not have the same perception of their environment, and as a result of it the data differs and lacks objective measurements. There are several approaches to improving this uncertainty factors, but the degree of automation varies. In some cases, a full automated process will work, while in others the human factor of contribution is required. Godchild introduced three main concepts of quality assurance for VGI data (Goodchild and Li, 2012):

Crowdsourcing

This concept is based on several meanings. On the one hand, it is assuming that a group of people can identify, validate and correct errors that individuals make. On the other hand, some problems can only be solved at the first place if they were processed by a group, which in some cases can outperform even experts. Some researchers have also thought about to compare it to Linus Law, named after the famous developer of Linux, Linus Torvalds, who said '*given enough eyes, all bugs are shallow*' (Raymond, 1999). Granted this approach is true in many cases like Wikipedia[8]. Nevertheless, it cannot be derived probably to services with a geographical component like Wikimapia[9] and Humanitarian OpenStreetMap[10], or only in parts. Prominent geographic features will be identified and validated correct, while latent or obscure features in less known parts of the earth might not.

[8] www.wikipedia.org (latest access: 05.04.2016)
[9] www.wikimapia.org (latest access: 05.04.2016)
[10] https://hotosm.org/ (latest access: 05.04.2016)

The Social Approach

In this concept, the users are separated into more or less trusted contributors, who act as moderators. Prominent platforms are Wikipedia, Wikimapia or OSM, but their hierarchy system is implemented in different ways. Some reward high contributions others implemented some sort of reputation system to become an advanced user or moderator. It can be observed that in all of these platforms a few individuals are responsibly for most of the data input. For example, the evaluation of the OSM contributors in the area of London over six years showed, that twenty users are responsibly for more of the half of all entries (Mooney et al., 2010). Furthermore, there are also some sorts of hybrids which combine classical authoritative data with volunteered user contributions. Google's Mapmaker falls into this category, as it reserves the top tier editing tools to their employees (Goodchild and Li, 2012). There are several other companies like TomTom who fall back upon volunteers, but their geographic features are only accepted/trusted from their own experts in a second step.[11]

The Geographic Approach

This approach evaluates purported geographic facts if they are true or false based on the broad body of geographic knowledge. It is based on rules, while the most important are the First Law of Geography, *'All things are related, but nearby things are more related than distant things'* (Tobler, 1970). For this approach, it is very important that the purported facts be consistent with its geographic context or vicinity. For example, a report of a flooded area is more likely to be true if floods have already been reported nearby. The geographic approach can be extended to hundreds of rules and could become therefore very complex, while the implementation in a rule-based system is very challenging. Research has to be

[11] https://www.tomtom.com (latest access: 20.07.2016)

performed to combine VGI with geographic knowledge and build a proper framework around for an effective implementation. (Goodchild and Li, 2012)

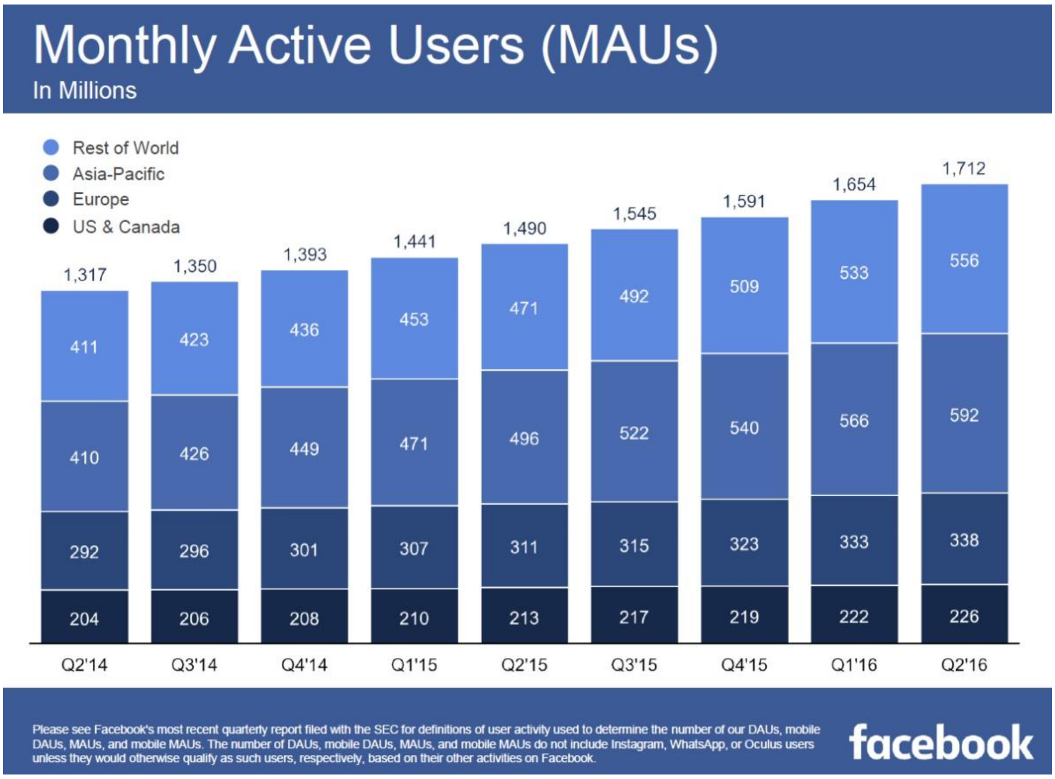
2.3 Data Quality of Georeferenced Tweets

There are many temporal und semantic uncertainties in analysing Twitter data. Geospatial accuracy is depending on the mobile device dependencies and the GPS signal itself. Furthermore, urban environments like high buildings are not novel to disturb the data quality and precision.(Zandbergen and Barbeau, 2011) Another important uncertainty is that the amounts of Tweets are not equally distributed and is mainly concentrating in highly populated areas. As an example of the later used data, the city of Berlin is responsible for one eight of all georeferenced Tweets in Germany, while there are big rural areas in Brandenburg where only a few Tweets are posted. (**see Fig. 10**.) This can lead to an over-, underrepresentation or even an exclusion of whole population groups. (Miller and Goodchild, 2015) The same phenomena can be observed with OSM data, and one of the main differences between VGI and commercial data sources is that the quality of the VGI datasets decreases considerably as the distance from the dense populated areas increases (Neis et al., 2011). In the case of Social Media, the semantical uncertainty of the Twitter speech and the unknown a priory activity of users are responsibly that they might be a weak indicator for real-world observations. (Steiger et al., 2015b)

It is also important to emphasise that only a very small number of Tweets are georeferenced and this differs further by country. Several studies have estimated a percentage of 2-4% of georeferenced messages for the United States and only around 1% for Germany. (Fuchs et al., 2013) The later cannot be affirmed in total by this thesis, but the measurements in **Chapter 6** suggest not that much more than 2% for Germany.

2.4 Social Network Sites (SNS) and their Spatial Factor

Social Media services like Twitter, Facebook, Google + or Instagram have become very popular around the world with billions of users. To understand the phenomenon Social Media, the expected data and penetration rates a brief review is given. The official stock market report of Facebook shows that by now it is the biggest social network site in the world, with around 1.712 billion monthly active users (**see Fig. 2.**). Also impressive is that 967 million people use Facebook only on their mobile device and 1.5 billion monthly active mobile users. [12]



Monthly Active Users (MAUs)

In Millions

Please see Facebook's most recent quarterly report filed with the SEC for definitions of user activity used to determine the number of our DAUs, mobile DAUs, MAUs, and mobile MAUs. The number of DAUs, mobile DAUs, MAUs, and mobile MAUs do not include Instagram, WhatsApp, or Oculus users unless they would otherwise qualify as such users, respectively, based on their other activities on Facebook.

Figure 2. Facebook Monthly Active Users in Millions

As speaking of 'global', it is very important to point out that there is still a so-called digital divide between the well and lesser developed regions of the world. The degree of participating in the information and communication technologies is determining the access to socio-economic prosperity. Against this background

[12] https://s21.q4cdn.com/399680738/files/doc_presentations/FB-Q216-Earnings-Slides.pdf (latest access: 05.07.2016)

the fast rising ICT markets in Asia and Africa are decreasing the global gap and are responsible for that even in the poorest regions of the world like sub-Saharan Africa, there is a penetration rate of 43% having mobile access to the internet. On a global scale, it is at least 63% of around 4.7 billion unique subscribers in 2015 and an estimated 5.6 billion subscribers in 2020 (GSMA report 2016).

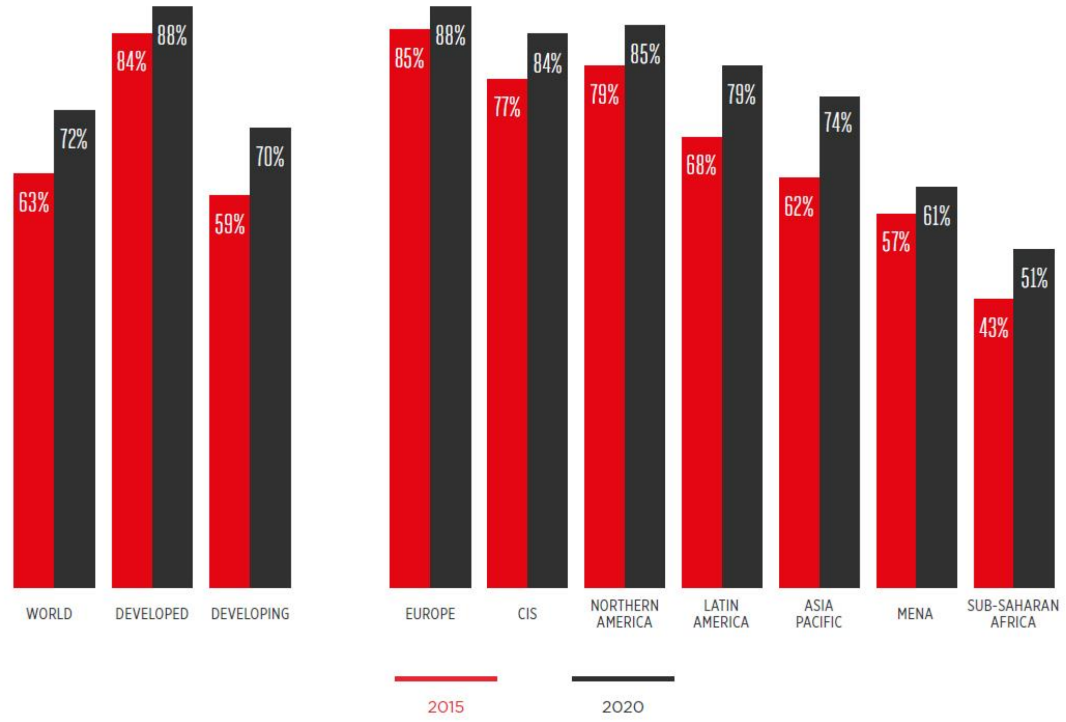


Figure 3. Unique subscriber penetration by region (GSMA 2016)

With this overwhelming numbers and the fast-growing rates of 15-20% per year it is now clear that in combination with GPS devices like in most common smartphones, SNS services have become a huge spatial factor. Furthermore, it is possible to geocode normal messages or photos on Facebook or Twitter, and therefore the communication becomes location based. The ubiquity around the world, even in poorly developed countries is one of the most exciting aspects of

SNS. Phenomena like the so-called Arabian Spring[13] or movements against TTIP[14] would have been inconceivable, without self-organising peoples over Social Media. The high amount of personal user data is a great opportunity in the field of VGI, but it also contains many risks.

2.5 Ambient Geographic Information (AGI)

For the first time, it is possible to observe human activities in unprecedented scales and to research new concepts of social interaction. In combination with traditional mapping or GIS solutions, these data sources are becoming highly popular especially in the fields of disaster management. Despite scale or resolution, there is the main difference between traditional VGI platforms like Wikimapia and Social Media sites like Facebook or Twitter. Firstly, people on e.g. Twitter mostly do not have the intention to generate geographic features for additional geographic knowledge or databases. Secondly, they are often not aware that their data is used secondarily for different studies or even commercial applications. Stefanidis(et al., 2013) argues that data acquired from these sources represents a deviation from volunteered geographic information because of its ambient character. As it is not the primary intention of the user to participate in the way of traditional crowdsourcing or citizen sensing, he coined the concept as Ambient Geographic Information and sees it as '*a second step in the evolution of geospatial data availability, following on the heels of volunteered geographical information*'. In contrast to VGI, AGI focuses on passively contributed data, and there are many synergies with Resch's (2013b) concept of 'Collective sensing', which is an integral part of this thesis.

[13] https://www.theguardian.com/world/interactive/2011/mar/22/middle-east-protest-interactive-timeline (latest access: 12.12.2016)

[14] http://www.bmwi.de/DE/Themen/Aussenwirtschaft/Freihandelsabkommen/TTIP/was-ist-ttip.html (latest access: 12.12.2016)

2.6 Twitter

The microblogging service gained worldwide popularity in the past years from 30 million active users to average 313 million users worldwide in the second quarter of 2016. Similarly to Facebook, Twitter has a very high rate of mobile users and the figures reveal that 82% of the people use the service from their mobile device.[15] Despite the facts that Twitter is slowly – but constantly – growing in the last years, the overall rate of messages (Tweets) per day could be declining as several application developers measured. While the all-time peak of Tweets was reached in August 2014 with nearly 661 million, the average number of Tweets per day decreased to 303 million Tweets in January 2016. It must be noticed that these data was measured via the Twitter own API by app developers and not published by Twitter itself. Furthermore, there is an ongoing dispute how the figures should be evaluated.[16] While the author can see the point of the critics, it has no surplus benefit for this thesis to go too much into detail about the user and Tweet statistics. Despite these discussions, the figures show that there is at least an average of 300 million assured Tweets per day and maybe much more relying on Twitter company announcements.

However, **Figure 4** shows some important statistics, what is Twitter all about.

[15] https://www.sec.gov/Archives/edgar/data/1418091/000156459016021918/twtr-10q_20160630.htm (latest access: 06.09.2016)

[16] http://www.businessinsider.de/tweets-on-twitter-is-in-serious-decline-2016-2?r=UK&IR=T (latest access: 06.09.2016)

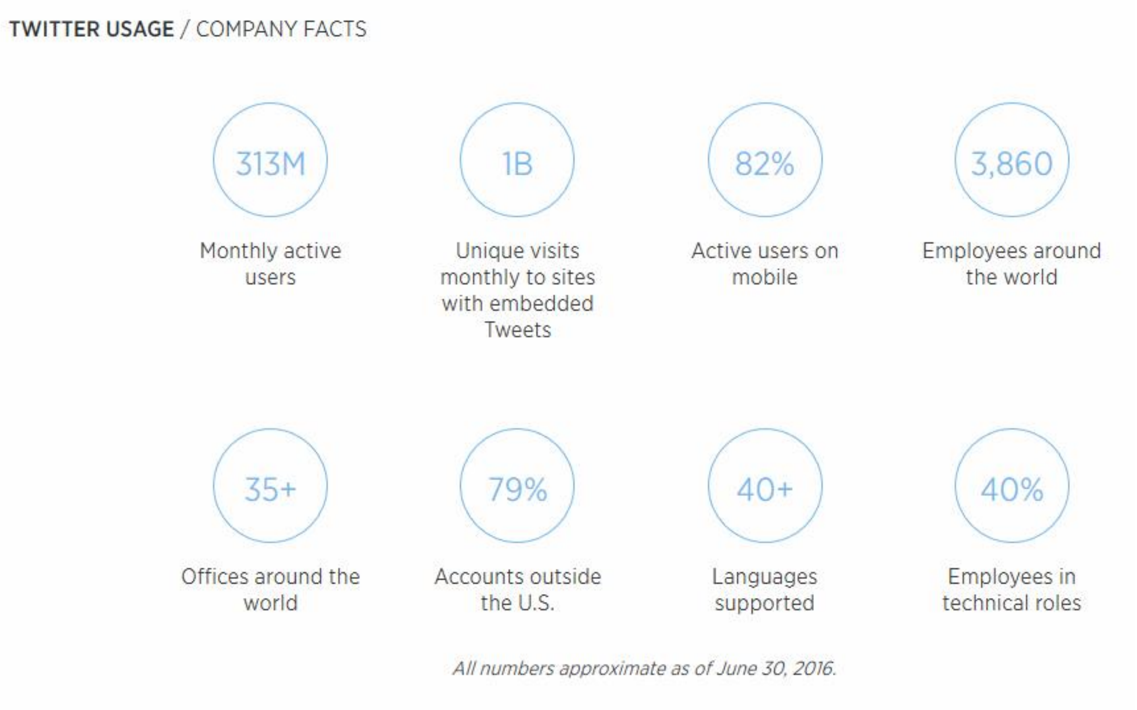

TWITTER USAGE / COMPANY FACTS
313M
Monthly active users
1B
Unique visits monthly to sites with embedded Tweets
82%
Active users on mobile
3,860
Employees around the world
35+
Offices around the world
79%
Accounts outside the U.S.
40+
Languages supported
40%
Employees in technical roles
All numbers approximate as of June 30, 2016.

Figure 4. Official Twitter usage/company facts

2.7 The Twitter Message Format

The microblogging service offers the user to share messages with others within their web application, mobile devices or third-party application based on the Twitter API.[17] A single Tweet consists of a maximum of 140 characters and is per default publicly visible, but the user is given the ability to restrict access to e.g. their own followers. The messages can also contain web links, photos or videos. Because of the popularity of the service, the community coined several semantic terms, which took place in our everyday language. For the sake of completeness, some of the terms are referred. Messages are called 'Tweets', and it is possible for other users to share and forward those Tweets, which are then called 'Retweets'. Twitter users can group their postings by topic with so-called 'Hashtags', which are words, phrases or sequences of characters that are prefixed with the '#' sign. Another prefixed sign is '@' for replying or mentioning to a user.[18] One of the primary purposes of Twitter is to connect people and give them the ability to debate whatever they like from their everyday life to their political opinions. It

[17] https://dev.twitter.com/streaming/overview (latest access: 05.12.2016)
[18] https://support.twitter.com/articles/473379# (latest access: 10.06.2016)

can be used as chat-tool or to brief their surrounding in a more self-portrayal way. Furthermore, not only individuals use these features but also companies, enterprises or organisations. It has become unthinkable for modern organisations to not share their activities on Twitter and most of the companies operate their own Social Media department for image and marketing purposes. Furthermore, Twitter is well known for its near real-time character. From modern news agencies to automated weather bots there is a broad variety of contemporary data, which makes Twitter a good platform for many research cases like event detection. The broad diversity of the posted topics leads to many different applications for different purposes. From influenza illness surveillance (Culotta, 2010; Gesualdo et al., 2013) to earthquake detection (Crooks et al., 2013; Sakaki et al., 2010) or stock market indicator prediction (Zhang et al., 2011) to detecting forest fires (De Longueville et al., 2009), Twitter has become the probably most researched Social Media platform.

3 Disaster Management

Disasters have been part of the humanity since the beginning and people, therefore, try to handle all kinds of hazards, but there is a slight definition difference between hazards and disaster. The first can occur everywhere but to become a disaster it has to harm human population and their legacies. Emergency Management, Disaster Management or Disaster Risk Management refers to all coordinated measures before, during and after a disaster event. Since the 1930s authorities and researchers developed four to eight phases to help to describe and understand ongoing processes of emergency management before four phases became the standard up to now (Neal, 1997). If Disaster Management can be described as an ongoing process before, during and after an event, this master's thesis concentrates mainly on the phases of response, recovery, and their subassemblies. These four phases can be found in every modern training book or on most of the authorities' homepages concerning disaster management. (Baharin et al., 2009)

- **Mitigation and Prevention**[19]

The first phase tries to minimise the impact of the disaster. It includes any activities that prevent or reduce the chance of an emergency. The mitigation process takes place before and after a crisis and can be described as the reduction of vulnerability of peoples and communities, injury and loss of life and property.

- **Preparedness**

The second phase tries to develop emergency training, warning systems and to prepare for the worst scenarios. It is a process of identifying the personnel training and equipment for every potential risk and harmful incidents. Normally,

[19] There are organisations like the National Fire Protection Association, who separate between mitigation and prevention as two discrete phases, but most of the modern agencies summarize it. (Baird, 2010)

this phase involves all levels of government, non-governmental organisations up to the private sector to identify and determinate vulnerabilities or threats.

- **Response**

The third phase tries to respond to the actual disaster impact, which includes search and rescue missions or emergency relief efforts. It uses all strategies and plans from the second phase to preserve the health and safety of the communities and their property. The phase also includes basic human needs, medical care, emergency shelters or the protection of the environment.

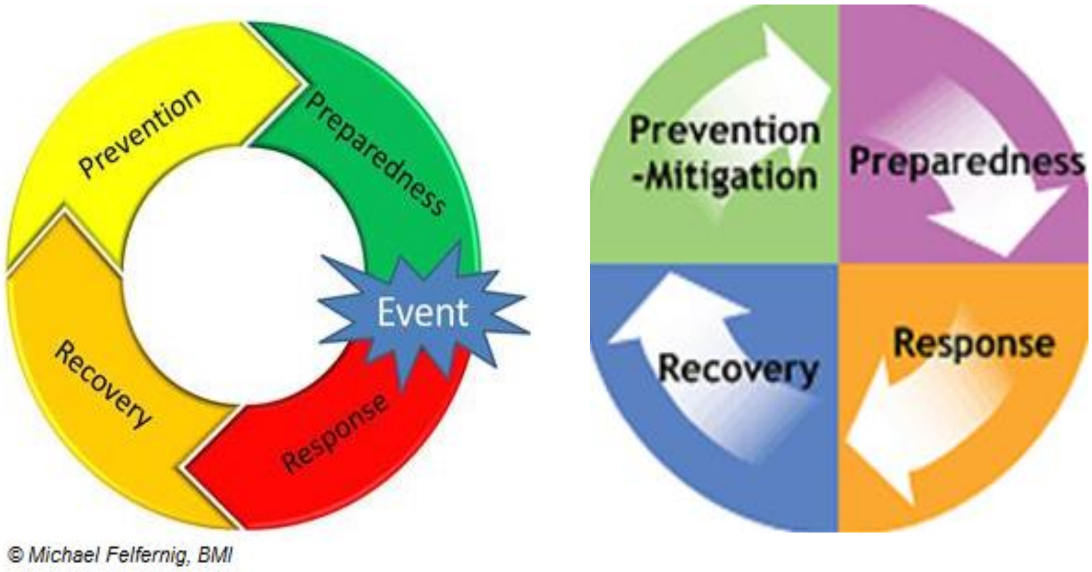
- **Recovery**

The last phase tries to bring the community back to normal. Typical measures for this phase are temporary housing, clean-up efforts and to restore institutions and relevant infrastructure. The actions often extend long after the disaster event and try to include mitigation processes to prevent future harm.

The first and the last phase show that we cannot think about clearly separated phases of emergency management. There are interrelationships between all four phases, and as a result, most organisations visualise the phases as an overlapping circle (**Fig. 5a**[20] and **Fig. 5b**[21]). The topic floods, as in this thesis, can be a good example how the right mitigation plans directly influence and improve the respond phase, because if development in flood plains is restricted it probably reduces the challenges in the response phase. (Baird, 2010)

[20] http://www.bmi.gv.at/cms/BMI_Zivilschutz_en/management/start.aspx (latest access: 22.06.2016)

[21] https://em.countyofdane.com/mitigation_plan.aspx (latest access: 22.06.2016)



© Michael Felfernig, BMI

Figure 5a-b. Emergency Management Circles

3.1 The Spatial Factor in Disaster Management

Gaining spatial information is very important for all phases of Emergency Management and the role of GIS cannot be overestimated. From risk assessment and the development of mitigation plans to preparations and quick response after the impact or the reconstruction plans, geographic information plays a central role. In adaption to the Disaster Management circle before, Cova (1999) added the most important spatial elements to the circle (**Fig. 6**). Under these circumstances, the phases have to be extended relating to their spatial component.

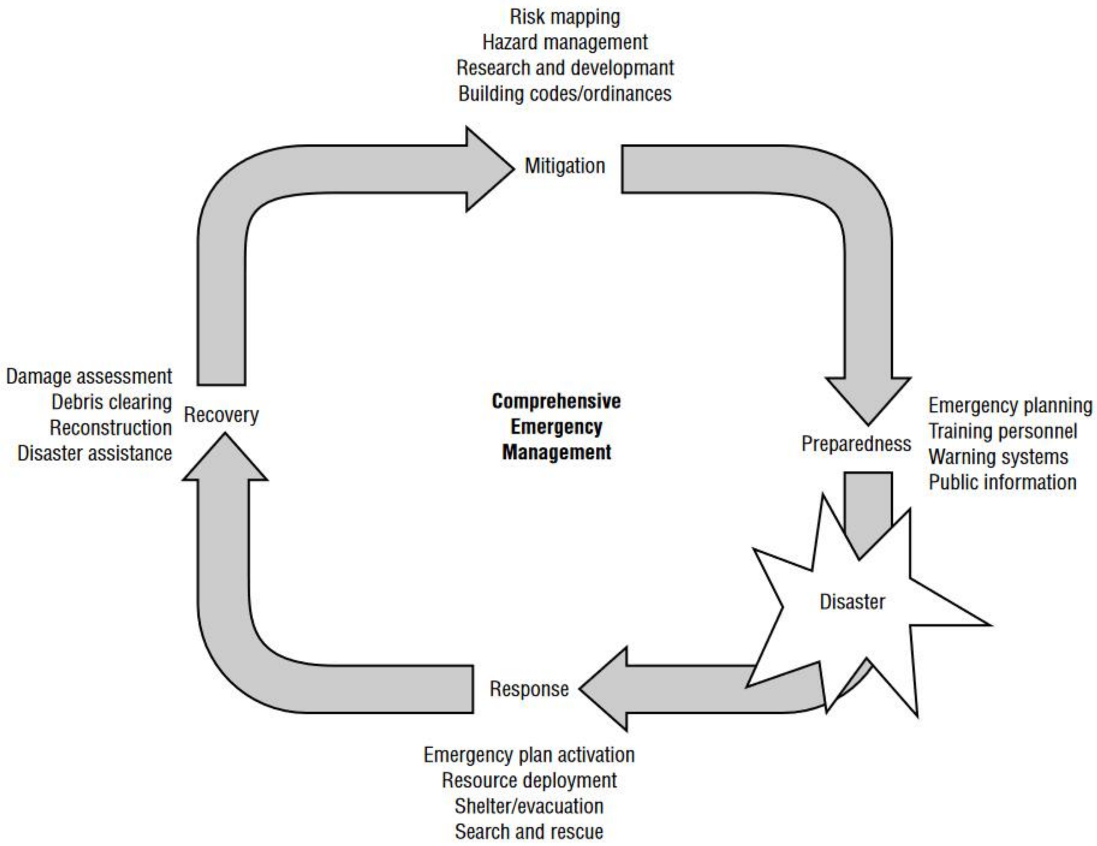


Figure 6. Examples for GIS tasks in the traditional Disaster Management circle by Cova (1999) and Godschalk (1991)

The mitigation and prevention phase is set temporary between two disasters and aims to calculate risks and vulnerability. GIS has the role of analytical models to create long-term assessments, forecasting or management. Typical GIS applications are vulnerability assessment – and natural hazard assessment mapping. Another important research field is to analyse how to mitigate the effects of a particular hazard phenomena like floods or fires or to develop strategies to reduce the vulnerability of a community to a certain hazard like hurricanes or earthquakes. Therefore, the environment, the population and the hazard itself can be seen as layers furthermore combined with risk maps. **Figure 7** shows a good example for modelling the concepts of vulnerability, hazards and risks. (Cova, 1999)

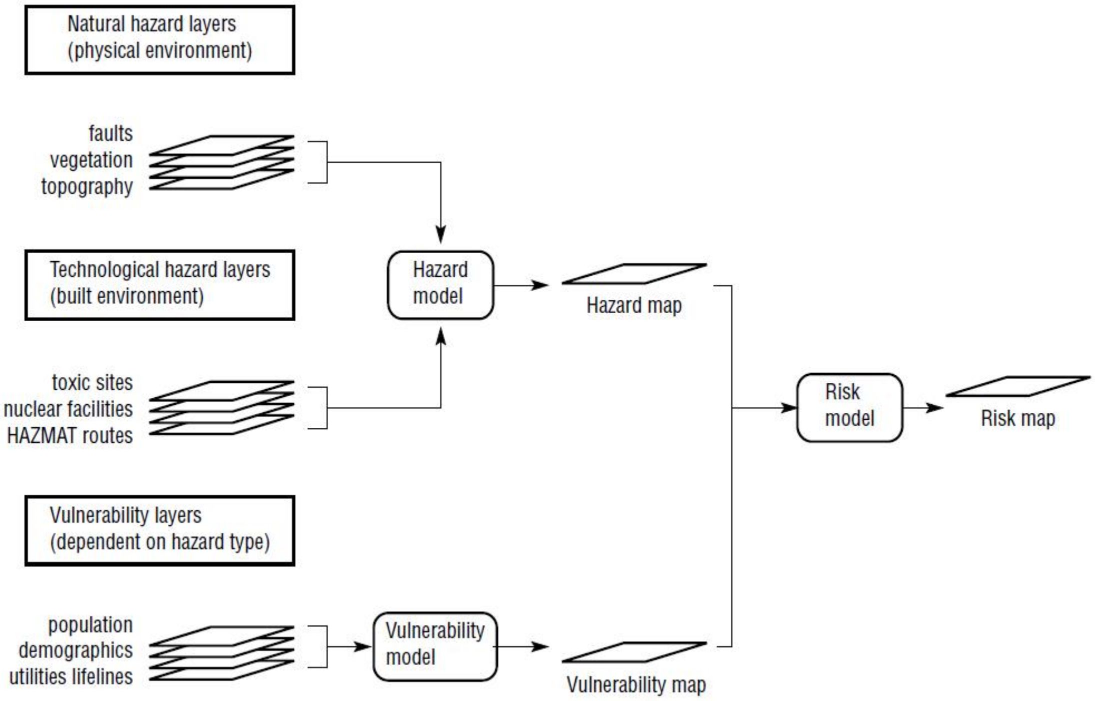


Figure 7. Natural hazard layers (Cova, 1999)

The preparedness and response phases are the most critical one because relief organisations need to know where an event is occurring and who is affected. They need to know the impacted area as well as the estimated losses. The calculation of emergency routes or plans of the evacuating area is very important and sometimes has to be extemporised in near-real time. Disaster Management also has a socio-economic aspect. In contrast to wealthy communities, who often plan and build stronger infrastructure and houses in safer areas, poorer people do not have the possibility and therefore suffer more often from disasters (Khan et al., 2008). An example would be the earthquake of Haiti in 2010, where large parts of the capital Port au Prince collapsed because in this poor city many people cannot afford concrete houses (MacEachren et al., 2011). Furthermore, it is a vicious circle because poor people often do not have the money to rebuild their homes or settle in a safer place. In this described context VGI could help to improve outcomes of traditional Disaster Management tasks.

4 Related Work – Disaster Management and Twitter

In the research field of Disaster Management and (A-)VGI the microblog Twitter is by far the most studied platform probably due to the good availability of data, the shortness of the messages and its near the real-time character. The following chapter tries to give a small overview of the latest research approaches and their objectives. While the concepts are described separately, they are often combined. Furthermore, current research gaps are to be highlighted.

In the topic of analysing Twitter messages, it is a common practice to filter the Tweets that refer to the disaster event by keywords (Landwehr and Carley, 2014). Most of the time keyword-lists relating to the particular events are created. In the case of floods a typical list would include the terms ‘flood’, ‘high-water’ and ‘sandbags’, or - in the event of an earthquake - the terms ‘quake’ and ‘earthquake’ (Albuquerque et al., 2015; Crooks et al., 2013). Many recent studies combine the keyword-based filtering with authoritative data like flood masks or sensor data from gauging stations (Fuchs et al., 2013; Peters and Albuquerque, 2015). Admittedly it is a very fast and straightforward method of gaining event related information and to classify the messages on- and off-topic. Therefore, it reflects the hashtag system of Twitter, that is also organised by keywords. On balance it also has some major drawbacks. Firstly, every keyword list has to be adapted for every single event and disaster type. (Hurricanes, earthquakes, floods, ‘all kind of disasters’ …) Secondly, many Tweets containing critical information are missed, because the poster did not write or misspelt a specific keyword (**see Fig. 12.**). For this reason, Peters and Albuquerque (2015) added the Levenshtein distance to their study parameters, which calculates the cost of two strings, in order to convert in the second one. On that account misspelt words or typing errors can be considered as well. However, Twitter-messages may also carry spatial and/or time-related information, since local individuals often create their own hashtags such as ‘#Elbeflut’ (in our dataset), ‘#SandyNyc’, ‘ or ‘#frankenstorm’ for the Hurricane Sandy, which hit New York in 2012 (Imran et al., 2013). Understanding of the data is therefore very crucial because Twitter hashtags change very quickly and can be very goofy. The above-mentioned '#frankenstorm', for example, is a hint on the Hurricane Sandy hitting New York on Halloween. These spatiotemporal semantics have to be taken into account when a keyword search is carried out. As a subpart of keyword searches, some studies filter the messages by hashtags, referring to the very popular Twitter Trending topics sites. However, it is not recommended to do so without a supporting keyword list. In our sample extracted from the public Twitter Streaming API **(see Chapter 6)** only 49.1% of the harvested Tweets dispersed over the period of a week included hashtags. Thus, it is in all probability to miss many on-topic messages.

One of the most popular topics in analysing Tweets is event- and disaster-detection. As the near real-time character of Twitter is evident, researchers such as Sakaki et al. (2010) and Crooks et al. (2013) tried to detect and estimate the trajectory of earthquakes. Crooks et al. (2013) showed the effect of the earthquake on 23rd August 2011 in the United States on Twitter with impressive figures. From a one percent sample his results revealed, that within two minutes approximately 100 accurately geolocated Tweets were sent off, while nearly 1,000 similar Tweets followed within five minutes. Extrapolating this random sample to the whole Twitter traffic, it is reasonable to expect around 100.000 geolocated Tweet reports within the first five minutes after the earthquake. Beside these numbers, the reporting system USGS DYFI collected 125.000 reports within eight hours (Crooks et al., 2013). In contrast to the earthquake of the East Coast, Fuchs et al. (2013) showed that event detection did not work for the severe floods in Germany of 2013. The main reason is the much lower density of Twitter usage compared to the United States and that only a small number of Tweets related to the floods were sent off (**see Fig. 14).** Secondary both disaster types, flood and earthquake, can hardly be compared, because of their time rate of change. While an earthquake hits an area within a very short time interval at a specific place, floods can last for months. Regarding density and quantity of Tweets there is, however, good news for german Twitter, because aforementioned human-sensor-networks densify exponentially, due to the heavily growing smartphone market. Therefore, the importance of developing new algorithms to successfully extract the signal from the noise becomes apparent, as supposed by this paper. A very similar study tried to analyse the impact of location based Social Media sites in comparison with specialised VGI apps in the use case of a forest fire near Marseille in France (De Longueville et al., 2009). The researchers showed that in contrast to the earthquake in the US the fire was first reported by media and two hours later on Twitter. In this case, citizens and aggregators used the microblog for sharing event related data and summing up already known facts. The main difference between the two studies rests in the density problem and the discrepancy between different disaster types. Forest fires occur per definition in sparsely populated regions and are therefore communicated less than earthquakes, which can be sensed over dozens of kilometres. It seems that not only the difference in population density between Europe and the US plays a key role, but that it also depends on the event how many people are tweeting and for which purpose. Therefore not only the sheer numbers of Tweets matters as well as the relation between the users and their own social network (Bakillah et al., 2015). There is a clear evidence that most of the Tweets correspond spatially with the disaster event itself. This perception is approved by the former studies of Albuquerque and Peters(Albuquerque et al., 2015; Peters and Albuquerque, 2015). They demonstrated that flood-related Tweets were 11 times more likely to occur near (<10 km) flood affected areas in Germany than 30 km away. Kryvasheyeu (et al., 2015) and his team take the same line by analysing the Twitter data before, during and after Hurricane Sandy. They showed that geo-location of users within or outside of the affected area play a significant role.

5 Case Study: The German Flood in May/June 2013

5.1 Meteorological Development

The heavy rainfalls of 2013 in Central Europe and the subsequent dramatic floods form the thematic basis of this master's thesis. Many meteorological and hydrological factors play a role in the development of large-scale flood events such as in June 2013. In May there were hardly any changes between two stable high-pressure areas over the eastern Atlantic/Western Europe and those above the White Sea. This meteorological constellation had determined the weather for weeks and was responsible for one of the biggest flood scenarios in the modern history of Germany (Stein and Malitz, 2013). Two gravure areas can be specified as responsible for the heavy rainfalls. The first depression development started on the 29th May above the northern Balkans, and on the 31st May 2013, the low-pressure area was above the north of the Czech Republic. From there, it migrated south-westwards to the south in the direction of the Alps, where it dissolved on the 2nd June. At the same time, the second low-pressure area was analysed prior over Poland on the 1st June and moved to Eastern Europe on the 3rd. On the same date, the rainfalls declined. (Thieken and DKK, 2015)

Around these depths, warm and especially humid air flowed from the south of Europe towards Germany. There, the labile-stratified air, which had a large enough liquid water content, glided on the much cooler air masses with the northern stream on the edge of the Atlantic high to Germany. In May, rainfall fell within an average of one and a half to two months, whereby the soils could hardly absorb more water. At the end of May, about 40% of the surface area of Germany had such high soil moisture values, which have never been observed since the start of measurements in 1962. From Thursday, the 30th May until Sunday, the 3rd June there was a weather situation with intensive continuous rain, which subsequently resulted in considerable floods in many parts of Germany. In a broad stretch from southern Schleswig-Holstein to northern Bavaria, 250% of the monthly rainfall was reached, in some landscapes even more than 300% (**see Fig. 8**.) (DWD, 2013).

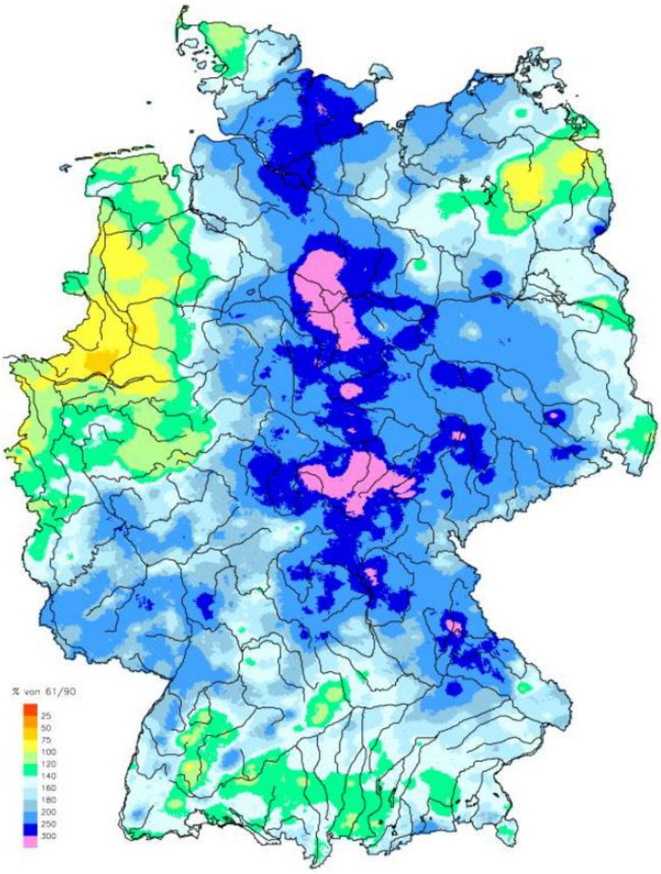


Figure 8. May 2013, height of precipitation in percent of the quarterly mean 1961-1990 (DWD, 2013)

The contribution of the snow covers, as a natural result of the annual snowmelt, played a comparatively small role in the formation of the floods and was only about 5%. The main focus of the storm events was in Bavaria and Baden-Württemberg, Saxony and Thuringia. Also the eastern parts of Hesse and southern parts of Lower Saxony were affected. The official report of the DWD shows a 10% increase of humidity compared to the 2002 flood (Stein and Malitz, 2013). With an average of 127 l/m^2 in Germany, it reached 178% of the perennial average. It was the wettest May since the start of measurements in 1881.[22]

In summary, it can be said that from the hydrological point of view, the flood has arisen due to two factors: First, there were high precipitation volumes due to several days of continuous rainfall, especially in the south-eastern half of Germany. Second, the exceptionally high humidity in large parts of Germany as a result of heavy rainfall. (Schröter et al., 2015)

5.2 The Centennial Flood of 2013

In Germany the flood of June 2013 loomed over a period of several days in the river basins, whereby long-stretched floodwaters with large discharge volumes emerged from the rivers. Virtually all regions close to watercourses were affected by the floodwaters accumulating from north-west to south east. The flood spread in several waves along the rivers, which partially overlapped and thus increased. For a better division the German waterways and river basins are classified into three main regions by virtue of their connections and topography:

1. Rhine/Main/Weser

2. Danube/Inn/Isar

3. Elbe/Saale

Furthermore, the classified regions are described in order of their appearance, however, the periods of flooding overlap in many cases.

As a hypothesis, this work assumes that because of the near real-time character of Twitter the spatiotemporal distribution of the Tweets talking about die floods should correlate approximately with the occurring water peaks. For this reason,

[22] http://www.dwd.de/DE/presse/pressemitteilungen/DE/2013/20130529_DeutschlandwetterimMai.pdf?__blob=publicationFile&v=3 (latest access: 01.10.2016)

the chronological sequence of the individual flood events has to be addressed particularly.

5.2.1 Rhine/Main/Weser

The first floods were measured along the Weser on 28th May and in combination with Werra and Fulda. The high waters in the Weser region continued until the 13th June. Along the Rhine, the flood that came from the Alps was measured on the 1st June in Rheinfeld. On the 2nd June, the crest was observed in Maxau and overlaid on the 3rd June with the floodwaters coming from the Neckar. On the same date, another flood crest was coming from the river Main, which was joined by a second wave three days later. The situation was different with Mosel, Lahn and Nahe, which did not bring any appreciable water increases to the river Rhine (Thieken and DKK, 2015).

5.2.2 Danube/Inn/Isar

In the Danube basin the floods had a lot more impact and inglorious new records of drains and water levels were recorded. The flood crest was reached on the 3rd June in Passau with a peak of 12.89 m, triggered by the connection of the tidal waves of the Inn, Ilz and Danube.[23] The height of the water surpassed even the unprecedented flood of 1501. While the floods of the Inn were mainly generated by the tributaries of Salzach, Alz and Rott, the waters of the Danube were increased from the tributaries of Regen, Naab and Vils, as well as Iller, Lech and Isar. On the 4th June 2013, the peak of the river Isar overlaid with the floodwaters of the Danube, causing the break of a dike at the Isar estuary near Steinkirchen-Fischerdorf (Deggendorf).[24] The Danube floodwaters reached Passau on June 6th, but the outflow of the Inn had already advanced so far, that the situation did not get any worse. (Thieken and DKK, 2015)

[23] The GIF on Imgur shows the dramatic comparison of Passau before and during the flood 2013. http://imgur.com/iFrPF1Q (latest access: 08.07.2016)

[24] http://www.news.de/panorama/855425286/hochwasser-2013-deggendorf-versinkt-in-der-donau/1/ (latest access: 30.08.2016)

5.2.3 Elbe/Saale

At the same date the floods appeared in the city of Dresden, marking the beginning of the severe Elbe/Saale floods. On the one hand, the waters of the river Moldau (Czech Republic) led to the ramparts and on the other hand the tributaries through the river Saale and Halde were mainly responsible for the floods. On the 8th June, the flood discharge of the river Saale joined with the Elbe waters. Several records were measured from the city Halle (Saale) to the estuary in the river Elbe. The day after the flood peak of the Elbe reached the capital city of Saxony-Anhalt, Magdeburg. Downriver of Magdeburg, the tidal wave was diminished by the tremendous dike break in Fischbeck on the 10th June. The event had such an impact on the landscape that five days later three ships had to be sunken along the break to stop the massive water amounts.[25] It is noteworthy that along the river Elbe, which passes Germany from the Czech border at Schönau and reaches the North Sea at Hamburg, nearly all administrative districts exclaimed disaster alert (**see Fig. 9**). (Thieken and DKK, 2015)

[25] http://www.spiegel.de/panorama/hochwasser-bei-fischbeck-drittes-schiff-nach-deichbruch-versenkt-a-906047.html (latest access: 30.8.2016)

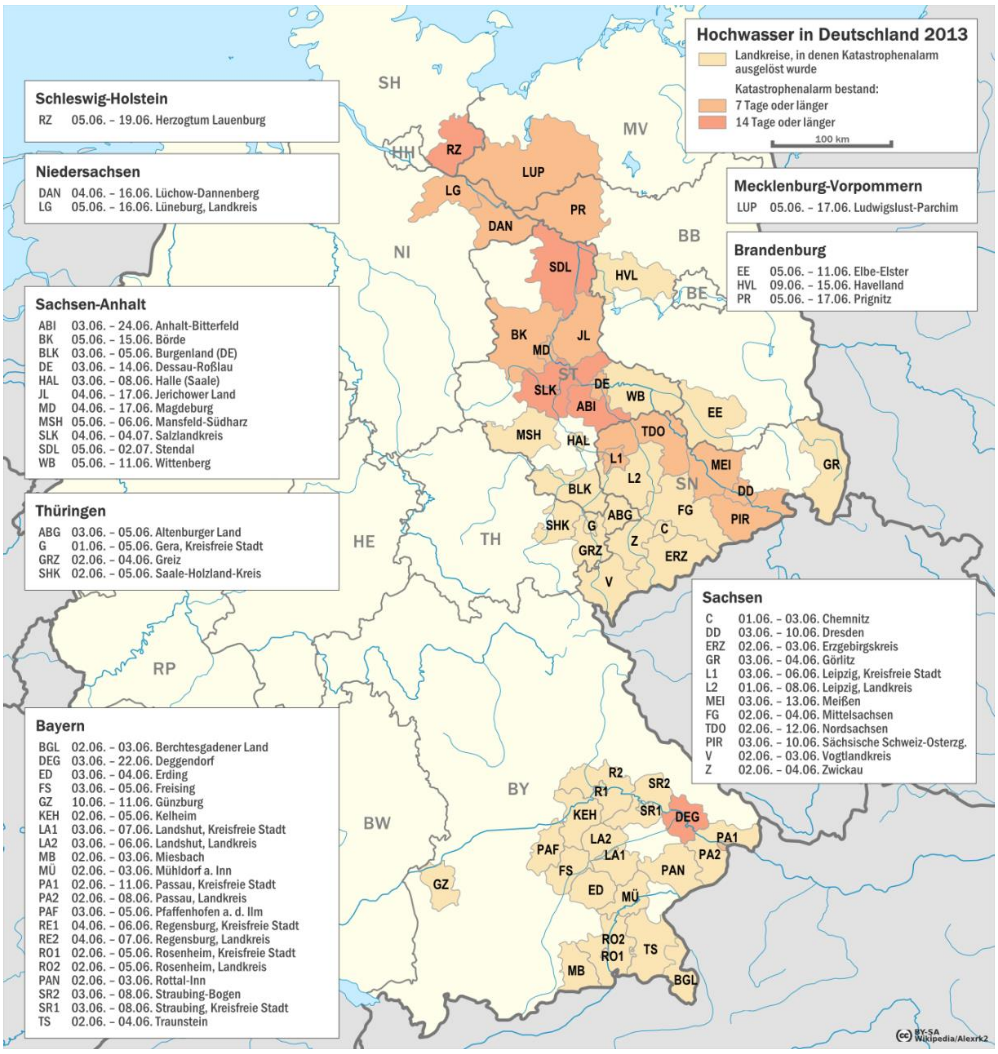

Hochwasser in Deutschland 2013
Landkreise, in denen Katastrophenalarm ausgelöst wurde
Katastrophenalarm bestand:
7 Tage oder länger
14 Tage oder länger
100 km
Schleswig-Holstein
RZ 05.06. – 19.06. Herzogtum Lauenburg
Niedersachsen
DAN 04.06. – 16.06. Lüchow-Dannenberg
LG 05.06. – 16.06. Lüneburg, Landkreis
Sachsen-Anhalt
ABI 03.06. – 24.06. Anhalt-Bitterfeld
BK 05.06. – 15.06. Börde
BLK 03.06. – 05.06. Burgenland (DE)
DE 03.06. – 14.06. Dessau-Roßlau
HAL 03.06. – 08.06. Halle (Saale)
JL 04.06. – 17.06. Jerichower Land
MD 04.06. – 17.06. Magdeburg
MSH 05.06. – 06.06. Mansfeld-Südharz
SLK 04.06. – 04.07. Salzlandkreis
SDL 05.06. – 02.07. Stendal
WB 05.06. – 11.06. Wittenberg
Thüringen
ABG 03.06. – 05.06. Altenburger Land
G 01.06. – 05.06. Gera, Kreisfreie Stadt
GRZ 02.06. – 04.06. Greiz
SHK 02.06. – 05.06. Saale-Holzland-Kreis
Mecklenburg-Vorpommern
LUP 05.06. – 17.06. Ludwigslust-Parchim
Brandenburg
EE 05.06. – 11.06. Elbe-Elster
HVL 09.06. – 15.06. Havelland
PR 05.06. – 17.06. Prignitz
Sachsen
C 01.06. – 03.06. Chemnitz
DD 03.06. – 10.06. Dresden
ERZ 02.06. – 03.06. Erzgebirgskreis
GR 03.06. – 04.06. Görlitz
L1 03.06. – 06.06. Leipzig, Kreisfreie Stadt
L2 01.06. – 08.06. Leipzig, Landkreis
MEI 03.06. – 13.06. Meißen
FG 02.06. – 04.06. Mittelsachsen
TDO 02.06. – 12.06. Nordsachsen
PIR 03.06. – 10.06. Sächsische Schweiz-Osterzg.
V 02.06. – 03.06. Vogtlandkreis
Z 02.06. – 04.06. Zwickau
Bayern
BGL 02.06. – 03.06. Berchtesgadener Land
DEG 03.06. – 22.06. Deggendorf
ED 03.06. – 04.06. Erding
FS 03.06. – 05.06. Freising
GZ 10.06. – 11.06. Günzburg
KEH 02.06. – 05.06. Kelheim
LA1 03.06. – 07.06. Landshut, Kreisfreie Stadt
LA2 03.06. – 06.06. Landshut, Landkreis
MB 02.06. – 03.06. Miesbach
MÜ 02.06. – 03.06. Mühldorf a. Inn
PA1 02.06. – 11.06. Passau, Kreisfreie Stadt
PA2 02.06. – 08.06. Passau, Landkreis
PAF 03.06. – 05.06. Pfaffenhofen a. d. Ilm
RE1 04.06. – 06.06. Regensburg, Kreisfreie Stadt
RE2 04.06. – 07.06. Regensburg, Landkreis
RO1 02.06. – 05.06. Rosenheim, Kreisfreie Stadt
RO2 02.06. – 05.06. Rosenheim, Landkreis
PAN 02.06. – 03.06. Rottal-Inn
SR2 03.06. – 08.06. Straubing-Bogen
SR1 03.06. – 08.06. Straubing, Kreisfreie Stadt
TS 02.06. – 04.06. Traunstein
SH
MV
HH
NI
BB
BE
ST
SN
HE
TH
RP
BY
BW
RZ
LUP
LG
PR
DAN
SDL
HVL
BK
MD
JL
SLK
DE
ABI
WB
EE
MSH
HAL
TDO
L1
MEI
GR
L2
BLK
DD
ABG
FG
PIR
SHK
G
C
GRZ
Z
ERZ
V
R2
R1
SR2
KEH
SR1
DEG
PA1
LA2
PAF
LA1
PA2
FS
PAN
GZ
ED
MÜ
RO2
RO1
TS
MB
BGL
(cc) BY-SA Wikipedia/Alexrk2

Figure 9. Administrative districts in which the disaster alarm was triggered (source: Wikipedia)

5.2.4 Summary

All in all it can be observed that the first region Rhine/Main/Weser was much less affected than the Danube/Inn/Isar and the Elbe/Saale areas. Furthermore, the spatiotemporal extent of the flood peaks is shown in the following table:

Region	**Rhine/Main/Weser**	**Danube/Inn/Isar**	**Elbe/Saale**
Flood Peaks	2nd to 3rd June	3rd to 6th June	6th to 13th June
Duration Disaster Alert	28th May-13th June	3rd to 10th June (Deggendorf: 22th June)	6th to 17th June
Highest Peak	Maxau, 2nd May	Passau, 3rd June	Magdeburg, 9th June

Of course some administrative districts – like Deggendorf – were affected much longer because of dike breaks, but that would exceed the investigation period of this thesis.

It was by far the largest fire brigades efforts since the existence of the Federal Republic of Germany. The German Fire Services Association had compiled from all the federal states around 70.000 fire fighters, who recorded together about 804.000 person-days of work.[26]

Not only the flood peaks were tremendous also the financial loss. According to the rating agency Fitch, Germany has caused economic losses of around 12 billion euros due to floods. In addition, there are insurance losses of around three billion euros.[27]

5.3 Twitter Data and Corpora

The Twitter dataset providing a 100% sample of all posted messages in a given time period was obtained by the company's own Firehose Streaming API.[28] All Tweets were collected using a bounding box covering Germany between the 27th May and the 13th June, reflecting the most affected flood period. Despite the fact that the recovery measures lasted much longer, this thesis investigates the flood events itself and their spatiotemporal impact on the Social Media platform Twitter. Analysing any accompany measures in the recovery phase would lead too far, given the amount of the thousands of single messages per day. All Tweets were stored in a PostgreSQL DB and filtered by their location field. Only Tweets which contain a point coordinate (with longitude and latitude) and intersect the official administrative borders of the Federal Republic of Germany were stored.

[26] http://www.fireworld.at/cms/story.php?id=45782 (latest access: 02.09.2016) The original press announcement of the German Fire Service Association) is no longer accessible. (latest access: 12.01.2017)

[27] http://www.zeit.de/wirtschaft/2013-06/hochwasser-schaeden-versicherungen/komplettansicht (latest access: 30.08.2016)

[28] Many thanks to Dr.B. Resch for organising and providing the Twitter data.

All other messages, which did not fulfil these criteria, were removed from the database. For the intersection query, the official borders of Germany were extracted from Open Street Map.[29] All datasets were delivered as CSV-files and were imported by the CSV-import method of the DB. **Figure 10** represents the entire Tweet corpus in the given study area and leads us to some first conclusions.

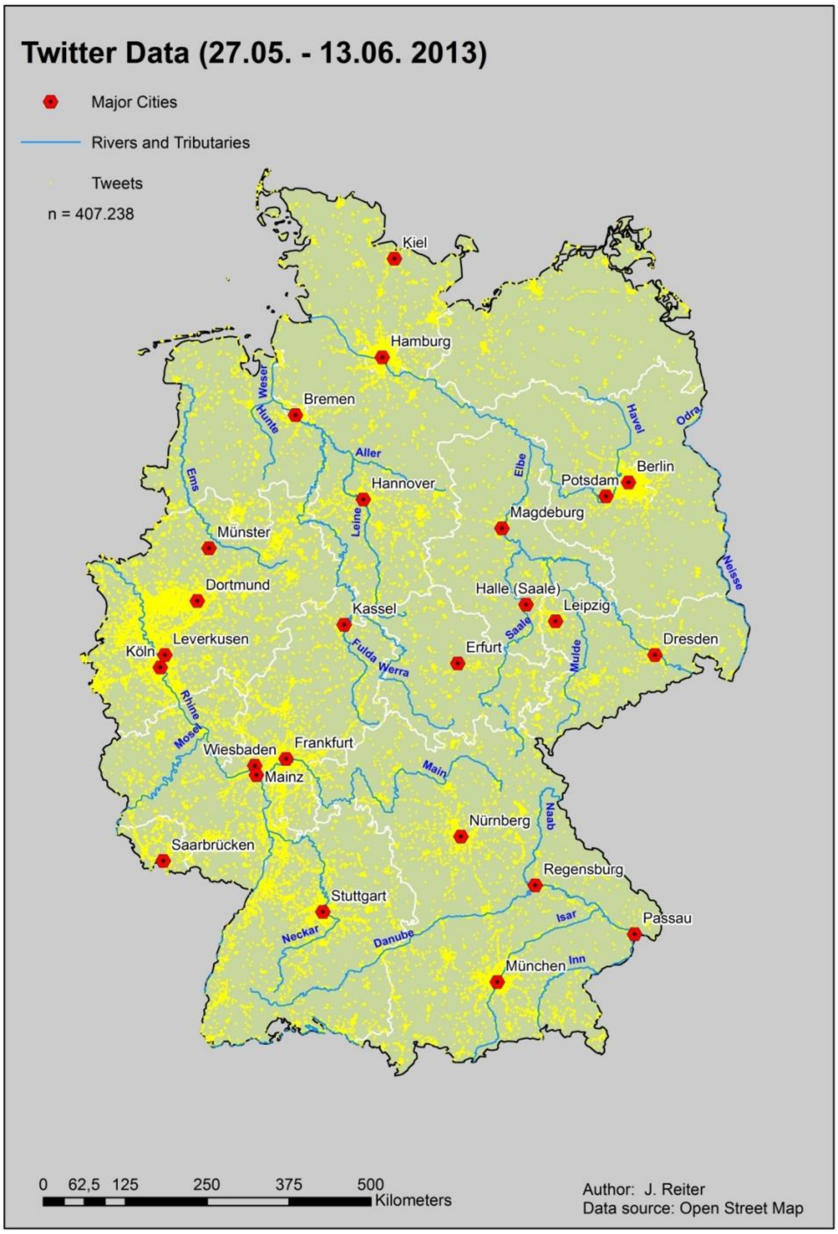


Figure 10. Visualisation of all georeferenced Tweets in Germany between the 27th May and the 13th June 2013

[29] https://openstreetmap.at (latest access: 12.12.2016)

There is a clearly defined pattern to the Tweet-Map indicating a correlation of a region's population density with the amount of Tweets located at the same place. As expected the distribution pattern of Tweets rougly matches the distribution of people. As predicted the densest social communication can be observed around the regional capitals of the German states. The colouring was deliberating chosen to remind the reader of a light pollution map and **Figure 11** shows the direct comparison of how much the two patterns are similar. Both maps are clear markers of civilising activities. The Satellite image was recorded 2014 by the Operational Linescan System (OLS) on behalf of the Defense Meteorological Satellite Program (DMSP) and provided by the National Oceanic and Atmospheric Administration (NOAA).[30]

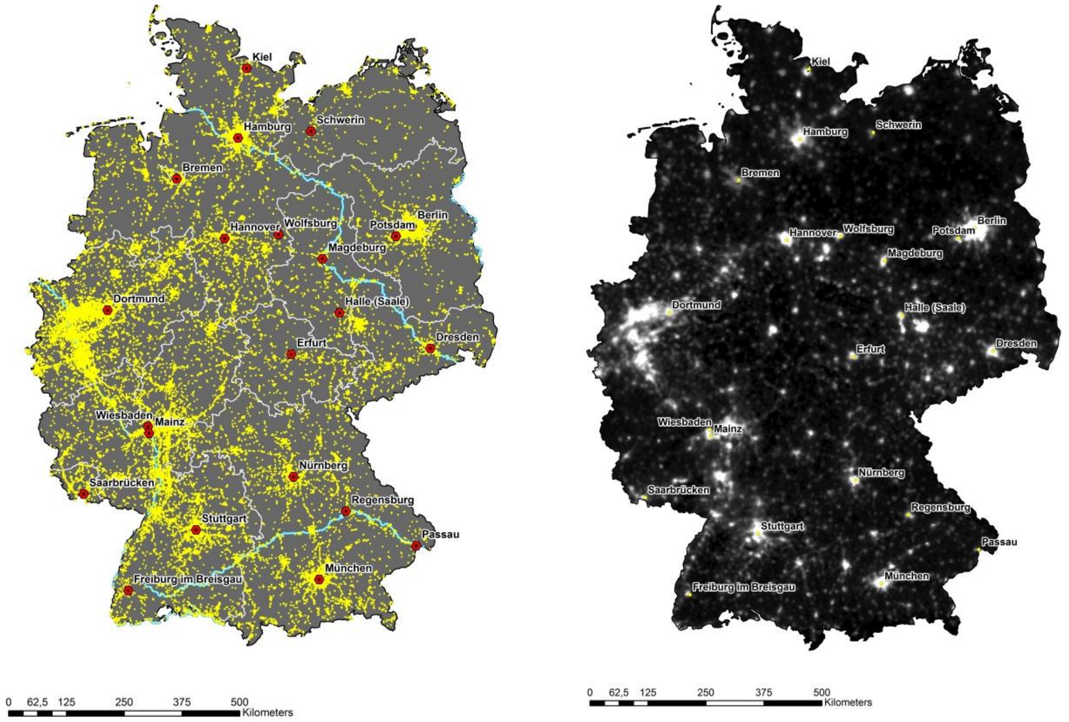
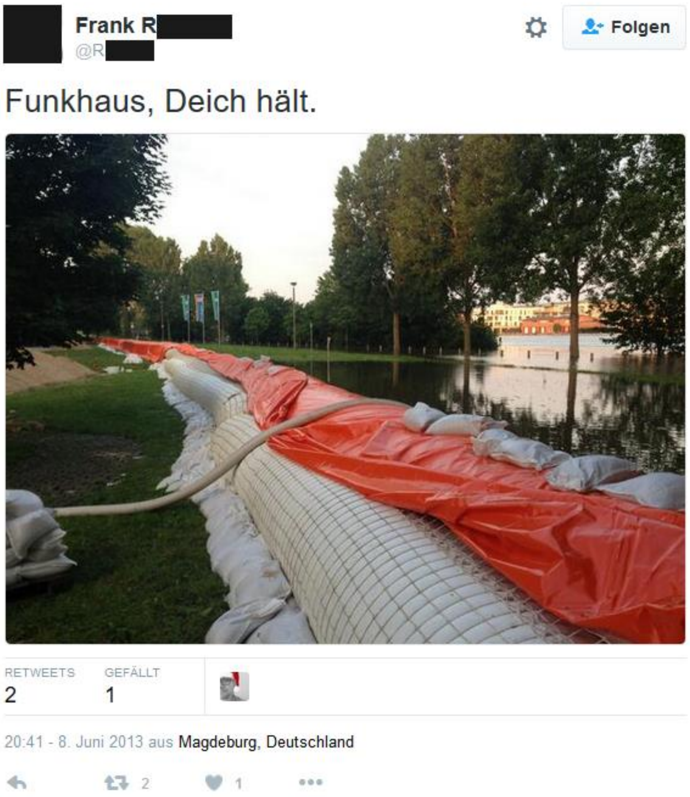


Figure 11. Comparison of the Tweet visualisation (l) with a light pollution map of Germany (r)

[30] https://www.ngdc.noaa.gov/eog/dmsp/download_radcal.html (latest access: 12.09.2016)

As mentioned every Tweet collected by the Firehose API has a lot of additional metadata. To give a real-world example of what has been collected in the PostGIS/PostgreSQL database, the following flood Tweet of Frank R. **(Fig. 12.)** shall be examined. For privacy concerns his real name and account photo have been blackened.[31]



[redacted] Frank R[redacted]
@R[redacted]

Funkhaus, Deich hält.

RETWEETS 2 GEFÄLLT 1

20:41 - 8. Juni 2013 aus Magdeburg, Deutschland

Figure 12. Tweet about the dike near the Funkhaus in Magdeburg[32]

The Tweet from Frank R. contains much useful information about the flood event. Beside the Tweet text, which means that the dike near the Funkhaus is holding, the situation is approved by a photo about the dike itself. Time, date and the location of the event at Magdeburg are also visible for everybody visiting his

[31] As described in Chapter 2, all Twitter messages are public domain as long as they are not set to private by the owner. However, the author of this thesis did not want to offend anyone by unfolding his/her Twitter communication with assignable names.

[32] https://twitter.com/Rugullis/status/343573563950891008/photo/1 (latest access: 10.10.2016)

post on Twitter. Furthermore, there is a lot of additional metadata, which can be accessed via Twitter's API. The table below shows the collected data and their data type stored in the PostgreSQL database:

Tweet ID	Time	Latitude	Longitude	User ID
3435xxxx	08.06.2013 20:41:35	52, 1225912	11, 63927958	783xxxx
numeric	timestamp with time zone	double precision	double precision	numeric

User Name	Operating System	Reply to User	Reply to Tweet	Place ID
Rxxxxx	IPhone	0	0	4d6e86b07e4383c8 (Magdeburg)
varchar	varchar	numeric	numeric	varchar

Tweet Message	Hashtags
Funkhaus, Deich hält. http://t.co/XokgbAGH2c	none
varchar	varchar

Therefore, the table shows that each Twitter message has his own unique identifier and also every Twitter user can be identified by a unique ID. As long as the post is online available and not private, this IDs can be used to look up the original Tweet. For practical reasons and further computations, a serial data type field is added to the table as the primary key. In addition to the coordinate-fields 'Latitude' and 'Longitude', both are combined and constructed as Point geometry with the OGC[33]-Well-Known text representation methods 'ST_SetSRID(ST_MakePoint(lon, lat),4326)' in a separate column. As the number 4326 in the spatial reference method implies, the coordinates system is set to the World Geodetic System 1984, which is also the standard for GPS.[34] Furthermore, the near real-time aspect of Twitter must be considered. For this reason, the

[33] http://www.opengeospatial.org/ (latest access: 14.12.2016)
[34] http://spatialreference.org/ref/epsg/wgs-84/ (latest access: 14.01.2017)

dataset was split up into one table per day, to correlate and compare each day with the development of the flooding.

5.3.1 Data Measures

For a better understanding of the collected datasets, the following Chapter addresses a basic description of the Twitter corpora of 2013 and some essential measurements, which are invaluable for the further presented methods. The semantic content of the Tweets is wilful excluded.

From **the 27th May to the 13th June 2013**, **407.238** Tweets were collected from **34.235** individual users. One of the major issues in analysing Tweet corpora is the immense usage of Twitter bots, which are currently under fire because of their influence on the presidential election 2016 of the United States.[35] On the other hand in 2013 those bots are heavily used by radio stations or cinema centres announcing every played song or movie and therefore generating hundreds of Tweets per day. Despite the fact that bots are part of the Twitter data generation in this Twitter corpus the seven most frequently posting Twitter bots were removed because their text similarity distracts most of the common Natural Language Processing tasks, which are described further in **Chapter 6**. These automatic generated messages are also unrelated to the floods and can be omitted without any concerns of losing critical information. On the contrary, these spam messages can be identified very easily through frequency queries, and their removal speeds up the follow-up computations. To give an example for a removed Tweet bot:

Tweet ID	**User Name**	**Tweet Text**
338843115526	BB_RADIO_MUSIC	#nowplaying #adele ~ Adele | Set Fire To The Rain ||| BB RADIO - In #Potsdam #Brandenburg #GER auf 107.5

According to the removal, it is most remarkable that the number of Tweets decreases by nearly one-fifth to **327.714** single messages, in other words, seven

[35] http://www.theatlantic.com/technology/archive/2016/11/election-bots/506072/ (latest access: 15.12.2016)

automated bots were responsible for a one-fifth of the georeferenced Twitter data in Germany. The next step of data pre-processing and measurement is to filter the posts per day. This is necessary because floods are highly dynamic disaster events, which can last for months but change their active zone in a few hours. The classification interval of days was chosen because the data was collected back in 2013. It would also be worth considering reducing the intervals to six hours or even a one-hour interval, to refer to the near real-time character of Twitter in an ongoing disaster alert scenario. The following diagram (**Fig. 13.**) represents the number of messages in Germany per day and sets them in combination with an ordinary keyword search for the related terms ‘Hochwasser’ and ‘flood’. Due to server issues, only a few thousand Tweets could be collected on the 30th May. For that reason, the graphs are extremely low on this date and the day cannot be compared to all others. The area diagram below (**Fig. 13.**) is showing a good representation of the posted Tweet numbers in Germany from the 27th May to the 13th June. Furthermore, a weekly rhythm of posting frequencies can be observed, which have their peaks on the weekend from the 31st May to the 2nd June and the 7th-9th June. The Mondays, the 27th May and the 10th June, have the lowest Twitter activity.

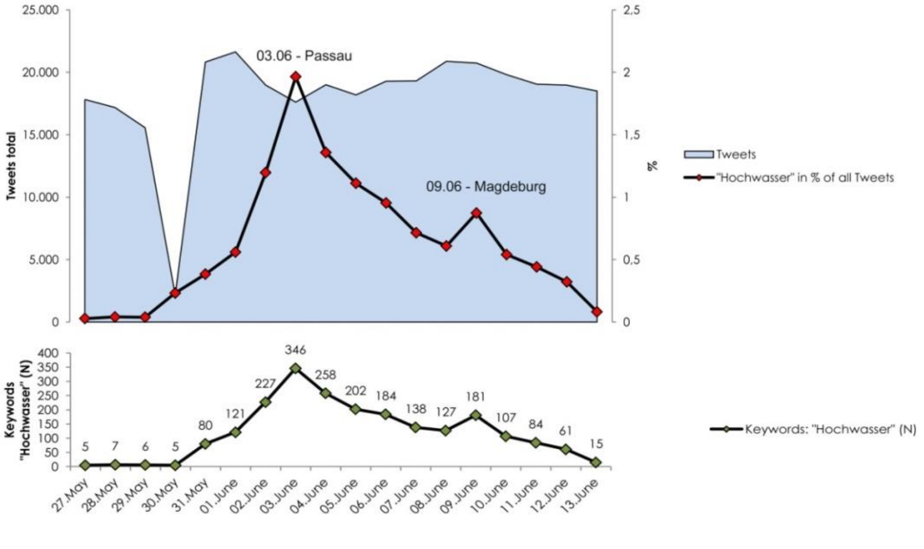


Figure 13. Twitter messages per day from the 27th May to the 13th June comparison to the number of Tweets with the keywords ‘Hochwasser’ and ‘flood’ in percent and below the total number per day.

As on many other recent Twitter studies, a basic keyword search was carried out to get an overview, what critical information can be expected (Albuquerque et al., 2015; Kongthon et al., 2014; Peters and Albuquerque, 2015). Tweets containing the German keywords 'Hochwasser', 'Deich', 'Flut' and 'Überschwemmung' as well as their English representations based on the Oxford English Dictionary were filtered.[36] The diagram reveals in accordance with the bespoken flood process that only a few Tweets were posted during the heavy rain falls and the beginning of the floods along the Rhine/Main/Mosel area. As recently as the 31st May the number of Tweets increased and showed a steady enlargement until the peak on the 3rd of June, which marks the highest number of Tweets per day in the dataset. From that date the figures show a continuous decrease until the 8th of June, before the graph increases again and a second crest can be observed on the 9th of June. Afterwards, the graph shows a steady decrease until the 13th of June. The map below (**Fig. 14.**) displays all 2173 Tweets, which are related to the mentioned flood keywords.

[36] http://www.oed.com/ (latest access: 02.11.2016)

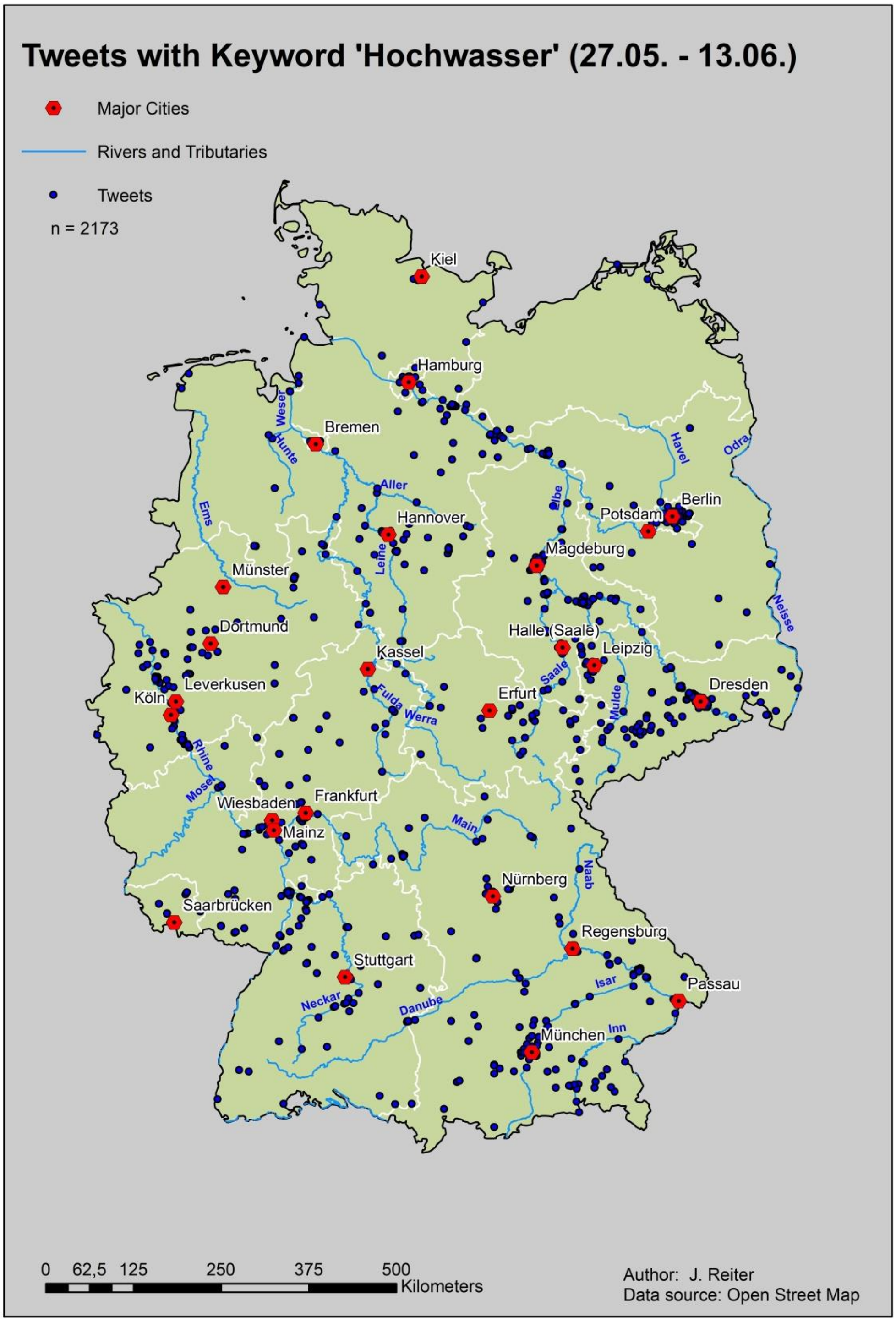
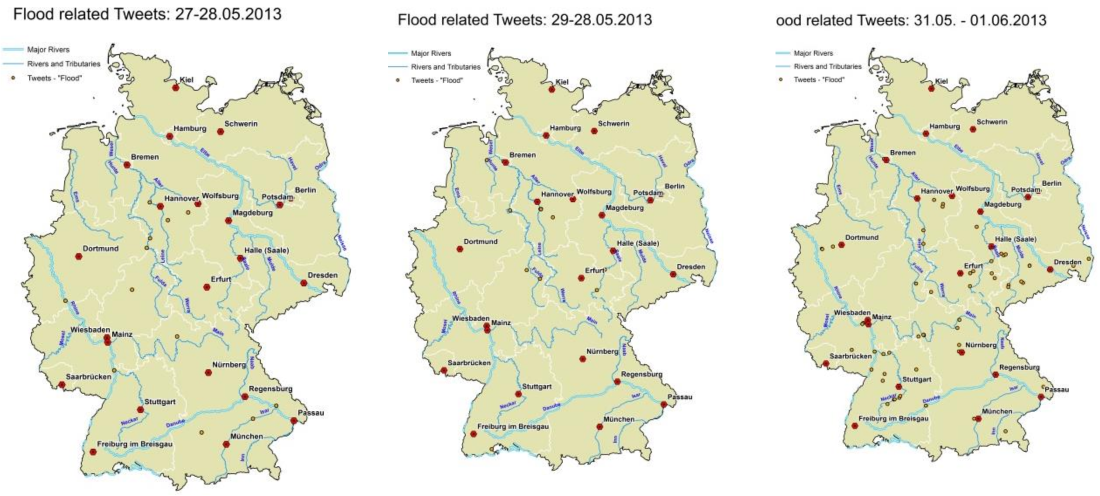


Figure 14. Display of all georeferenced Tweets containing one of the bespoken keywords.

There is a clearly defined pattern that most of the Tweets correspond spatially with the major rivers of Germany. This perception is approved by the former studies (Albuquerque et al., 2015; Peters and Albuquerque, 2015). They showed that flood-related Tweets were 11 times more likely to occur near (<10 km) flood affected areas than 30 km away. Of course, these considerations do not include any semantic contents, which are presented later, but the pure numbers of the Tweet appearance.

Another important observation can be made from the spatiotemporal relation of the single Tweets with the different flood events along the three main affected regions. The nine maps (**Fig. 15**) display the whole event in a summarised two-day interval. Without obtaining to the semantic content, two clear patterns can be described. From the 27th May to the 6th June the Tweets are concentrating in the Rhine/Weser/Main area and along the Danube/Inn/Isar region, while from the 6th to the 13th June the Elbe/Saale region has the most Tweet appearance. This corresponds to the chronological order of the several flood events presented in **Chapter 5**. Of course, there are many Tweets over the entire Republic of Germany on every single day, but the concentration along the rivers to the specific flood peaks is evident.



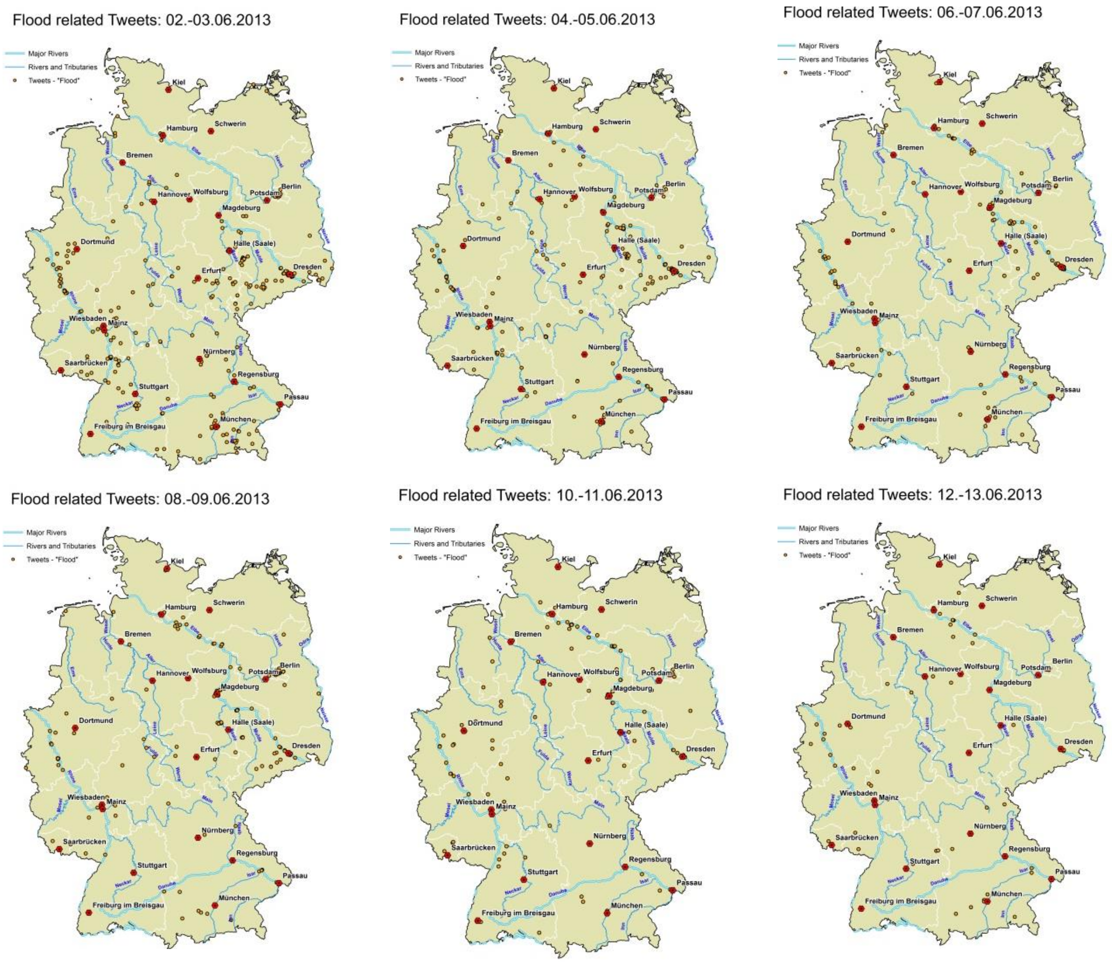
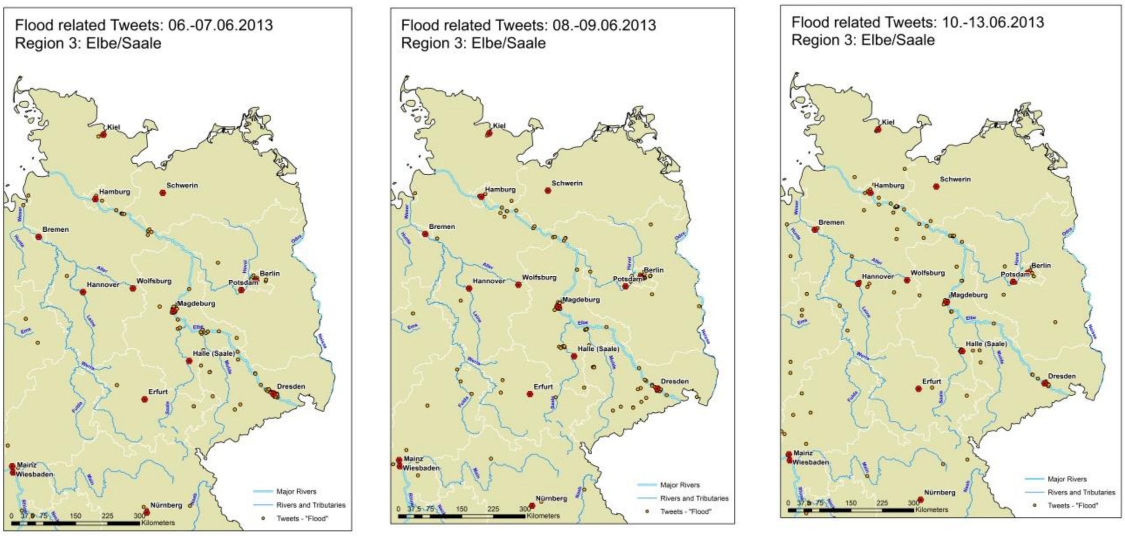


Figure 15. Spatiotemporal development of the flood-related Tweets in two-day intervals

Furthermore, these time slides can be presented on a larger scale for the most affected region Elbe/Saale. On the 6th June the flood reaches the city of Dresden with a water peak on the 8th. There is also a Tweet concentration in and near Dresden for that period. The situation is similar from the 6th to the 13th June where the Tweet appearance follows the tidal waves with nearly no delay and several small clusters are emerging along the rivers Elbe and Saale. This conclusion has to be constrained in several points. While the cluster concentration has a clear spatiotemporal pattern, they, of course, occur mainly in the densely populated cities with a high Twitter activity like Dresden, Magdeburg or Halle (**Fig. 16**).


Flood related Tweets: 06.-07.06.2013
Region 3: Elbe/Saale

Flood related Tweets: 08.-09.06.2013
Region 3: Elbe/Saale

Flood related Tweets: 10.-13.06.2013
Region 3: Elbe/Saale

Figure 16. Time series of Tweets in the Elbe/Saale region

5.4 Remote Sensing Data – ZKI (DLR)

The water masks of the flood affected areas in Germany are provided by the Center for Satellite Based Crisis Information[37] (ZKI-DE), which presents a service of the German Remotes Sensing Data Center[38] (DFD) and is, therefore, a department of the German Aerospace Center (DLR)[39]. Furthermore, the ZKI provides a 24/7 service for the rapid processing and analysing of satellite and remote sensing data during natural and environment disasters. The resulting products are provided to public authorities or relief organisations worldwide, in order to support Disaster Management operations or civil security issues.

In 2013 the German Joint Information and Situation Centre activated the International Charter ‘Space and Major Disasters’, which is a unified system of space data acquisition.[40] The Charter allows the participating organisations to get

[37] https://www.zki.dlr.de/de (latest access: 02.12.2016)

[38] http://www.dlr.de/eoc/en/desktopdefault.aspx/tabid-5278/8856_read-15911/ (latest access: 02.12.2016)

[39] http://www.dlr.de/ (latest access: 02.12.2016)

[40] https://www.disasterscharter.org/web/guest/home;jsessionid=F391FDE41E9BECFE2E9273CA3B1089F1.jvm1 (latest access: 02.12.2016)

cost-free and quick remote sensing data from public and private operators of earth observation satellites. During the activation, the DLR acted as the Charter project manager as well as a satellite data provider, with the TerraSAR-X[30] system. The ZKI-DE was tasked to create and deliver satellite-based information of the most affected flooded areas for the German Federal Ministry of the Interior[41].

5.4.1 Sensor Systems

The satellite data was mainly captured by the radar satellites TerraSAR-X[42], Radarsat 2[43] and RapidEye[44], which is equipped with optical cameras. For some of the ZKI-DE map products, a base map was rendered from the Pleiades[45] satellite program. Most of the captured flooded areas were made by TerraSAR-X and Radarsat 2. All radar sensors provide high-resolution and wide-area radar images with an average resolution between 3 and 5m. Furthermore, there are single areas like Passau (**see Fig. 17.**), where the high-resolution sensor TerraSAR-Spotlight was used with a resolution of 1.5m and only a few products were generated with resolutions higher than 20m.

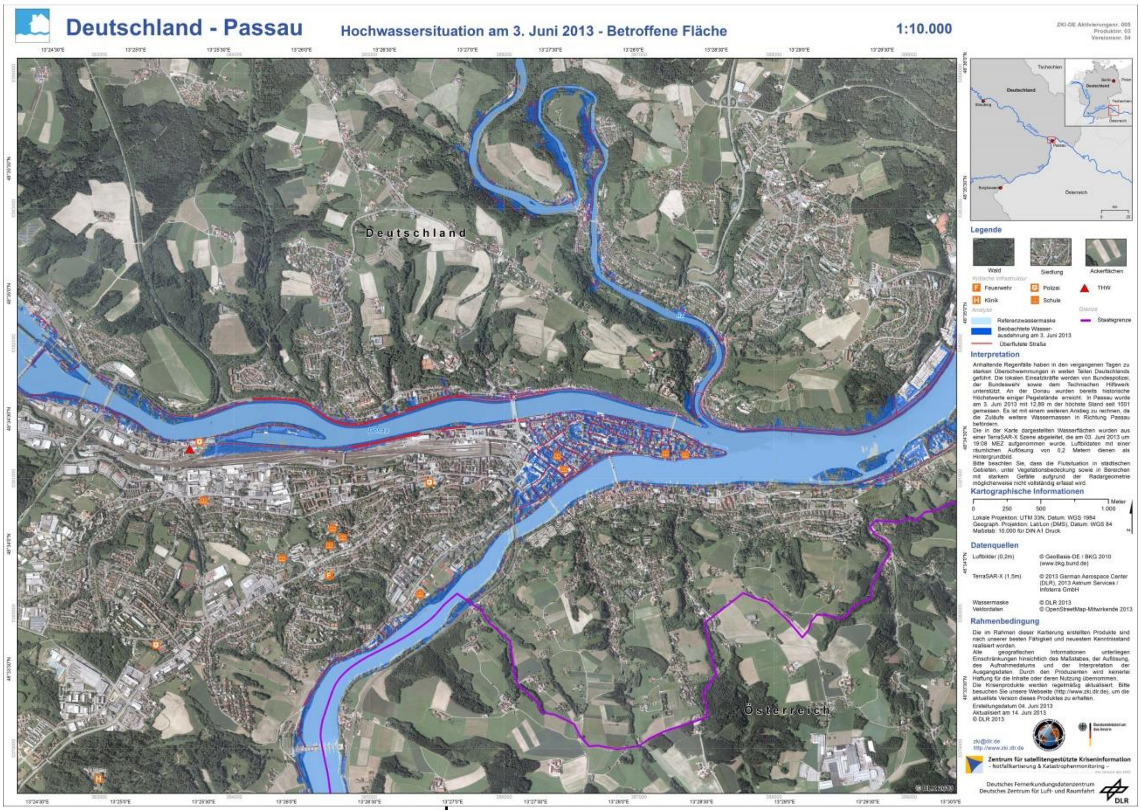
[41] http://www.bmi.bund.de/DE/Themen/Bevoelkerungsschutz/Zivil-undKatastrophenschutz/zivil-undkatastrophenschutz_node.html (latest access: 02.12.2016)

[42] http://www.dlr.de/dlr/desktopdefault.aspx/tabid-10377/565_read-436/#/gallery/350 and http://www.intelligence-airbusds.com/terrasar-x/ (latest access: 02.12.2012)

[43] http://www.asc-csa.gc.ca/eng/publications/default.asp#RADARSAT (latest access: 02.12.2012)

[44] http://www.dlr.de/rd/desktopdefault.aspx/tabid-2440/3586_read-5336/ (latest access: 02.12.2016)

[45] http://www.intelligence-airbusds.com/pleiades/ (latest access: 03.12.2016)


Deutschland - Passau
Hochwassersituation am 3. Juni 2013 - Betroffene Fläche
1:10.000
ZKI-DE Aktivierungsnr. 005
Produktnr. 03
Versionsnr. 04
Deutschland
Österreich
Tschechien
Polen
Berlin
Straubing
Passau
Burghausen
Donau
Inn
km
0
20
Legende
Wald
Siedlung
Ackerflächen
Kritische Infrastruktur
Feuerwehr
Polizei
THW
Klinik
Schule
Analyse
Grenze
Referenzwassermaske
Staatsgrenze
Beobachtete Wasserausdehnung am 3. Juni 2013
Überflutete Straße
Kartographische Informationen
Meter
0
250
500
1.000
N
zki@dlr.de
http://www.zki.dlr.de
Bundesministerium des Innern
Zentrum für satellitengestützte Kriseninformation
– Notfallkartierung & Katastrophenmonitoring –
Deutsches Fernerkundungsdatenzentrum
Deutsches Zentrum für Luft- und Raumfahrt
DLR
© DLR 2013

Interpretation

Anhaltende Regenfälle haben in den vergangenen Tagen zu starken Überschwemmungen in weiten Teilen Deutschlands geführt. Die lokalen Einsatzkräfte werden von Bundespolizei, der Bundeswehr sowie dem Technischen Hilfswerk unterstützt. An der Donau wurden bereits historische Höchstwerte einiger Pegelstände erreicht. In Passau wurde am 3. Juni 2013 mit 12,89 m der höchste Stand seit 1501 gemessen. Es ist mit einem weiteren Anstieg zu rechnen, da die Zuläufe weitere Wassermassen in Richtung Passau befördern.

Die in der Karte dargestellten Wasserflächen wurden aus einer TerraSAR-X Szene abgeleitet, die am 03. Juni 2013 um 19:08 MEZ aufgenommen wurde. Luftbilddaten mit einer räumlichen Auflösung von 0,2 Metern dienen als Hintergrundbild.

Bitte beachten Sie, dass die Flutsituation in städtischen Gebieten, unter Vegetationsbedeckung sowie in Bereichen mit starkem Gefälle aufgrund der Radargeometrie möglicherweise nicht vollständig erfasst wird.

Kartographische Informationen

Lokale Projektion: UTM 33N, Datum: WGS 1984
Geograph. Projektion: Lat/Lon (DMS), Datum: WGS 84
Maßstab: 10.000 für DIN A1 Druck

Datenquellen

Luftbilder (0,2m)	© GeoBasis-DE / BKG 2010 (www.bkg.bund.de)
TerraSAR-X (1,5m)	© 2013 German Aerospace Center (DLR), 2013 Astrium Services / Infoterra GmbH
Wassermaske	© DLR 2013
Vektordaten	© OpenStreetMap-Mitwirkende 2013

Rahmenbedingung

Die im Rahmen dieser Kartierung erstellten Produkte sind nach unserer besten Fähigkeit und neuestem Kenntnisstand realisiert worden.

Alle geografischen Informationen unterliegen Einschränkungen hinsichtlich des Maßstabes, der Auflösung, des Aufnahmedatums und der Interpretation der Ausgangsdaten. Durch den Produzenten wird keinerlei Haftung für die Inhalte oder deren Nutzung übernommen.

Die Krisenprodukte werden regelmäßig aktualisiert. Bitte besuchen Sie unsere Webseite (http://www.zki.dlr.de), um die aktuellste Version dieses Produktes zu erhalten.

Erstellungsdatum 04. Juni 2013
Aktualisiert am 14. Juni 2013
© DLR 2013

Figure 17. Flood situation on the 3rd June according to the official map product of the ZKI-DLR

One of the main advantages of acquiring radar data via optical satellites is their capability to capture data and gather information at any irrespecting of cloud cover. 49 maps and satellite products[46] were generated by the ZKI-DE between the 3rd and the 18th June covering the two most affected regions in Germany – Region 1: Donau/Isar and region 2: Elbe/Saale **(see Fig. 18.)**

[46] https://www.zki.dlr.de/de/article/2374 (latest access: 03.12.2016)

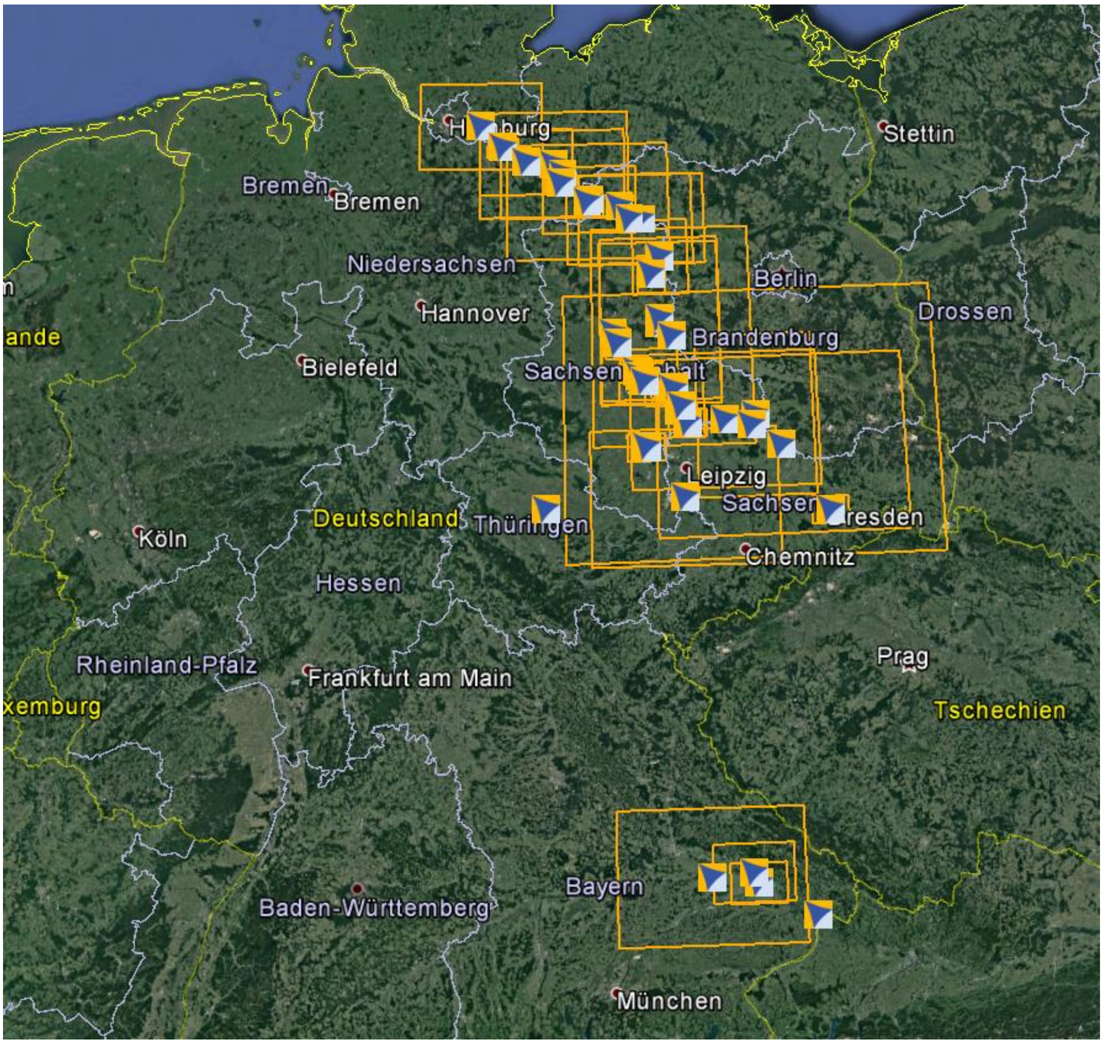


Figure 18. Display of all 49 map products provided by the ZKI-DE (DLR) during the flood of 2013 in Germany

5.4.2 Danube/Isar – Region 1

The first region covers the flood affected areas in the federal state Bavaria from the 3rd–7th June 2013. The ZKI-DE published five official map products covering the city of Passau, the Danube/Isar basins and the area around Deggendorf and Straubing. Additional many parts of the Danube and Isar tributaries, like Amper or Regen, are also provided. However, a relatively large piece of the Danube is missing between Neustift (Passau) and Herzogau (Künzing). The published water masks last from Vohburg (Pfaffenhofen) in the west to Rampsau (Regenstauf) in the north and Helfenbrunn (Kirchdorf) in the south. **Figure 19** shows all water masks in the federal state of Bavaria.

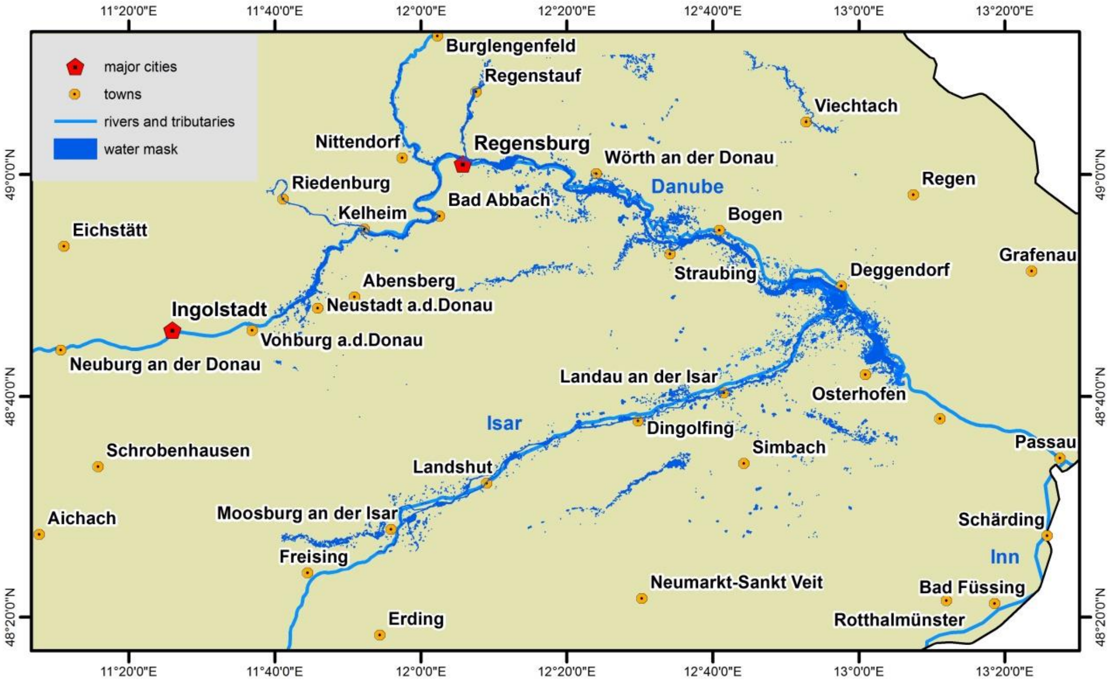


Figure 19. Water masks created by the ZKI-DE (DLR) in Bavaria from the 3rd–7th June 2013

5.4.3 Elbe/Saale – Region 2

The second region covers almost the entire course of the river Elbe and its tributaries along the federal states Saxony, Saxony-Anhalt, Thuringia and Hamburg. All data products were captured between the 3rd and the 18th of June. However, the small part of the river Elbe between Dresden and the German border to the Czech Republic was not recorded. In the opposite direction, the recording of the flooded areas was stopped shortly after the city of Hamburg, which is not surprisingly because this area was hardly affected (**Fig. 20**).

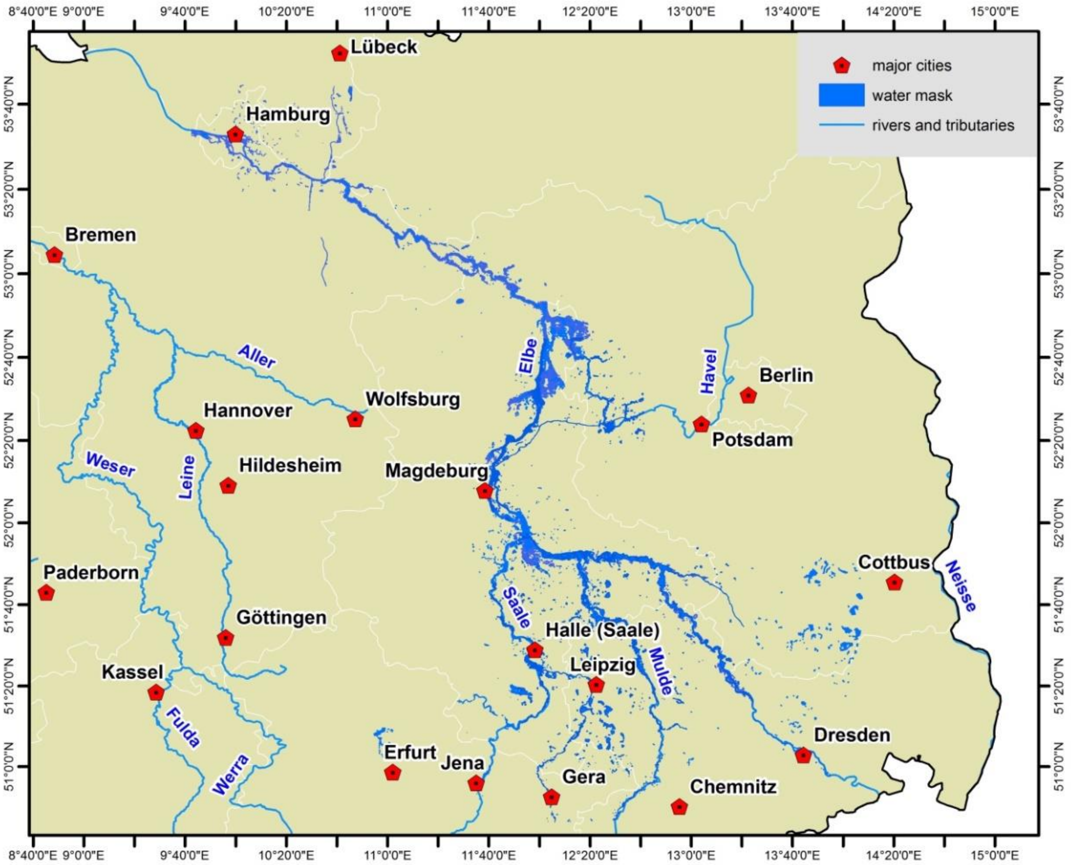


Figure 20. Water masks created by the ZKI-DE (DLR) in the Elbe/Saale region from the 3rd–18th June 2013

5.4.4 Data Processing

All flood masks are provided as shapefiles in vector format and are stored in the previous mentioned PostGIS/PostgreSQL database as a polygon datatype. It must be pointed out that of course not only the flood affected areas were captured by the satellites, but all water expanses like lakes or ponds. On balance, this has to be taken into account when investigating specific areas like the river harbour of Hamburg, which was nearly unaffected by the floods of 2013. A second restriction to the data is the temporal progression. As mentioned before the data was collected between the 3rd and the 18th June but the spatial extent of the captured data differs heavily. As an example the entire course of the river Elbe can only be displayed with data from several different days and a spatiotemporal development of the flood affected areas cannot be derived from all places along the two main regions Elbe/Saale and Danube/Isar.

6 Methodology

The following section provides the methods of extracting, organising, filtering and analysing Tweet corpora in general and further describes Natural Language Processing as well as probabilistic topic modelling in the context of Twitter. Furthermore, the author then reviews the factors of fusing remote sensing data with Twitter data as well as the additional benefits comparing traditional keyword searches. As mentioned in **Chapter 5** the dataset was collected in near real-time via the Twitter Firehose API[47], within a predefined Boundary Box covering Germany. Only georeferenced Tweets were extracted and further organised in a PostGIS/PostgreSQL database. **Figure 21** shows the analysis framework and describes the four main stages of this thesis, from data mining to the final visualisation.

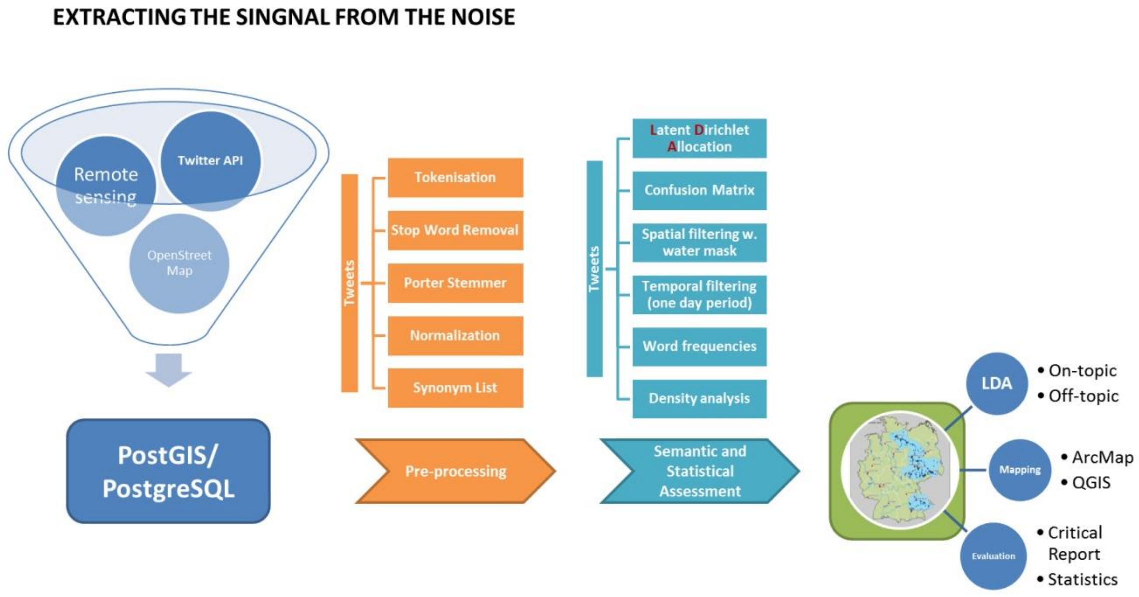


Figure 21: Analysis Framework: The four main phases

[47] https://dev.twitter.com/streaming/overview (latest access: 10.06.2016)

All steps are carried out with Python 3.5.1[48] on a Windows 10 workstation with an Intel i7 Core with 8 Cores and 32 GB-RAM. For the organisation of the coded scripts the integrated development environment of Pycharm: Professional Edition is used under an educational license.[49] The following Python modules and libraries are used:

Import Abbreviations	Name/Description	Source
gensim	Topic modelling for humans	https://radimrehurek.com/gensim/
json	JSON (JavaScript Object Notation)-Encoder and Decoder	https://docs.python.org/3/library/json.html
logging	Flexible event logging system for applications and libraries	https://docs.python.org/3/library/logging.html
matplot	Matplotlib is a 2D plotting library	http://matplotlib.org/
nltk	Natural Language Toolkit	http://www.nltk.org/
pandas	Pandas is providing high-performance, easy-to-use data structures and data analysis tools for Python	http://pandas.pydata.org/
numpy	NumPy is the fundamental package for scientific computing with Python	http://www.numpy.org/
re	Regular expression operations	https://docs.python.org/3/library/re.html
tweepy	Access the Twitter Streaming API	http://www.tweepy.org/
psychopg2	PostgreSQL adapter for the Python	http://initd.org/psycopg/

[48] https://www.python.org/ (latest access: 10.06.2016)

[49] https://www.jetbrains.com/pycharm/features/ (latest access: 10.06.2016)

6.1 Data Mining via Twitter Streaming API

Although all further analyses is carried out with data collected in 2013, this thesis tries to attend all steps right from the start. Furthermore, we need to look at the phenomenon Social Media in relation to Disaster Management in its entirety. This also comprises the data mining process itself. Of course, the author has no extremely cost intensive Firehose account and therefore uses the public accessible Streaming API.[50] However, the process of app development and coding differs only in the account constraints and not the coding itself. Thus an adaption to the 100% sample of the Firehose API is easy to accomplish if the necessary financial resources are given. In contrast to the holistically aspect, there is another central question of how many georeferenced Tweets can be collected with the public Streaming API compared to the full access of the major Firehose. As bespoken previously in **Chapter 2** the public Streaming API collects 1% of all Twitter messages in near-real time. As shown in other studies the amount of georeferenced Tweets (with coordinates: latitude and longitude) of this 1% is very low around 2-4% in the United States (Burton et al., 2012; Crooks et al., 2013; Steiger et al., 2015b) and even lower for Germany with nearly 1% (Fuchs et al., 2013). The later author supposed that, if someone filters the public Streaming API only for the georeferenced Tweets in the data collecting process, he would not only get 1-4% but the majority of all geolocated messages in Germany. The next sub-Chapter addresses this assumption.

6.1.1 Accessing the Twitter Stream

It is a precondition for using the Twitter API to have an own Twitter account, which was created for this thesis. Secondly, a new application was registered online on www.twitter.com/apps to get the necessary consumer-keys and access-tokens. Like the rest of the coding, all steps were carried out with Python 3.5.1.

[50] https://dev.twitter.com/streaming/overview (latest access: 05.12.2016)

The next step was to connect the application with the newly created PostGIS/PostgreSQL database, which was performed with the free Python PostgreSQL database adapter library **psycopg2**[51]. There are many other DB adapters for PostgreSQL, but psycopg2 is known for its stable handling of multi-threaded applications with lots of inserts or updates. Furthermore, to connect to the Streaming API **tweepy**[52] was used, which is a straight forward python library to connect and handle the Twitter Streaming API. As the data stream from Twitter is provided in JSON format also the **JSON** library was imported. As mentioned every Tweet contains a lot of metadata. Furthermore, it is really important to know the JSON data structure of Twitter, how to query and transmit the stream to the database. Previous knowledge of the obtained data from 2013 did speed up the process and the same parameters could be reused and enhanced in the script.[53] After the database and the table have been established an exception was coded to filter only those Tweets, where the coordinate key of the JSON data is populated with latitude and longitude and which are within a bounding box covering Germany. The bounding box query had the following extent:

WGS-84	latitude	longitude
NE-corner	47.2982950321	5.077004901
SW-corner	55.0039819676	15.0403900256

The data was collected during the period beginning with Monday, the 28th of November 00:00 am and ending on the 4th of December midnight, while the data collected through the Firehose was collected from the 3rd to the 8th of June. After the successful data retrieval, all Tweets were removed, which were not within or intersect the borders of Germany. This step was carried out in the database, and as source data layer the borders of Germany were extracted from Open Street Map data.

[51] http://initd.org/psycopg/ (latest access: 09.12.2016)

[52] http://www.tweepy.org/ (latest access: 09.12.2016)

[53] For further details see the field guide for Tweets: https://dev.twitter.com/overview/api/tweets (latest access: 08.12.2016)

It must be noticed that comparing Tweet amounts by time has some restrictions. The first is that Social Media sites are growing every year and the total number of Tweets is fluctuating. Secondly, the Tweet rate is also depending on seasonal deviations. Despite those restrictions, a weekly period was chosen to get approximate values, in order to get values comparable with the 100% Firehose sample. (**see Fig. 22.**).

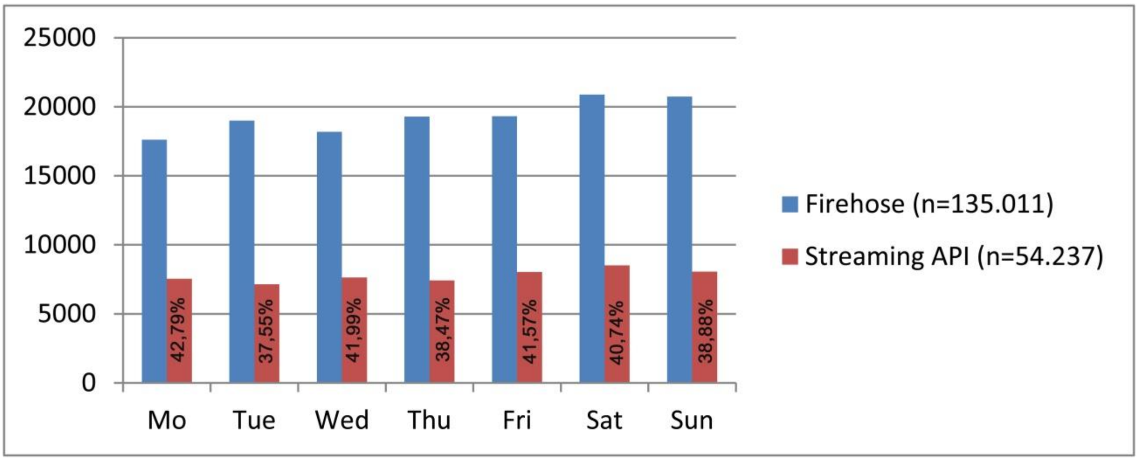


Figure 22. Comparison of the Tweet amount from the different API approaches

This diagram represents the direct comparison and reveals that while the Firehose stream collected 135.011 single Tweets the public Streaming API did only gather 54.237 single messages in a one week period. Fuchs' (et al., 2013) estimation that it might be possible to get the majority of all georeferenced Tweets in Germany cannot be confirmed as a whole, but the numbers show that we get at least around 40%. This is all the more impressive as the public access only grants 1% of the current Twitter stream (both types: with and without geolocation), but if the stream is filtered by coordinates during the extraction process, forty times as many single messages can be retrieved for free. The results of the present study can be visualised best with a kernel density map, because the display of the Tweets as single points is looking very similar, because of the small scale. The kernel density method calculates the density of point features around each output cell. **Figure 23** shows a comparison between the Firehose data from 2013 and the public Streaming API from 2016. Although the density centres overlap in the major cities, the Firehose data has a bigger spread.

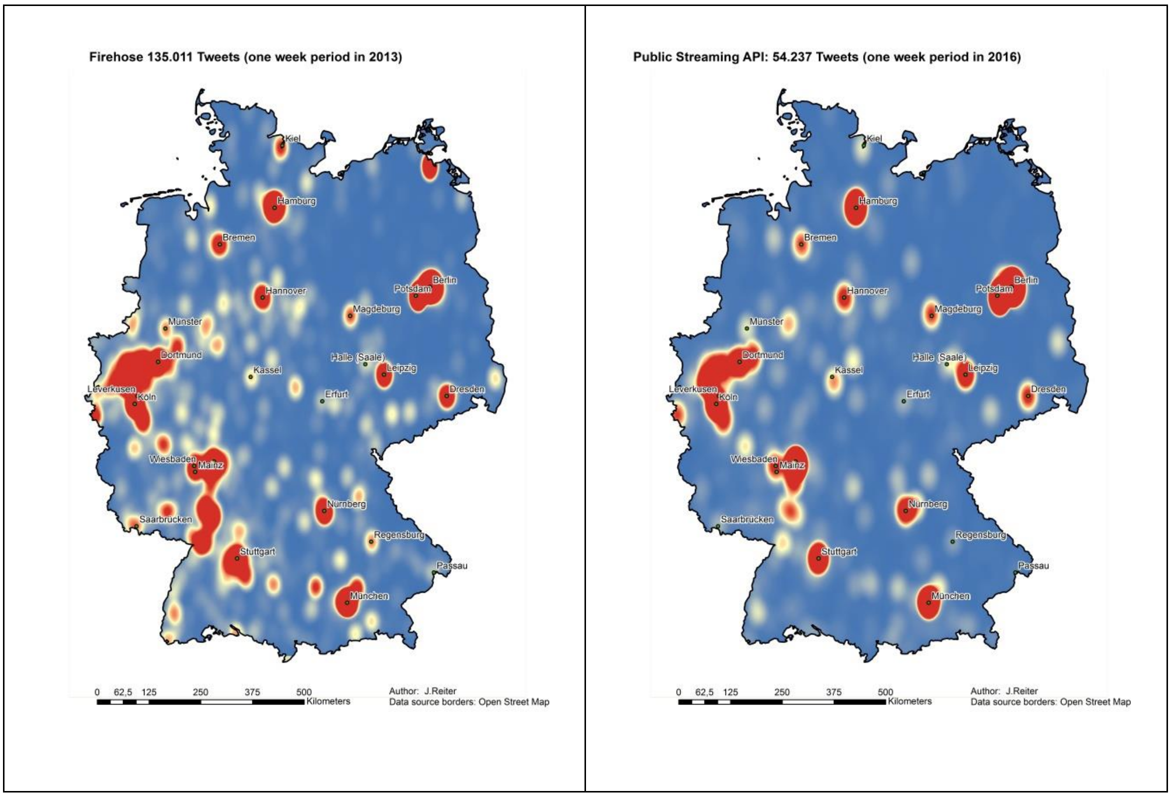


Figure 23. Kernel density maps of the Firehose data 2013 and the public Streaming API (cell-size: 100m, standard deviation = 1.5)

Further investigations and comparisons have to be made to verify these observations with not only actual public Streaming API data but also actual Firehose messages. However, one can clearly see that further studies, interested only in georeferenced Tweets (or lacking financial resources) should query the API with a fixed study area to get the most out of the extraction process.

6.2 Pre-processing and Natural Language Processing (NLP)

The semantic dimension of Twitter feeds is enormous and so the amount has to be reduced first in a pre-processing step for further calculations. One has to extract the signal from the noise. To carry out processing on texts several Natural Language Processing (NLP) techniques are applied. NLP is related to the interaction between natural language and the computer. With the help of machine learning, computers get the ability to understand human speech and therefore it is an integral part of Artificial Intelligence (AI). The following sub Chapters represents the technical implementation, the pre-processing steps and further probabilistic topic model analysis for the Firehose data from 2013.

6.2.1 Normalization

After the connection to the PostGIS/PostgreSQL database has been established several normalisation steps have to be performed to the Tweet text fields. All punctuations, whitespaces and numbers are removed with regular expression patterns. In a second step, all URL's or @ mentions are as well removed because they do not contain semantic information. The next step is to convert the entire text into lowercase, but it would also be possible to do the otherwise as long as it is uniform.

The first measure is to convert the text sentences in single tokens through tokenisation. For this step, the tokenisation method from the Python Natural Language Toolkit was used.[54] After splitting the sentences several filter techniques are performed, most of them through self-made Python functions, which are all part of the text normalisation. One of the most important tasks is stop word removal. Stop words like 'a' and 'or' (and many more) do not contribute much to the overall meaning of a sentence and would be overrepresented in every frequency distribution. The filter process consists of two steps: the first uses the standard stop word list from the above-mentioned NLTK module, which has been slightly adapted to remove the stop words in the most common languages like 'English', 'Spanish', 'Finnish', 'French', 'German', 'Hungarian', 'Italian', 'Norwegian', 'Portuguese', 'Russian', 'Swedish', 'Turkish', 'Dutch', 'Danish'. The second phase performs a second run with a self-made stop word list, backbone of which consists of the Cornell University stop word list,

[54] http://www.nltk.org/ (latest access: 11.10.2016)

developed by G. Salton and C. Buckley.[55] This is especially important for Twitter texts, because of their special slang behaviour. Tweets often contain no real sentences, and many are full of shorthands, codes or weird formatting, which is owed to the maximum of 140 characters a Tweet can contain. Because of the briefness people are forced to be concise or creative, which results constantly changing terms and codes.[56] (**see Fig. 24.**)[57]

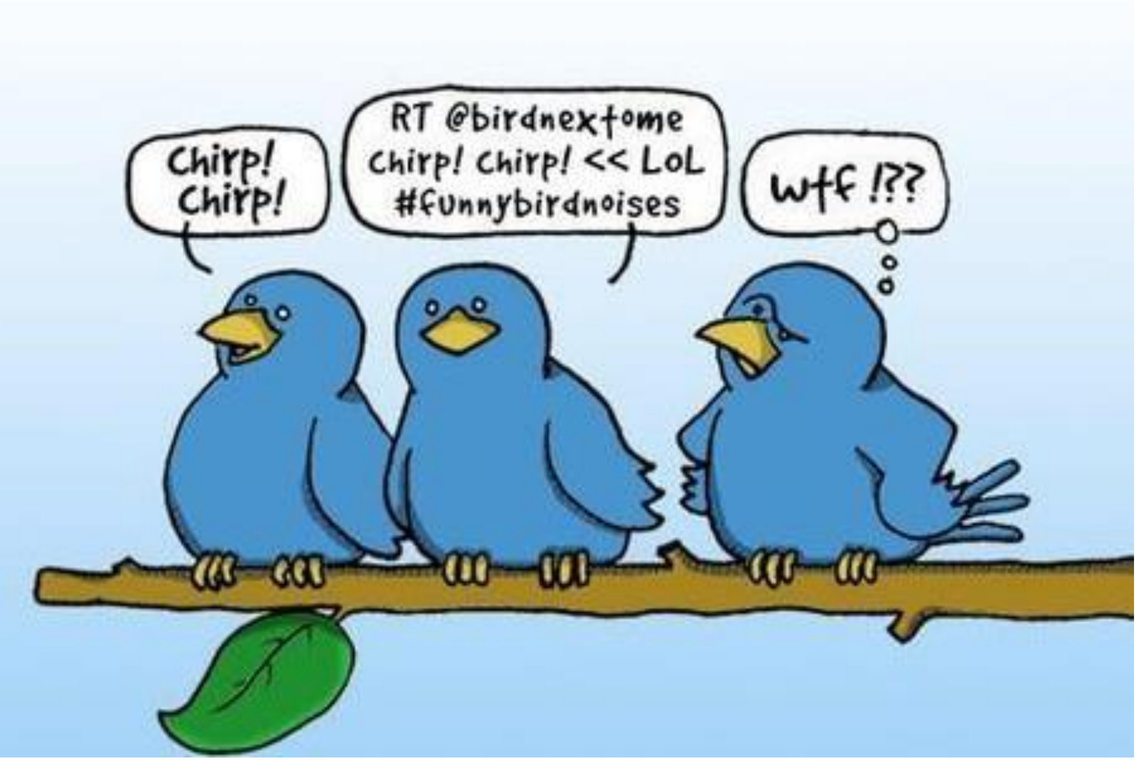


Figure 24. Cartoon 'Even Tweeting Birds Do It' visualising the Twitter slang problem

Traditional stop word lists cannot handle this slang richness, and a manual adaption is invaluable. Frequency Distributions can help to identify these unintended words and are a good way of becoming acquainted and familiar with the data. Another important step to reduce the semantic dimension is to remove

[55] http://www.lextek.com/manuals/onix/stopwords2.html (latest access: 11.10.2016)

[56] http://www.socialmediatoday.com/content/top-twitter-abbreviations-you-need-know (latest access: 10.09.2016) There are no official dictionaries or abbreviation lists but this homepage has a quick overview about common abbreviations.

[57] Image courtesy of Twittonary.com

words with less than three characters because these words behave like stop words and are very unlikely to contain much meaning to the sentence. In some cases, it is worth considering excluding words with even less than four characters, but it highly depends on the given dataset. In the same way, words, which occur only twice in corpora of thousands of words, are very likely to have no meaning or result out of heavy typing errors, the slang mentioned above words or emoticons.

The next processing step is morphology. To get the smallest unit of language that has meaning (morpheme[58]) the words need to be reduced to their word stems. They are also called free morphemes, since they can exist without adding affixes. To give an example, the word 'snowing' would be reduced/stemmed to the word root 'snow'. In this thesis, the stemming was processed with the well-known Porter stemming algorithm (Porter, 1997) and a snowball-stemmer for different languages. Both algorithms are designed to eliminate known suffixes in English, German, French and much more. The author used the same languages as for the stop word lists to remain consistent.

In the case of the Twitter sample, the number of Tweets relating to the floods is very low and can hardly be compared with the high numbers of Tweets talking about Hurricane Sandy in New York, for example. Despite the fact that the City of New York has a lot more daily Twitter messages than Germany it also relies on the character of the event. Floods can last for months and therefore do not produce those short-term peaks, which are so characteristic for earthquakes or the mentioned Hurricane Sandy. In fact, Fuchs (et al., 2013) proved that event detection relating to the high concentration of Tweets does not work for the current flood dataset. These findings reinforced the author's assertion how to push the few Tweets related to the floods to be a bit more represented in the data, which is especially important for the later topic modelling. For this reason, a synonym list was created and implemented into a separate python function in order to replace the terms 'flood, Hochwassersituation, Hochwasser2013,

[58] https://en.oxforddictionaries.com/definition/morpheme (latest access: 12.12.2016)

Elbehochwasser, Flut, Deich, Sandsäcke' with the word 'Hochwasser'.[59] This resulted in a significant improvement of the LDA classification process and is a sophisticated way to highlight topics, which are otherwise underrepresented in the data. Of course, all stop words or other filtered tasks are only removed for the later calculations and still remain untouched in the database. The development of the filter techniques is one of the most important and time-consuming tasks, but cannot be underestimated. This follows the principles of 'Garbage in, Garbage out' in the field of computer science.[60]

6.3 Probabilistic Topic Models

As mentioned before, most of the current Twitter research studies concentrated mainly on keyword lists and further manual classification processes. Of course, these methods refer to how most of the modern internet pages are organised by links and keywords. For small samples like the one percent Twitter data it would be possible to manually overlook the entire corpus, identify and classify the whole set in decent time. One of the drawbacks is the time consumption, but the accuracy is, of course, much higher than with machine-based algorithms. In contrast to the small Twitter sample of Germany of the public Streaming API, Social Media is exponential densifying, and if the data is provided by the Firehose API, it is highly possible to expect millions of Tweets per day, of course depending on the research extent. This amount of text messages can no longer be searched and classified by hand because it is most likely that the current disaster is gone for weeks, before the results can be presented. One has to rely on new computational ways to solve these so-called Big Data problems. There are many concepts of searching, organising and analyse these vast amounts of data and one of the most promising approaches is Latent Dirichlet Allocation (LDA), developed by Blei and his team in 2003.

[59] Many thanks to Florian Usländer, who had the original idea using synonym lists and the productive discussion on performing LDA.
[60] https://en.wikipedia.org/wiki/Garbage_in,_garbage_out (latest access: 12.07.2016)

6.4 Latent Dirichlet Allocation

LDA is a generative probabilistic model based on a three level Bayesian model, '*in which each item of a collection is modelled as a finite mixture over an underlying set of topics. Each topic is, in turn, modelled as an infinite mixture over an underlying set of topic probabilities. In the context of text modelling, the topic probabilities provide an explicit representation of a document.*' (Blei et al., 2003)

Furthermore, it is an unsupervised machine learning model, which identifies latent topics and assesses semantic similarities. There has been much successful research in analysing vast amounts of Twitter data in the last years, which proved that this method could reduce the semantic dimensions tremendously. (Steiger et al., 2015a) LDA is also predestined for large unseen discrete data sets like the existing Twitter text corpora but is also successfully applied for population genetics (Pritchard et al., 2000) or image classification (Sivic et al., 2008).

In the probabilistic model, topics are defined as a multinomial distribution over words and are assumed to have been drawn from a Dirichlet distribution $\beta_k \sim Dirichlet\ (\eta)$. LDA assumes a generative process for each document d. It first draws a distribution over topics $\theta_d \sim Dirichlet\ (\alpha)$. Furthermore for every word I in the document, a topic index $z_{di}\ \epsilon\ \{1, \dots, K\}$ is drawed from the topic weights $z_{di} \sim \theta_d$ and therefore draws the observed word w_{di} from the selected topic $w_{di} \sim \beta_{zdi}$. (Hoffman et al., 2010)

The generative process for LDA is corresponding to the following joint distribution:

$$p(\beta_{1:k}, \theta_{1-k}, z_{1:D}, w_{1:D}) = \prod_{i=1}^{K} p(\beta_i) \prod_{d=1}^{D} p(\theta_d) \left(\prod\nolimits_{n=1}^{N} p\left(z_{d,n} \middle| \theta_d\right) p\left(w_{d,n} \middle| \beta_{1-k}, z_{d,n}\right) \right)$$

The graphical representation of the above description can be displayed in the following way according to Blei (2012):

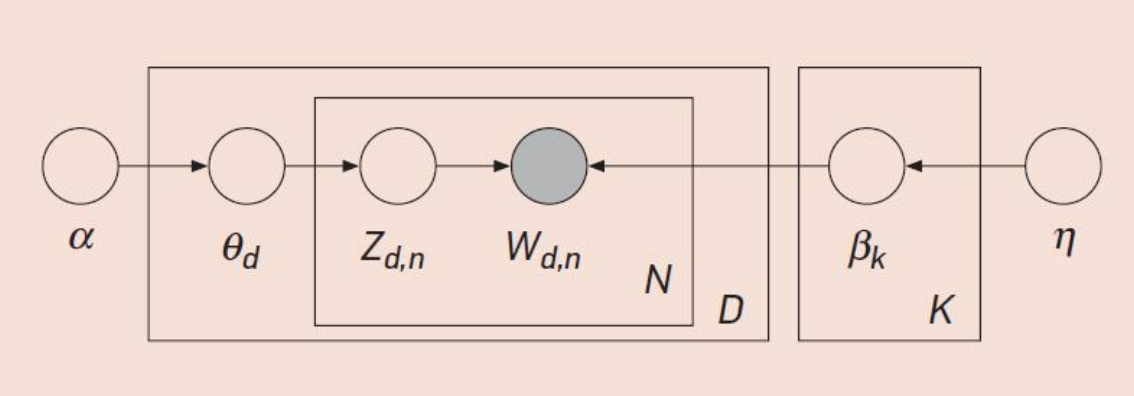


Figure 25. Graphical LDA model according to Blei(2012)

The model assumes that the Tweet corpora contain a random number of latent topics per document α, whereby each topic is characterized by a distribution over words β. Both parameters are corpus-level parameters, assumed to be sampled once in generating the corpus. The θ_d variables are represent the document-level parameter, which is also sampled once per document, while the variables z_{dn} and w_{dn} are the word-level variables which are sampled for each word in each document. The rectangles stand for 'plate' notation, which denotes replication. The outer plate **D** denotes the collection of documents, within the collection while **N** is denoting the words within the documents. One of the main advantages of LDA is the scalability and that millions of datasets can be calculated, while especially simple keyword filtering techniques reach their limits rather quickly. Another assumption of LDA is the 'bag of words' model, which assumes that the order of the words in the document does not matter. On the one hand this assumption is quite unrealistic, while on the other hand it makes perfect sense for uncovering the coarse semantic structures of documents (Steiger et al., 2015b).

For this paper, the online LDA model of Hoffman was used because of its ability to stream and run in constant memory, two factors not to be underestimated, when performing large unseen datasets (Hoffman et al., 2010). The training documents may come in sequentially, and the algorithm can perform Twitter corpora larger than RAM and has the capability of distributed computing to speed up the calculation. One of the main difficulties is the intractable character of the posterior distribution and the parameter estimation of the hidden variables. The posterior distribution cannot be computed because the number of possible topic structures is exponentially large. Online LDA tries to solve the problem with variational Bayes inference, which optimises a parameterized family of distributions over the hidden structure and finds that one that is closest to the posterior. (Blei, 2012)

6.5 Technical Implementation of LDA

After performing the main pre-processing steps on each day of the used sample, the actual computing of the topic modelling was taken into account. All parts of the LDA computing were performed with the **genism** library, which is one of the most efficient topic modelling software is for Python and fits many ways of unsupervised semantic modelling from plain texts.[61] As stated before the LDA parameter estimation can be a very difficult task, because there is no standardised way how to determine the three main parameters:

k = the number of topics in the corpora

α = indicates how many topics a document potentially has. The lower the value of alpha is chosen the lower the number of topics per document is estimated.

β = indicates the number of words per document. The lower the value of Beta is, the lower the number of words per topic.

The default method to estimate the hyperparameters (alpha and beta) is a symmetric 1.0/number_of_topics (k) prior. Several additional hyperparameters

[61] https://radimrehurek.com/gensim/models/ldamodel.html (latest access: 02.01.2017)

values ranging from 0.1 to 0.001 were computed both for the alpha and beta parameters, but the most valuable results were achieved with the default method. There is no best practise to determine the number of topics (k) in topic modelling, and so the author tested the range of 3 to 50 topics and matched the results with the containing flood Tweets, to get a first overview, which can be evaluated and classified in further steps. Too few topics (<6) resulted in very similar topics to one another while computing with more than 30 topics led into a wide spread of the flood relating Tweets in the calculated topics. The best results were achieved within the range of 6 to 10 topics, and the attempts with eight topics performed best on all days. Furthermore, the whole computing of the data with the various parameters lasted for nearly three weeks and generated hundreds of single result documents. For briefness the two days 3rd June and 9th June will be examined in detail with eight topics because they mark the two peaks of the severe flood of 2013 in Germany and are also reflecting the two main water peaks of the flood-regions Danube/Isar and Elbe/Saale. (**See Fig. 13**) In contrast to the batch LDA method, which processes the whole corpus in one pass, updates the corpus and then passes the next full run, the used online LDA method takes small chunks of documents and updates the LDA model in a self-chosen interval. The so-called chunksize was set to one thousand, which updates the LDA model every thousand documents. To give an example the dataset from the 3rd June contains 17.179 single Tweets (documents) and if one regular pass is executed the online LDA method will have done 18 updates. This helps to improve the model estimation and prevents from too much topic drifts. To get the most accurate results, all Tweet corpora were passed fifty times updating every thousand Tweets for a total of 900 times. The author also tried to improve the estimation by increasing the number of passes, but even 200 additional full runs did not improve/change the results. After the computation, the LDA model is first stored as a separate file, which is imported into the DB and every Tweet is matched over a predefined foreign key. Furthermore, the algorithm allocates every single Tweet with the calculated topic and estimates an intern probability how well the content of the Tweet matches the topic itself. To give an example, a SQL-Join query was carried out, and the table below represents a flood-related Tweet of the 3rd June, which was correctly matched by the algorithm.

gid	time	Tweet text	topic number	probability
162059	'2013-06-03 12:33:19'	'Harte Nummer '@tobiasgillen: Heftig! #Passau mit und ohne #Hochwasser als GIF http://t.co/DKtwJFo9bm via @christianmutter"	1 (flood-topic)	0.70833333163

The Tweet has a probability of 70.83% that the content is classified correct in topic one, which is the flood-related topic as one can see in the topic results of the 3rd June. (**see Fig. 30**)

6.6 Spatial Extraction and Filtering

Beside the attempts to evaluate and compute the datasets by time periods of one day, also a spatial filtering was provided based on the remote sensing data of the ZKI-DLR. Recent studies of the flood 2013 from (Albuquerque et al., 2015) and (Peters and Albuquerque, 2015) showed that '*At distances ≤10 km, tweets near strongly affected catchments with a relative water level of +0.75 m were 54 times as likely to be on-topic as tweets in proximity to unaffected catchments with a relative water level of -0.75 m.*' Moreover, this effect was measurable up to ≤30 km and reflects the geographic approach mentioned in **Chapter 2**. The author built on this analysis and created a 30 km buffer around the water masks. Furthermore, all Tweets which are intersecting the buffer zone were extracted and stored in the PostgreSQL database. Like the full dataset, the table was split into tables with a one day period. This filter technique reduced the amount of Twitter messages tremendously and therefore speeded up the computing time from hours to minutes.

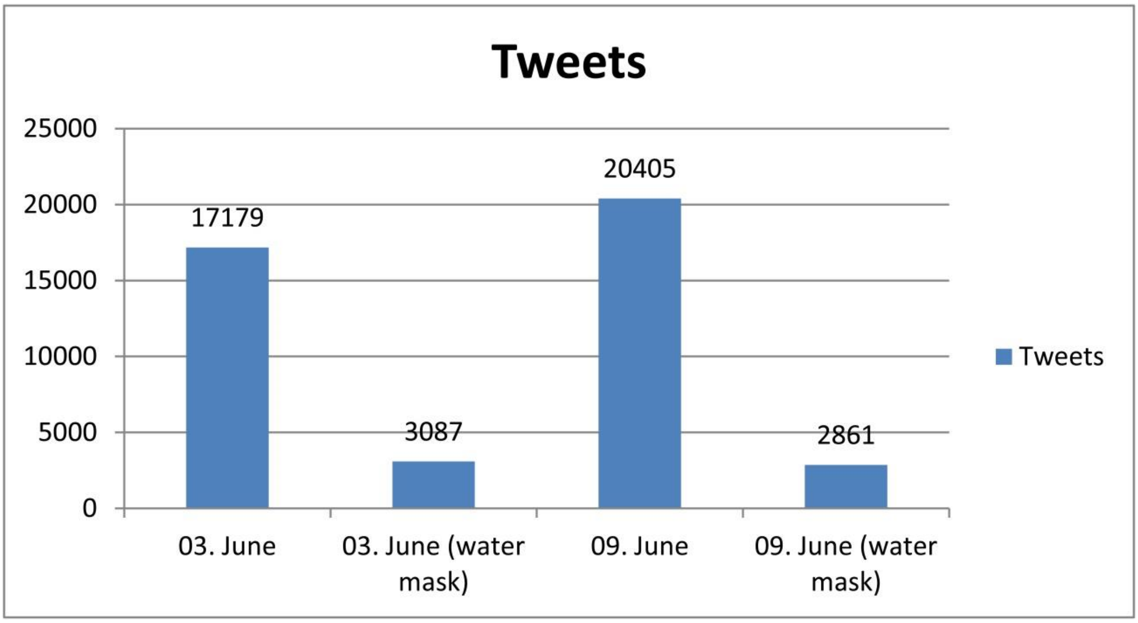


Figure 26. Comparison of the full dataset and the spatial extracted Tweets, which intersect the 30 km buffer of the ZKI-DLR water masks.

Concentrating on the spatial extent of the flood event has some major advantages compared with computing the whole dataset. Major cities like Berlin were not affected by severe floods of 2013, but the capital of Germany is responsible for nearly 1/8 off all Twitter messages in the dataset. Of course, many people in Berlin tweeted about the flood, most of the time showing their solicitousness with the flood affected people but it is very unlikely that a Tweet from Berlin contains critical on ground information. Therefore, concentrating on the buffer areas around the remote sensing data could provide faster and finer grained additional information on the hazard areas. This is especially important for relief units like the Bundesanstalt Technisches Hilfswerk (THW)[62] who already work with map products from the ZKI-DLR.

[62] https://www.thw.de/DE/Startseite/startseite_node.html (latest access: 12.06.2016)

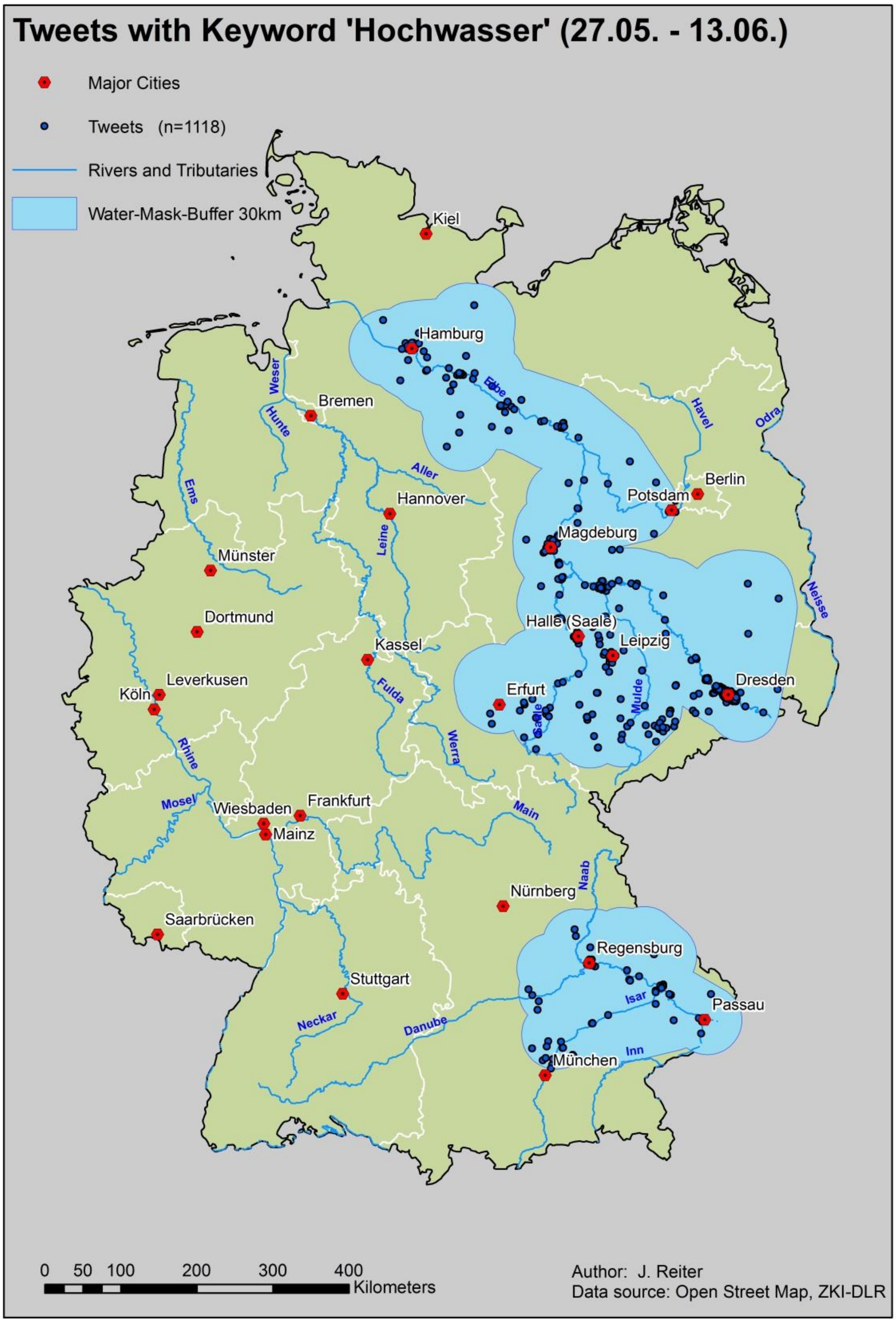


Figure 27. 30 km buffer on the water masks provided by the ZKI-DLR and visualisation of all Tweets containing the word 'Hochwasser' and are intersecting the buffer area.

6.7 Classification Process and Confusion Matrix

The whole dataset is therefore classified manually in on-topic Tweets which are related to the floods and off-topic Tweets, which are not related. Under these circumstances, the low number of Tweets per day in the dataset of Germany allowed a manual classification, which would otherwise be extremely time-consuming. The evaluation process of the manual classification and the classification by the LDA computation is therefore reviewed by using a confusion matrix, in order to check how well the pre-processing and LDA algorithm automatically classified the Tweets in the calculated topics. The confusion matrix is a common practise for checking the performance of an algorithm in the field of machine learning. Not only does it possible to visualise the table for a better understand, but also it allows checking if the algorithm mismatches or confuses two or more predicted classes. This mislabelling can be further measured with established classification functions like *recall, precision, negative predicted value, specificity, Cohen's kappa* and the *overall accuracy*. The layout of the matrix depends on the number of classes defined (N x N). In our case, a 2 x 2 matrix was established with two rows and two columns that report the number of *true positives*, *false negatives*, *false positives* and *true negatives*. This leads to a more detailed analysis of the data and prevents from misleading labelling if the classes are very unbalanced in size (Stehman, 1997). This is the case in our Twitter dataset, where the flood-related Tweets are only a small share

The **predicted condition positive** can be formulated as follows (**see Fig.28 and Fig.29**):

True positive (t_p) = the Tweet is on-topic (relating to the flood) and the algorithm has classified it correct

False negative (f_n) = the Tweet is on-topic, but the algorithm has classified it misleadingly as off-topic

The **predicted condition negative** can be formulated as follows:

False positive (f_p) = the Tweet is off-topic, but the algorithm has classified it misleadingly as on-topic

True negative (t_n) = the Tweet is off-topic, and the algorithm has classified it correct

		Predicted condition	
	Total Tweets	Predicted condition positive/ on-topic	Predicted condition negative/ off-topic
True condition	Condition positive/on-topic	True positiv	False positve
	Condition negative/off-topic	False negative	True negative

Figure 28. Confusion Matrix

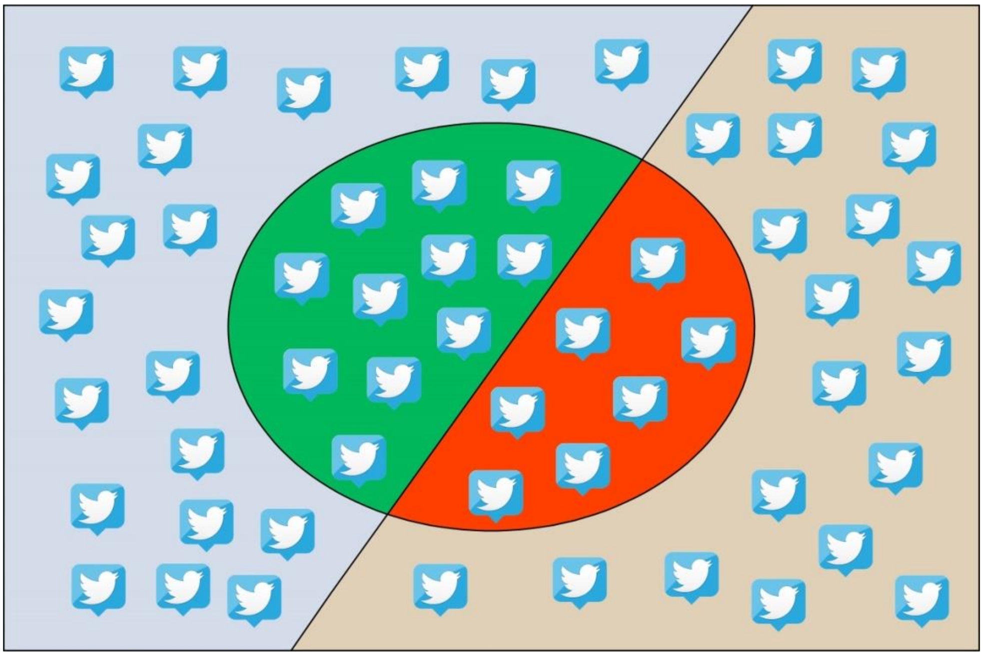


Figure 29. Assuming to the explanation above every Twitter-logo is representing a single Tweet. The oval represents the flood-topic calculated by the LDA algorithm, with *true positives* and *false positives*, while the blue and light grey areas represent all other topics with the false negatives and true negatives.

6.7.1 Classification Functions of the Confusion Matrix

Recall/Sensitivity = $\frac{tp}{tp+fn}$

The *recall/sensitivity* measures the proportion of positives that are correctly classified by the algorithm. (For example, the percentage of flood-related Tweets, which are correctly identified.) The function is also known as the *true positive rate*.

Specificity = $\frac{tn}{tn+fp}$

The *specificity* measures the proportion of negatives that are correctly classified by the algorithm. (For example, the percentage of all off-topic Tweets, which are correctly identified.)

Precision = $\frac{tp}{tp+fp}$

Precision indicates the proportion of the results correctly recognised as positive for all the positive results. (For example, it shows how many Tweets of the flood-topic are relevant.)

Negative predicted value (NPV) = $\frac{tn}{tn+fn}$

The *negative predicted value* indicates the proportion of the results correctly recognised as negatives for all the negative results. (For example, the proportion of the true off-topic Tweets, without the classified on-topic Tweets.)

Overall accuracy = $\frac{tp+tn}{tp+fn+fp+tn}$

The *overall accuracy* provides the proportion of correct guesses. Therefore, it is very important to take into account that the accuracy value cannot be considered alone as a metric of the performance of the algorithm. Especially if the dataset classes are very unbalanced (like the small flood-topic), the result should only be contemplated in combination with the preceding functions.

Cohen's Kappa = $\frac{P_o - P_e}{1 - P_e}$

The *Cohen's* (1960) *kappa* coefficient measures inter-class agreement for qualitative items, in our case the manual evaluation and the unsupervised LDA algorithm.

$P_o = \frac{(tp+tn)}{(tp+fp+tn+fn)}$

Class a: $\frac{(tp+fp)*(tp+fn)}{(tp+fp+fn+tn)}$

Class b: $\frac{(fn+tn)*(fp+tn)}{(tp+fp+fn+tn)}$

$P_e = (a + b) / (tp + fp + fn + tn)$

Beside the before addressed statistical methods to evaluate the confusion matrix the k-method is thought to be more robust than simple percentage agreement calculations because the kappa coefficient takes into account the possibility of the agreement occurring by chance. Furthermore, it compares the accuracy of the system to the accuracy of a random system.[63] If k = 1 the classes are in complete agreement while k <= 0 means that there is no agreement other than what would be expected by chance (Bortz, 1999). The interpretation of k is considered by (Landis and Koch, 1977) into five classes: 0-0.2 slight, 0.21-0.40 as fair, 0.41-060 as moderate, 061-0,8 as substantial and 0.81-1 as almost perfect, while Fleiss (et al., 2003) considers k > 0.75 as excellent, 0.40-0,75 as fair to good and < 0.40 as poor. In the case of machine learning, all values above 0.40 are exceptional.

[63] http://standardwisdom.com/softwarejournal/2011/12/confusion-matrix-another-single-value-metric-kappa-statistic/ (latest access: 12.01.2017)

7 LDA Computing and Evaluation in a Confusion Matrix

The results of the present study, especially the LDA computing, were evaluated within a Confusion Matrix testing the performance of the used probabilistic topic modelling algorithm. As every chosen dataset is very different compared to each other, the single results are discussed immediately after the Confusion Matrix. However, a general discussion is provided in **Chapter 8**.

7.1 3rd June Corpus

dataset	**3rd June-corpus**
documents (Tweets)	17.179
features	9.408
non-zero-entries	54.556
passes	50
update-chunks	every 1.000 documents
k (number of topics)	8
alpha	0.125
beta	0.125

17.179 single documents (Tweets) with 9408 features and 54556 non-zero entries were accepted by the LDA algorithm. Therefore, an online LDA training with eight topics and 50 passes was performed over the supplied corpus. The algorithm updates the model once every 1000 Tweets evaluating perplexity every 17.000 documents and iterating 50 times with a convergence threshold of 0.001. The Alpha and Beta parameters used the calculated values at 0.125 each in a symmetric prior. A simple word frequencies analysis reveals that on the 3rd June the word 'Hochwasser' is the second most used word behind Berlin. It is also remarkable that Passau, one of the most flood-affected cities in Germany is the 9th most used word. In this respect, there are two flood-related words in the top ten, two because the very small city in Lower Bavaria normally would be irrelevant in a Twitter datasets containing the whole Republic of Germany.

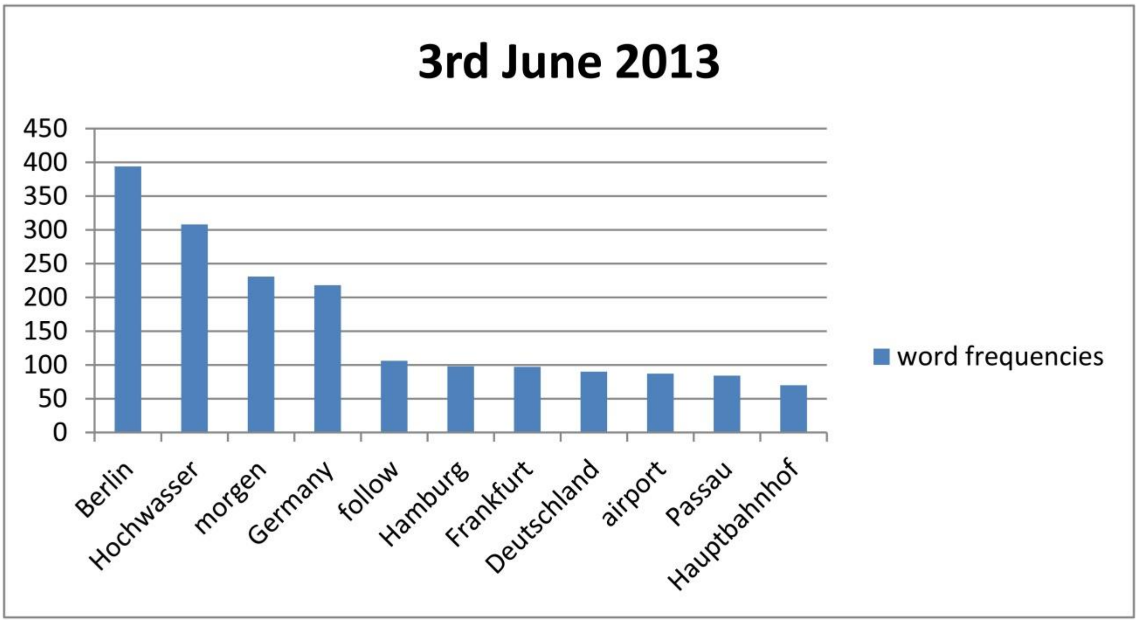


Figure 30. Most frequent words of the 3rd June dataset

LDA results and interpretation – Topics: 3rd June 2013[64]

*(0, '0.023*kwhkwp + 0.023*stund + 0.022*aktuel + 0.022*today + 0.021*kommt + 0.021*leistung + 0.021*gesamt + 0.020*happi + 0.018*schule + 0.018*menschen + 0.016*glück + 0.016*warum + 0.014*hoffentlich'*

(1, '0.183*hochwass + 0.050*dresden + 0.048*passau + 0.028*wasser + 0.020*regensburg + 0.020*magdeburg + 0.013*steigt

*(2, '0.040*magento + 0.035*hauptbahnhof + 0.026*great + 0.025*hamburg + 0.020*ashton + 0.016*party + 0.013*repli + 0.011*isemarkt*

*(3, '0.078*morgen + 0.029*justinnoticestratfordwav + 0.018*hotel + 0.017*summer + 0.013*thing + 0.012*goodnight + 0.010*torwandschiessen + 0.010*straub + 0.009*facebook'*

*(4, '0.027*airport + 0.024*munich + 0.019*berlin + 0.015*intern + 0.014*strauß + 0.014*franzjosef + 0.011*english + 0.010*bayern'*

*(5, '0.081*hamburg + 0.065*germani + 0.045*leipzig + 0.022*water + 0.018*follow + 0.016*gruenderhh + 0.015*school + 0.012*alert + 0.009*weather*

[64] The topic results are presented after the stemming to show the exact computed result. Because of this the word 'hochwasser' is visualized as 'hochwass' or the word 'happy' as 'happi'. Moreover, all words are lowercase. (For an explanation of stemming see **Chapter 6**.)

*(6, '0.028*birthday + 0.018*german + 0.012*moment + 0.011*stehen + 0.010*schlafen + 0.009*endomondo + 0.009*schönen + 0.009*geburtstag + 0.009*regeln*

*(7, '0.020*leben + 0.014*problem + 0.012*campus + 0.011*sicher + 0.009*sweet + 0.009*warten + 0.008*deutschland + 0.008*universe*

Twitter topics computed by LDA are never as accurate as performing the algorithm on big corpora like the Science magazine (Blei, 2012). One has to be very creative in labelling each topic with a headline. Nevertheless, on the 3rd June, the flood-related topic can be clearly identified in topic number one.

(1, '0.183*hochwass + 0.050*dresden + 0.048*passau + 0.028*wasser + 0.020*regensburg + 0.020*magdeburg + 0.013*steigt

As the LDA model embodies the Tweets (single documents) as a mixture of topics these words are the most likely to have generated the original Tweets (text corpora). It can be seen from the topic 'one', that all of the words could be potentially be related to the floods. The four mentioned cities are the most affected cities on that date and the additional words 'wasser' and 'steigt' can be dedicated to floods because it is evident that the water was rising. (for the development of the floods see **Chapter 5**) Furthermore, the likelihood is fairly good between the words, in the context of Twitter analysis, where the percentage is rather low compared to long text documents.

Confusion Matrix for the 3rd June

3rd June	Predicted condition positive/on-topic	Predicted condition negative/off-topic	Classification overall	
Condition positive/ on-topic	295	482	777	*Precision: 37.967%*
Condition negative/off-topic	142	16260	16402	*NPV: 99.134%*
Truth overall	437	2848	17179	
Overall accuracy: 96.368%	*Recall: 67.506%*	*Specificity: 97.121%*		*Kappa: 0.469*

The confusion matrix for the 3rd June shows how important it is to evaluate not only the overall accuracy, which with 96.368% would be an exceptionally good value. As the matrix reveals there are only 437 on-topic Tweets in the whole dataset of which 295 were classified correctly by the LDA algorithm and 142 were missed, which leads to a *recall* rate of 67.50%. In the flood-related topic one, 777 Tweets were labelled as on-topic of which there were 482 false positive. The figures here reveal a 37.69% *precision* rate. Furthermore, the *NPV* percentage of 99.13% and the *specificity* 97.12% shows that with 16260 true negative Tweets the LDA model identified the majority of all off-topic Tweets correct. It can be seen from the *kappa* value of 0.469, which is the most robust statistic, which for an unsupervised machine learning algorithms the results demonstrate a pretty good classification rate.

7.2 9th June Corpus

dataset	9th June-corpus
documents (Tweets)	20.405
features	11.093
non-zero-entries	66.753
passes	50
update-chunks	every 1000 documents
k (number of topics)	8
alpha	0.125
beta	0.125

20.405 single documents (Tweets) with 11.093 features and 66.753 non-zero entries were accepted by the LDA algorithm. Therefore, an online LDA training with eight topics and 50 passes was performed over the supplied corpus. The algorithm updates the model once every 1000 Tweets evaluating perplexity every 20.000 documents and iterating 50 times with a convergence threshold of 0.001. The Alpha and Beta parameters used the calculated values at 0.125 each in a symmetric prior. While in previous corpus the word 'Hochwasser' was in the second place of the most frequent words, this time it takes the fourth place with hardly over 200 entries. Sadly, no second word can be related to the floods, and even no flood affected city, like Magdeburg or Dresden, is mentioned in the Top 10.

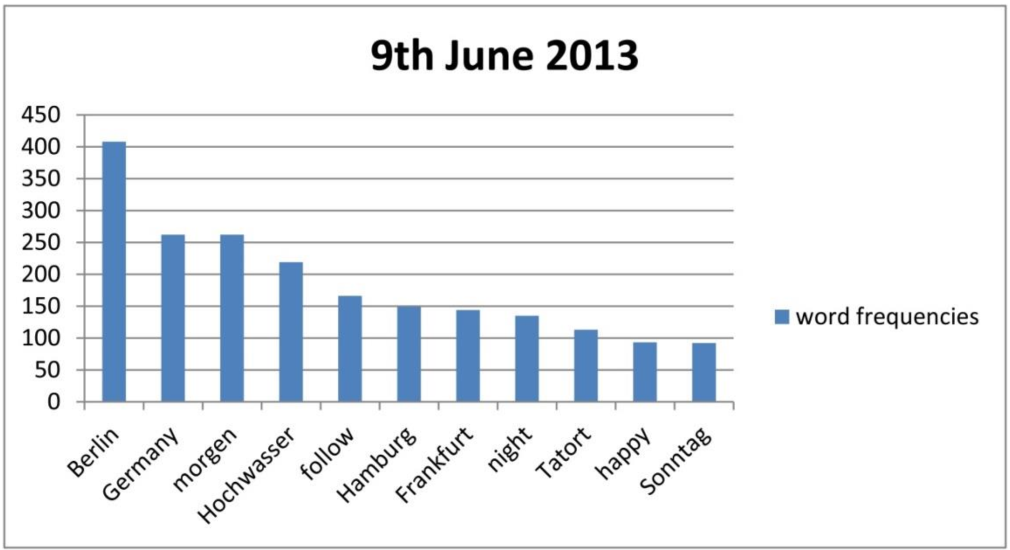


Figure 31. Most frequent words of the 9th June dataset

LDA results and interpretation – Topics: 9th June 2013

(0, '0.048*frankfurt + 0.032*airport + 0.019*hauptbahnhof + 0.017*richtig + 0.016*cancer + 0.015*weekend + 0.015*schule + 0.014*birthday')

(1, '0.035*tatort + 0.020*bunlar + 0.019*abend + 0.018*sonntag + 0.016*olympiastadion + 0.015*erkegi + 0.013*çabulcu + 0.013*türkiy

(2, '0.122*morgen + 0.112*berlin + 0.047*follow + 0.026*hamburg + 0.026*monday + 0.018*school + 0.017*arbeit + 0.012*tomorrow

(3, '0.025*starbuck + 0.023*schlafen + 0.019*sunday + 0.016*alemania + 0.016*gestern + 0.015*kommen + 0.015*lecker + 0.013*depechemod

(4, '0.088*germani + 0.032*munich + 0.027*deutschland + 0.023*mannheim + 0.015*schöne + 0.013*urlaub + 0.013*düsseldorf'+ 0.011*hamburg

(5, '0.024*folgen + 0.021*eigentlich + 0.018*menschen + 0.013*платье + 0.013*thought + 0.010*heidelberg + 0.011*session + 0.011*gesehen

(6, '0.030*montag + 0.028*münchen + 0.024*arbeit + 0.020*german + 0.016*besser + 0.016*endlich + 0.015*justin + 0.013*düsseldorf')

(7, '0.063*hochwass + 0.028*lostau + 0.024*frühstück + 0.021*awesom + 0.018*wetter + 0.018*magdeburg'+ 0.016*minuten + 0.015*fahren

Several noteworthy results can be made out of the LDA computation even if there are not that many flood-related words in the word frequencies analysis. Topic Seven can be clearly identified as the flood-topic, even if the results consists not entirely of distinct words. Lostau as a district of the city Magdeburg was heavily hit by the severe floods on the day before, and many Tweets refer to relief units and volunteers trying to build up dikes with sandbags. (**see Fig. 12.**) The belonging to the topic and the percentage of the likelihood of the single words is slightly worse than on the 3rd June. Words like '*Frühstück*', '*Minuten*' or '*fahren*' are difficult to interpret, because there are flood related Tweets, which contain those words. To give an example '*@welt Von Berlin bis Hannover plant die Bahn aktuell mit 70 Minuten Verspätung. Warte zurzeit auf #ice644. #hochwasser*' Knowing the data and the contents can help to improve the understanding of the LDA topics in the context of the small 140 character messages.

Confusion Matrix for the 9th June

9th June	Predicted condition positive/on-topic	Predicted condition negative/off-topic	Classification overall	
Condition positive/ on-topic	178	546	724	*Precision:* 24.586%
Condition negative/off-topic	75	19606	19681	*NPV:* 99.619%
Truth overall	253	20152	20405	
Overall accuracy: 96,957%	*Recall:* 70.356%	*Specificity:* 97.291%		*Kappa:* 0.352

As the results of the 3rd June, the *overall accuracy* of 96.95% can be misleading as the 253 on-topic Tweets are very small in number compared to the total of 20.405 Tweets. The recall rate of 70.35% is a bit higher, but the *precision* of 24.58% reveals that there are more *false positives* in the flood-topic. Nevertheless, the *NPV* and *specificity* percentage of 99.61% and 97.291% shows that like the 3rd June dataset the majority of all off-topic Tweets were labelled correctly. With a *kappa* value of 0.352 the LDA model performs a bit weaker than before but can be designated as fair good again in the case of machine learning. Due to the low number of on-topic Tweets compared to the 3rd June these numbers are not unsurprising.

7.3 30 km Buffer - 3rd June Corpus

dataset	3rd June-corpus
documents (Tweets)	3087
features	1910
non-zero-entries	8002
passes	50
update-chunks	every 1000 documents
k (number of topics)	5
alpha	0.2
beta	0.2

3.087 single documents (Tweets) with 1910 features and 8002 non-zero entries were accepted by the LDA algorithm. Therefore, an online LDA training with five topics and 50 passes was performed over the supplied corpus. The algorithm updates the model once every 1.000 Tweets evaluating perplexity every 3.087 documents and iterating 50 times with a convergence threshold of 0.001. The Alpha and Beta parameters used the calculated values at 0.2 each in a symmetric prior. In contrast to the frequencies analysis of whole Germany, the Tweets, that intersect the 30 km buffer of the DLR-ZKI water masks contain almost 6 of 10 words, which can be related to the flood event in one way or another.

‘Hochwasser’ almost doubles the second placed word ‘Hamburg’. Also Passau, Dresden and Leipzig can be associated with the flood because all three cities struggled with high waters. Of course, Dresden and Leipzig occur in off-topic messages, but the predominant theme is the flood on that day. The words ‘Wasser’ and ‘flood’ are therefore self-evident.

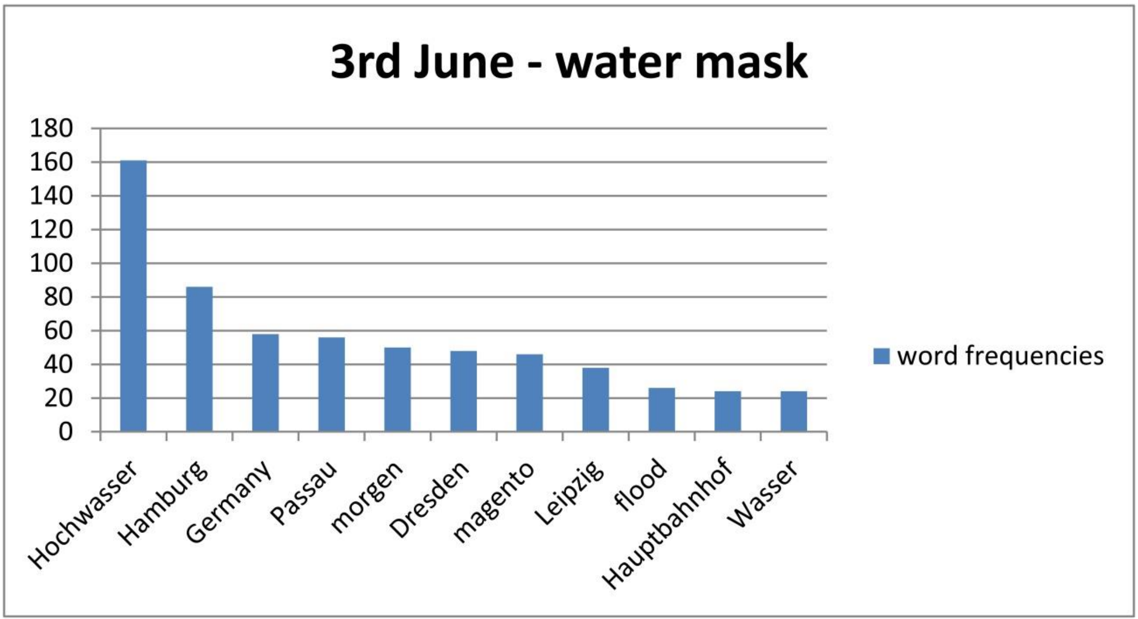


Figure 32. Most frequent words of the 3rd June buffer dataset

LDA Results and Interpretation – Topics: 3rd June 2013 (30 km buffer)

*(0, '0.018*today + 0.016*start + 0.016*kwhkwp + 0.016*stund + 0.016*solar + 0.015*birthday + 0.015*aktuel + 0.015*leistung*

*(1, '0.018*airport + 0.016*munich + 0.013*menschen + 0.013*berlin + 0.011*tomorrow + 0.010*friend + 0.009*strauß + 0.009*franzjosef*

*(2, '0.047*morgen + 0.037*germani + 0.028*leipzig + 0.019*magento + 0.010*hoffentlich + 0.010*summer + 0.010*glück + 0.010*school + 0.009*regen + 0.008*steht + 0.008*kommt + 0.007*schule + 0.007*alert')*

*(3, '0.067*hamburg + 0.018*justinnoticestratfordwav + 0.015*leben + 0.012*follow + 0.010*gruenderhh + 0.007*goodnight + 0.007*dicht + 0.007*isemarkt + 0.006*campus*

(4, '0.129*hochwass + 0.042*passau + 0.035*dresden + 0.024*flood 0.020*wasser + 0.014*regensburg + 0.014*magdeburg + 0.007*sachsen + 0.007*donau

Unlike the eight topics of the full dataset, only five topics were computed for the spatial extracted data, with respect to the much lower number of total Tweets. Topic four can be clearly identified as the flood-related topic. All of the nine most related words are imaginable with the high waters, and the percentage of the likelihood is almost comparable to the full dataset. Passau as the most affected city on this day is well represented and even the river Danube which caused the disaster is mapped in the results. Regensburg, Dresden, Magdeburg were also heavily hit by the waters. It is remarkable that even the English term flood is represented in the mostly German written data.

Confusion Matrix for the 3rd June (Water Masks)

3rd June (water masks)	*Predicted condition positive/on-topic*	*Predicted condition negative/off-topic*	*Classification overall*	*Precision*
Condition positive/ on-topic	*193*	*229*	*442*	*Precision: 45.735%*
Condition negative/off-topic	*46*	*2619*	*2665*	*NPV: 98.274%*
Truth overall	*239*	*2848*	*3087*	
Overall accuracy: 90.092%	*Recall: 80.753%*	*Specificity: 91.959%*		*Kappa: 0.538*

From the 437 on-topic Tweets dispersed all over Germany 239 flood-related Tweets remain after the spatial extraction by the 30km buffer around the water masks. The figures of the table reveal that the *recall* rate of 80.75% is almost 13% higher than in the full dataset and also the *precision* rate is about 7% increased. The *NPV* and *specificity* value, 98.27% and 91.95%, however, are slightly worse. This could be due to the low number of Tweets in total. According to the statistics, the *kappa* value of 0.538 marks the highest positive agreement of all provided datasets. This finding is consistent with the before made prediction that if there are many flood-related Tweets, the LDA model performs way more effective than in very unbalanced datasets. This special case is described further in the conclusion of this thesis.

7.4 30 km Buffer – 9th June Corpus

dataset	**3rd June-corpus**
documents (Tweets)	2861
features	2030
non-zero-entries	8275
passes	50
update-chunks	every 1000 documents
k (number of topics)	5
alpha	0.2
beta	0.2

2861 single documents (Tweets) with 2030 features and 8275 non-zero entries were accepted by the LDA algorithm. Therefore, an online LDA training with five topics and 50 passes was performed over the supplied corpus. The algorithm updates the model once every 1.000 Tweets evaluating perplexity every 2.861 documents and iterating 50 times with a convergence threshold of 0.001. The Alpha and Beta parameters used the calculated values at 0.2 each in a symmetric prior. Unlike the results of the spatial extracted dataset of the 3rd June, where six words could be related to the floods, only 'Hochwasser' and probably

‘Magdeburg’ are in the top ten of the most frequent words. At least the former is on place number one with around 160 times.

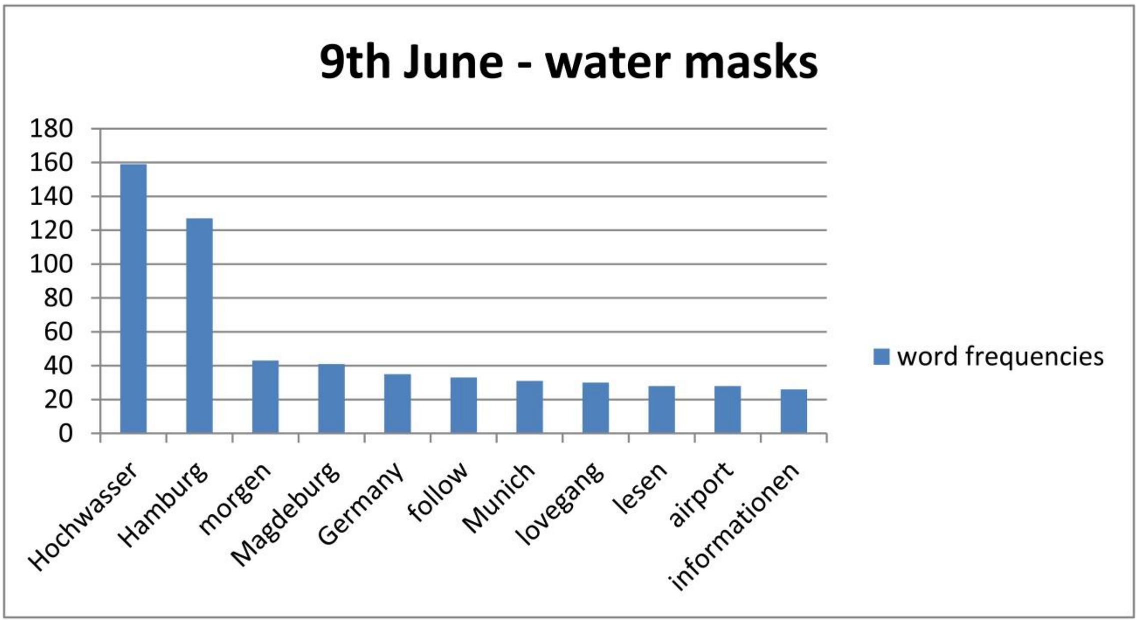


Figure 33. Most frequent words of the 3rd June buffer dataset

LDA Results and Interpretation – Topics: 9th June 2013 (30 km buffer)

(0, '0.033*weather + 0.021*nacht + 0.019*today + 0.017*aktuel + 0.015*hotel + 0.014*kwhkwp + 0.013*stund + 0.012*leistung

(1, '0.016*great + 0.013*friend + 0.011*endlich + 0.011*german + 0.010*psalm + 0.010*leipzig + 0.009*iphon + 0.009*kommt + 0.008*brefra + 0.008*fahren + 0.008*music + 0.007*niemand + 0.007*updat')

(2, '0.105*hochwass + 0.028*magdeburg + 0.015*dresden + 0.014*informationen + 0.014*sicherheitsrelevant + 0.012*lostau + 0.012*passau + 0.010*sandsäcke')

(3, '0.047*morgen + 0.025*munich + 0.015*airport + 0.012*berlin + 0.011*montag + 0.011*kreta + 0.010*abend + 0.008*sonntag

(4, '0.089*hamburg + 0.028*germani + 0.022*night + 0.018*lovegang + 0.017*school + 0.013*hauptbahnhof + 0.013*heart + 0.012*diamond + 0.012*happi)

The number two topic can be clearly assigned as the flood related. It is noticeable that unlike in the whole Germany dataset of the same date, all words of the flood topic are imaginable to have something to do with the high waters. All the called cities were hit by the waters and the words ‘sandsäcke’ and ‘Hochwasser’ are self-evident. Also the words ‘Informationen’ and ‘sicherheitsrelevant’ are present in many flood related Tweets. The percentage of the likelihood that those Tweets are the most relevant in rebuilding the original Tweets is comparable to the before mentioned researches on the other dates.

Confusion Matrix for the 9th June (Water Masks)

9th June (water masks)	**Predicted condition positive/on-topic**	**Predicted condition negative/off-topic**	**Classification overall**	
Condition positive/ on-topic	**156**	**481**	**637**	***Precision: 24.49%***
Condition negative/off-topic	**23**	**2201**	**2224**	***NPV: 98.966%***
Truth overall	**169**	**2692**	**2861**	
Overall accuracy: 82.034%	***Recall: 87.151%***	***Specificity: 82.066%***		***Kappa: 0.315***

From 253 on-topic Twitter messages in the Germany dataset 169 remain after the spatial extraction with the buffered water masks. The *recall* rate of 87.151% is pretty high compared to the full dataset, and only 23 flood-related Tweets were misclassified by the LDA model. On the other hand, the *precision* rate with only 24.49% is almost identically with 24.58% of the 9th June data. While the *specificity* is almost 15% lower, due to the many false positive Tweets, the *NPV* with 98.96% is comparable. With a *kappa* value of 0.315, the spatial extracted dataset of the 9th June holds the red lantern compared to the other results. Even after several additional runs and parameter changes, the results did not get any better. As one reason the much lower number of on-topic Tweets in the data compared to the 3rd June (water masks) can be assumed. Another interesting observation can be made if we concentrate on the flood topic itself. With 637 Tweets the LDA model classified a lot more Tweets to the flood-related topic number two than in the previous dataset and therefore producing many false positive classifications, which are mainly responsible for the lower *kappa* value.

7.5 Cascading LDA

After computing the LDA algorithm and classifying into independent topics, new considerations were made on how to further speed up the computation without losing too many on-topic messages. In the field of Disaster Management time is one of the most important keys especially in the response phase. While the manual classification in on and off topic messages did last for several workdays per day of the flood data, the classification process with LDA speeded up the process tremendously. Evaluation and manual classification of the LDA computed flood topic took about two hours. As mentioned above the calculated flood topic consisted of 25-45% flood-related Tweets. In the case of the 3rd June, the amount of 17.179 Tweets decreased to 777 Tweets in the flood topic, which had to be classified manually. The idea behind cascading LDA (**see Fig. 34**) was to investigate if the same algorithm passes over the predefined flood topic a second time and if it could classify the results any further and to what time and precision costs?

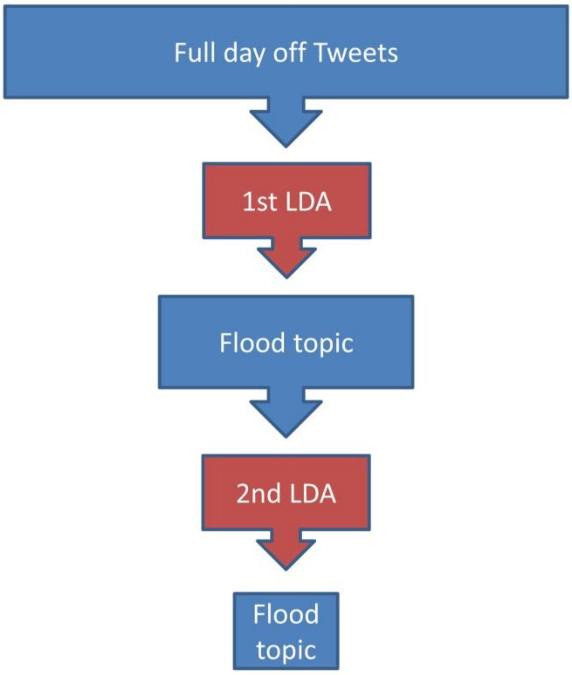


Figure 34. Classification process of the cascading LDA algorithm

For the attempt, the dataset of the 3rd June was taken and the flood topic was exported into a new table in the DB.

7.6 Cascading LDA – 3rd June Corpus (Flood Topic: No. 1)

dataset	**3rd June corpus (flood topic)**
documents (Tweets)	777
features	464
non-zero-entries	2492
passes	50
update-chunks	every 350 documents
k (number of topics)	4
alpha	0.25
beta	0.25

777 single documents (Tweets) with 464 features and 2492 non-zero entries were accepted by the LDA algorithm. Therefore, an online LDA training with four topics and 50 passes was performed over the supplied corpus. The algorithm updates the model once every 350 Tweets evaluating perplexity every 777

documents and iterating 50 times with a convergence threshold of 0.001. The Alpha and Beta parameters used the calculated values at 0.25 each in a symmetric prior. For the sake of completeness a word frequencies analysis was provided like in the previous examples but it is unsurprising that in the recalculated flood topic of the 3rd June the most used word is ‘Hochwasser’ or ‘Passau’ as most affected city on place three.

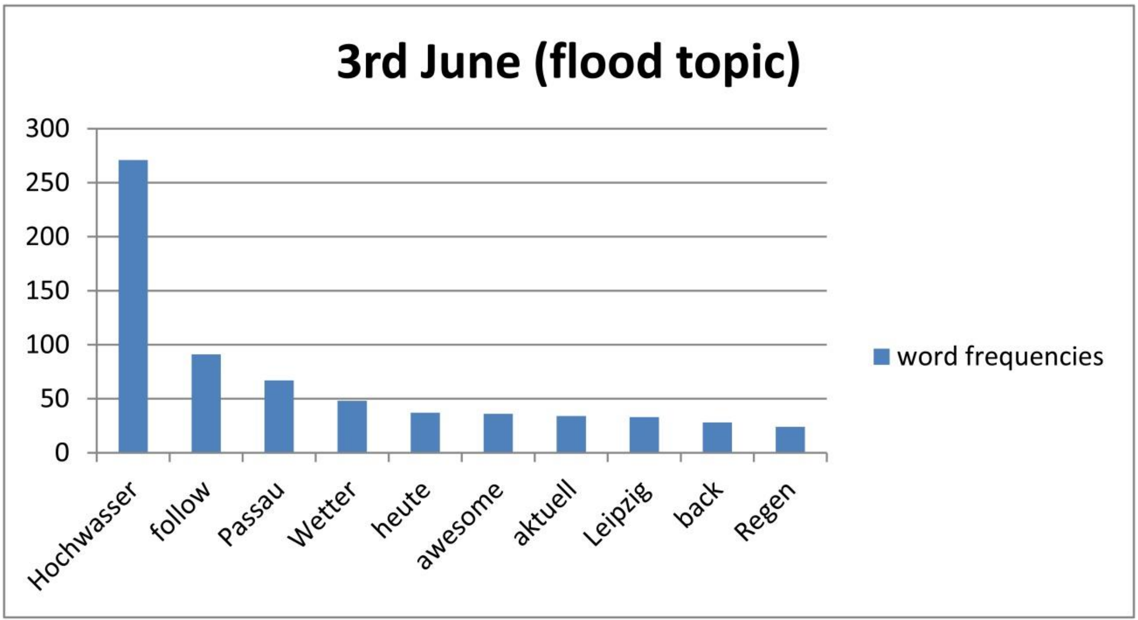


Figure 35. Most frequent words of the 3rd June (only flood topic) dataset

LDA Results and Interpretation – Topics: 3rd June 2013 (flood topic)

*(0, '0.135*follow + 0.042*back + 0.033*regen + 0.028*cologn + 0.023*awesom + 0.023*dormir + 0.023*love + 0.020*brejcha + 0.020*bori + 0.020*tribut*

(1, '0.275*hochwass + 0.024*ffw + 0.024*halle + 0.022*dresden + 0.015*regensburg + 0.013*wasser + 0.012*schlimm + 0.009*magdeburg + 0.009*elbe + 0.008*überschwemmung

*(2, '0.062*wetter + 0.046*leipzig + 0.033*weingarten + 0.028*hunger + 0.017*luftfeuchtigkeit + 0.017*messen + 0.015*rain + 0.015*pm + 0.014*michael*

```
(3, '0.061*passau + 0.039*heut + 0.030*freund + 0.027*aktuel + 0.026*mwh + 0.026*kwh +
0.026*kwhkwp + 0.026*leistung + 0.024*stund + 0.019*flood
```

In this case the topics are more misleading than in the previous first LDA run of 3rd June data sample. While it seems that topic one is unambiguously the computed flood topic with very high percentages within the topic itself, topic three starts with the word Passau, the most affected city. Nonetheless all words from topic one can be interpreted in the context of floods. The abbreviation ‘ffw’ means ‘Freiwillige Feuerwehr’ (German Fire Services Association) and is frequently used throughout all days of the disaster event. Furthermore all named cities were affected that day and the word ‘Überschwemmung’ is just another word for flooding. Several additional runs with different topic numbers showed that the topic drift is increasing quickly over < 5 topics and below 3 the spreading of flood related Tweets was unacceptable.

Confusion Matrix for the 3rd June (flood topic)

3rd June / Cascading LDA	Predicted condition positive/on-topic	Predicted condition negative/off-topic	Classification overall	Precision
Condition positive/ on-topic	199	23	222	Precision: 89.64%
Condition negative/off-topic	96	459	555	NPV: 82.703%
Truth overall	295	482	777	
Overall accuracy: 84.685%	Recall: 67.458%	Specificity: 95.228%		Kappa: 0.658

The figures reveal that from the primary 777 Tweets the algorithm classified 222 messages as on-topic. On the one hand, the *precision* rate increased about 52% to 89,64%, because 199 Tweets were *true positives*, while only 23 *false positives* were misclassified in the flood topic. On the other hand, the *recall* rate stayed almost at the same percentage of 67, 45% like in the original full dataset. The *specificity* value decreased around 2% while the *NPV* has fallen from 99.13% to 82.70%. Also, the overall accuracy suffered from the second LDA run and decreased around 11% to 84,68%. Whereas the *overall accuracy* is not as significant for very unbalanced datasets like the present the *kappa* value of 0.658 is by far the best result of all previous LDA runs. On Fleiss (et al., 2003) the value would be characterised as good, and it is without question significant.

8 Results and Discussion

The first research question discusses the assumption of Fuchs (et al., 2013), that, if the data available freely from the Twitter Streaming API providing a random one-percent-example of the total stream is filtered for geolocated Tweets only , one would get the majority of all geolocated Tweets of that particular day. Comparing the one-week Firehose sample (granting 100%) with self-harvested Tweets revealed, that an average of 40% of all geolocated Tweets in Germany can be collected. This is all the more impressive, as the public Twitter-interface only grants for 1% of the current overall Tweet traffic. So if, however, the stream is filtered directly for existing coordinates, the result comprises forty times as many Tweets as guaranteed for free in the first place. Although the datasets cannot be compared entirely, for they were mined in different years, they give at least a rough estimation on what to expect for the spatial Tweet distribution of Germany. Future works should evaluate this observation taking into account more Twitter-oriented countries such as the USA or Brasil, where the percentage of the geolocated Tweets at the moment is much higher than 1% of the total stream.

Furthermore, a classification with the probabilistic topic model called Latent Dirichlet Allocation was provided together with the attempt to extract the Tweets by the enlarged spatial extent of the water masks (ZKI-DLR). The results of the LDA computation and the subsequent manual classification were evaluated in a confusion matrix, so as to verify further and estimate the correctness of the algorithm.

How can probabilistic topic models like LDA be used for automated topic classifications in the field of Twitter messages and how reliable are the results?

Latent Dirichlet Allocation offers a sophisticated approach enriching traditional keyword searches as well as time- and people-intensive manual classification processes. Irrespective of water mask spatial filtering was applied, all presented case studies reveal, that the calculated flood topics are very stable. With Cohen's kappa values from 0.31 to 0.65 the statistical significance is without question and exceptionally fair to good in the case of machine learning algorithm (Fleiss et al., 2003). The *recall* rates, displaying the proportion rate of *true positives* and *false negatives*, vary from 67-70% for the two full data examples and 80-87% for the spatially filtered tables. The *precision*, which rates the proportion of the flood topic itself in *true positives* and *false positives*, reveal values between 24-38% for the full data and 25-45% in the spatial extracted tables. The *specificity*, displaying the proportion of the predicted condition negative, ranges from 97% for both full datasets and 82-92% for the spatially filtered tables. The *NPV* indicates the proportion of all true negatives and false negatives. In our case, the rates vary between 99% for both full and 98% of the small samples. As mentioned before the best results were made between 6-8 topics for a one day period. Future works employing LDA should also concentrate on improving the hyperparameter estimation of *beta* and *alpha* as well as the number of topics *k*. Several attempts were made from Gibbs sampling to introducing hierarchical Bayesian models (Heinrich, 2004). There are ways to automatically determine the number of topics like hierarchical Dirichlet process (HDP) but there is still no standardized way. By and large it can be said, that for the small flood topic in the sparse German dataset the implementation of the synonym list helped a lot, in order to improve the topic estimation. Of course one could argue that it slightly falsifies the LDA-modelling, but the rewards of the results and the statistical evidence prove the benefits.

What additional benefits can be obtained from the intersection of 'Collective sensing' with the proven remote sensing data?

The statistical values that are given above lead to several interpretations. After the spatial filtering, and the thereby reduction of the Tweets around a sixth, the *recall* and *precision* rates improved, while the *specificity* value decreased and the *NVP* almost stayed the same. An explanation for the better *recall* and *precision* rates of the spatial extracted samples can be seen in the word frequencies. The concentration on the flooded areas and the estimated buffer of 30km around the remote sensing displayed, that the hazard related Tweets showed significantly higher numbers in the filtered sets than in the others. This result is not surprising and correlates with the geographic approach, for the water masks already outlined the spatial extent of an on the ground problem. With the reduction of the overall number of Tweets and the proportional increase of the on-topic messages, it is easier for the online- LDA algorithm to identify the latent flood topic and to assess the semantic similarities. As mentioned in **Chapter 6** recent studies proved, that flood-related Tweets were 11 times more likely to occur near flood-affected areas (<10 km) (Albuquerque et al., 2015). Regarding this, the loss of the Tweets, which did not intersect the 30km buffer around the water masks, can be rather neglected, because it is very unlikely, that they contain critical on-ground information on the disaster event.

To what degree can the results of the LDA algorithm be improved, if only the flood-related topic is calculated a second time? Keyword: Cascading LDA

From the figures of the confusion matrix, it is apparent that the recall rate almost stayed the same, while the precision value increased about 52%. With a *kappa* value of 0.65, the cascading LDA run had the highest statistical significance of all tested samples. Nonetheless based on the original on-topic Tweets of 3rd June, nearly half of the flood-relating Twitter messages were identified correctly compared to the 67.5% of the single LDA run. However cascading LDA has one major advantage – **time**. In disaster management the response phase is critical, in order to preserve the health and safety of the communities and their property. To gain as much information as possible in the shortest time can sometimes be more valuable than knowing all the facts in detail. Cascading LDA offers this solution, even if there is the chance to misclassify about 50% of the on-topic messages. **Figure 36** shows the cascading steps and the measured time for the LDA computation as well as the manual classification, which had to be performed, in order to evaluate the algorithm itself.

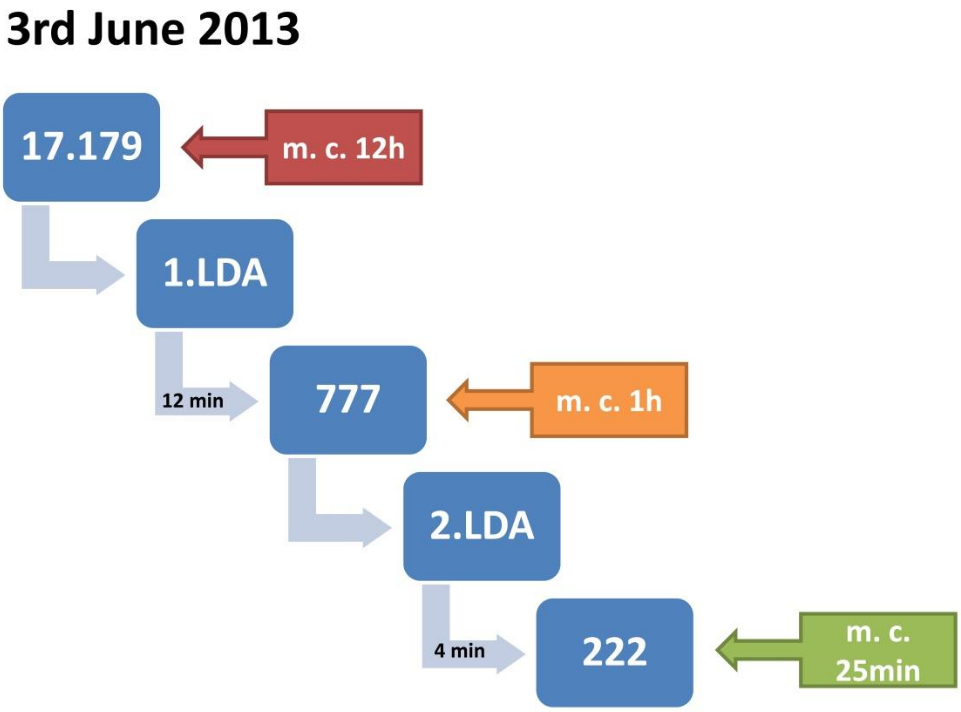


Figure 36. Cascading LDA - Comparison of the topic calculation time, with the manual classification in the PostgreSQL DB.

Calculating both LDA runs in the PostgreSQL DB took around 16 minutes and the manual classification in on- and off-topic messages, another 25 minutes, while calculation and manual evaluation of the first LDA run took around one hour and 12 minutes.[65] By and large we can see, that cascading LDA could speed up information gathering especially in very big datasets. While Germany is not the most Tweeting region in the world, countries like the USA with millions of single Tweets per day could be further promising fields of application. Of course there are additional steps are necessary in order to improve the computation time. While Python is one of the quickest object oriented languages there are LDA algorithms written in pure C, which could speed up the computation process. However the genism library, which was used by this thesis, counts as one of the most stable libraries in terms of probabilistic topic modelling.

[65] Time measurements based on the used Workstation: Intel I7 2600K 4.0 Ghz, 32GB-RAM

9 Conclusion and Future Works

This thesis exemplified a workflow reaching from automatically reading Twitter messages into a PostgreSQL database via script/application as far as classifying and filtering the Tweets in terms of their spatiotemporal significance, in order to highlight the benefits of VGI in the context of Disaster Management. This paper also serves as a window to an understanding of the process in the sense of 'Collective sensing' (Resch, 2013b) from harvesting Social Media data, fusing it with traditional remote sensing data and classifying the content with a sophisticated unsupervised topic model, in order to search for latent topics in unstructured data. However computing and classifying the Tweets with LDA is no one-way road. Once the model is calculated, it can be updated for every new flood, hitting the researched area, because it is in all probability that flood-related messages will not change their semantic content in the foreseeable future within specific spatial extents. We can see then, that a stable LDA-model can be used to classify Tweets, harvested by our application in near-real time. The arguments given above reveal that it is in all likelihood to get additional benefits on the topic of traditional Disaster Management. The combination of authoritative data with concepts of 'Collective sensing' in the case of Social Media provides an effective way to collect and gain critical information in a spatiotemporal context. While it cannot replace traditional keyword searches or time intense manual classification methods entirely, it suits as a powerful auxiliary tool to support relief organisations or public authorities. Future works could also concentrate on exploring the networks of the users and combining them with the spatial filtering through remote sensing data. Machine learning is developing rapidly as enterprises like Google and Facebook show. The potential to deliver real-time optimizations regarding different languages, network analysis, as well as different semantic contents is just starting to evolve and accelerating quickly. Therefore it is in all probability, that these revolutionising developments could solve many problems in the process of fusing Twitter corpora with remote sensing data, so as to lead to a new understanding of near real-time semantic analysis.

A References

Albuquerque, J.P. de, Herfort, B., Brenning, A., Zipf, A., 2015. A geographic approach for combining social media and authoritative data towards identifying useful information for disaster management. Int. J. Geogr. Inf. Sci. 29, 667–689. doi:10.1080/13658816.2014.996567

Baharin, S., Shibghatullah, A., Othman, Z., 2009. An Application Framework of Integrated Routing Application for Emergency Response Management System. Int. Conf. Soft Comput. Pattern Recognit. 716 – 719.

Baird, M., 2010. "The "Phases" of Emergency Management."

Bakillah, M., Li, R.-Y., Liang, S.H.L., 2015. Geo-located community detection in Twitter with enhanced fast-greedy optimization of modularity: the case study of typhoon Haiyan. Int. J. Geogr. Inf. Sci. 29, 258–279. doi:10.1080/13658816.2014.964247

Blei, D.M., 2012. Probabilistic topic models. Commun. ACM 55, 77. doi:10.1145/2133806.2133826

Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent dirichlet allocation. The Journal of Machine Learning Research 3, 993–1022.

Bortz, J., 1999. Statistik für Sozialwissenschaftler, 5., vollst. überarb. und aktualisierte Aufl. ed, Springer-Lehrbuch. Springer, Berlin.

Burton, S.H., Tanner, K.W., Giraud-Carrier, C.G., West, J.H., Barnes, M.D., 2012. "Right Time, Right Place" Health Communication on Twitter: Value and Accuracy of Location Information. J. Med. Internet Res. 14, e156. doi:10.2196/jmir.2121

Cohen, J., 1960. A Coefficient of Agreement for Nominal Scales. Educ. Psychol. Meas. 20, 37–46. doi:10.1177/001316446002000104

Cova, T., 1999. GIS in emergency management. Geogr. Inf. Syst. 2, 845–858.

Crooks, A., Croitoru, A., Stefanidis, A., Radzikowski, J., 2013. #Earthquake: Twitter as a Distributed Sensor System: #Earthquake: Twitter as a Distributed Sensor System. Trans. GIS 17, 124–147. doi:10.1111/j.1467-9671.2012.01359.x

Culotta, A., 2010. Towards detecting influenza epidemics by analyzing Twitter messages. ACM Press, pp. 115–122. doi:10.1145/1964858.1964874

De Longueville, B., Smith, R.S., Luraschi, G., 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. ACM Press, p. 73. doi:10.1145/1629890.1629907

Fleiss, J.L., Levin, B., Paik, M.C., 2003. Statistical methods for rates and proportions, 3rd ed. ed, Wiley series in probability and statistics. J. Wiley, Hoboken, N.J.

Fuchs, G., Andrienko, N., Andrienko, G., Bothe, S., Stange, H., 2013. Tracing the German centennial flood in the stream of tweets: first lessons learned. ACM Press, pp. 31–38. doi:10.1145/2534732.2534741

Gesualdo, F., Stilo, G., Agricola, E., Gonfiantini, M.V., Pandolfi, E., Velardi, P., Tozzi, A.E., 2013. Influenza-Like Illness Surveillance on Twitter through Automated Learning of Naïve Language. PLoS ONE 8, e82489. doi:10.1371/journal.pone.0082489

Godschalk, D., 1991. Disaster mitigation and hazard management. Emergency management: Principles and practice for local government.

Goodchild, M.F., 2007. Citizens as sensors: the world of volunteered geography. GeoJournal 69, 211–221.

Goodchild, M.F., Glennon, J.A., 2010. Crowdsourcing geographic information for disaster response: a research frontier. Int. J. Digit. Earth 3, 231–241.

Goodchild, M.F., Li, L., 2012. Assuring the quality of volunteered geographic information. Spat. Stat. 1, 110–120. doi:10.1016/j.spasta.2012.03.002

Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P., Ratti, C., 2014. Geo-located Twitter as proxy for global mobility patterns. Cartogr. Geogr. Inf. Sci. 41, 260–271.

Heinrich, G., 2004. Parameter estimation for text analysis.

Hoffman, M.D., Blei, D.M., Bach, F., 2010. Online Learning for Latent Dirichlet Allocation, in: Proceedings of the 23rd International Conference on Neural Information Processing Systems, NIPS'10. Curran Associates Inc., USA, pp. 856–864.

Imran, M., Elbassuoni, S., Castillo, C., Diaz, F., Meier, P., 2013. Practical extraction of disaster-relevant information from social media. ACM Press, pp. 1021–1024. doi:10.1145/2487788.2488109

Jovilyn, T., Fajardo, B., Oppus, M., 2010. A mobile disaster management system using the android technology. WTOC 9, 343–353.

Khan, H., VASILESCU, L., KHAN, A., 2008. DISASTER MANAGEMENT CYCLE – A THEORETICAL APPROACH. Management & Marketing - Craiova 1, 43–50.

Kongthon, A., Haruechaiyasak, C., Pailai, J., Kongyoung, S., 2014. The Role of Social Media During a Natural Disaster: A Case Study of the 2011 Thai Flood. Int. J. Innov. Technol. Manag. 11, 1440012. doi:10.1142/S0219877014400124

Kryvasheyeu, Y., Chen, H., Moro, E., Van Hentenryck, P., Cebrian, M., 2015. Performance of Social Network Sensors during Hurricane Sandy. PLOS ONE 10, e0117288. doi:10.1371/journal.pone.0117288

Landis, J.R., Koch, G.G., 1977. The Measurement of Observer Agreement for Categorical Data. Biometrics 33, 159. doi:10.2307/2529310

Landwehr, P.M., Carley, K.M., 2014. Social Media in Disaster Relief, in: Chu, W.W. (Ed.), Data Mining and Knowledge Discovery for Big Data. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 225–257.

MacEachren, A.M., Robinson, A.C., Jaiswal, A., Pezanowski, S., Savelyev, A., Blanford, J., Mitra, P., 2011. Geo-twitter analytics: Applications in crisis management.

Miller, H.J., Goodchild, M.F., 2015. Data-driven geography. GeoJournal 80, 449–461. doi:10.1007/s10708-014-9602-6

Mooney, P., Corcoran, P., Winstanley, A.C., 2010. Towards quality metrics for OpenStreetMap. ACM Press, p. 514. doi:10.1145/1869790.1869875

Neal, D.M., 1997. Reconsidering the Phases of Disaster. Int. J. Mass Emergencies Disasters 15, 239–246.

Neis, P., Zielstra, D., Zipf, A., 2011. The Street Network Evolution of Crowdsourced Maps: OpenStreetMap in Germany 2007–2011. Future Internet 4, 1–21. doi:10.3390/fi4010001

Peters, R., Albuquerque, J.P. de, 2015. Investigating images as indicators for relevant social media messages in disaster management. Proceedings of the ISCRAM 2015 Conference - Kristiansand, May 24 - 27.

Porter, M.F., 1997. Readings in Information Retrieval, in: Sparck Jones, K., Willett, P. (Eds.), . Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 313–316.

Pritchard, J.K., Stephens, M., Donnelly, P., 2000. Inference of Population Structure Using Multilocus Genotype Data. Genetics 155, 945–959.

Raymond, E.S., 1999. The cathedral & the bazaar: musings on Linux and open source by an accidental revolutionary, 1st ed. ed. O'Reilly, Beijing ; Cambridge, Mass.

Resch, B., 2013a. People as sensors and collective sensing-contextual observations complementing geo-sensor network measurements, in: Progress in Location-Based Services. Springer, pp. 391–406.

Resch, B., 2013b. People as Sensors and Collective Sensing-Contextual Observations Complementing Geo-Sensor Network Measurements, in: Krisp, J.M. (Ed.), Progress in Location-Based Services. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 391–406.

Resch, B., Mittlboeck, M., Lippautz, M., 2010. Pervasive Monitoring—An Intelligent Sensor Pod Approach for Standardised Measurement Infrastructures. Sensors 10, 11440–11467. doi:10.3390/s101211440

Sagl, G., Resch, B., Mittlboeck, M., Hochwimmer, B., Lippautz, M., Roth, C., 2012. Standardised geo-sensor webs and web-based geo-processing for near real-time situational awareness in emergency management. Int. J. Bus. Contin. Risk Manag. 3, 339–358.

Sakaki, T., Okazaki, M., Matsuo, Y., 2010. Earthquake shakes Twitter users: real-time event detection by social sensors. ACM Press, p. 851. doi:10.1145/1772690.1772777

Schröter, K., Kunz, M., Elmer, F., Mühr, B., Merz, B., 2015. What made the June 2013 flood in Germany an exceptional event? A hydro-meteorological evaluation. Hydrol. Earth Syst. Sci. 19, 309–327. doi:10.5194/hess-19-309-2015

Sivic, J., Russell, B.C., Zisserman, A., Freeman, W.T., Efros, A.A., 2008. Unsupervised discovery of visual object class hierarchies. IEEE, pp. 1–8. doi:10.1109/CVPR.2008.4587622

Stefanidis, A., Crooks, A., Radzikowski, J., 2013. Harvesting ambient geospatial information from social media feeds. GeoJournal 78, 319–338. doi:10.1007/s10708-011-9438-2

Stehman, S.V., 1997. Selecting and interpreting measures of thematic classification accuracy. Remote Sens. Environ. 62, 77–89. doi:10.1016/S0034-4257(97)00083-7

Steiger, E., Ellersiek, T., Resch, B., Zipf, A., 2015a. Uncovering Latent Mobility Patterns from Twitter During Mass Events. GI_Forum 1, 525–534. doi:10.1553/giscience2015s525

Steiger, E., Resch, B., Zipf, A., 2015b. Exploration of spatiotemporal and semantic clusters of Twitter data using unsupervised neural networks. Int. J. Geogr. Inf. Sci. 1–23. doi:10.1080/13658816.2015.1099658

Stein, C., Malitz, G., 2013. Das Hochwasser an Elbe und Donau im Juni 2013: Wetterentwicklung und Warnmanagment des DWD ; Hydrometeorologische Rahmenbedingungen, Berichte des Deutschen Wetterdienstes. Selbstverl. des Deutschen Wetterdienstes, Offenbach am Main.

Thieken, A., Deutsches Komitee für Katastrophenvorsorge (Eds.), 2015. Das Hochwasser im Juni 2013: Bewährungsprobe für das Hochwasserrisikomanagement in Deutschland, Schriftenreihe des DKKV. DKKV, Bonn.

Tobler, W.R., 1970. A Computer Movie Simulating Urban Growth in the Detroit Region. Econ. Geogr. 46, 234. doi:10.2307/143141

Zandbergen, P.A., Barbeau, S.J., 2011. Positional Accuracy of Assisted GPS Data from High-Sensitivity GPS-enabled Mobile Phones. J. Navig. 64, 381–399. doi:10.1017/S0373463311000051

Zhang, X., Fuehres, H., Gloor, P.A., 2011. Predicting Stock Market Indicators Through Twitter “I hope it is not as bad as I fear.” Procedia - Soc. Behav. Sci. 26, 55–62. doi:10.1016/j.sbspro.2011.10.562

B List of Figures