Geo-information Science and Remote Sensing

Thesis Report GIRS-2018-45

Retrieving critical location based information from Twitter for disaster management using LDA topic modelling

A case study with a chemical fire at Moerdijk

R.B. Nedkov

02-09-2018

WAGENINGEN
UNIVERSITY & RESEARCH

Retrieving critical location based information from Twitter for disaster management using LDA topic modelling

A case study with a chemical fire in Moerdijk

Rony Nedkov
900415597120

Supervisors:

Dr. ir. Ron van Lammeren
Dr. ir. Arend van Ligtenberg

A thesis submitted in partial fulfilment of the degree of Master of Science
at Wageningen University and Research Centre,
The Netherlands.

02-09-2018
Wageningen, The Netherlands

Thesis code number: GRS-804xx
Thesis Report: GIRS-2015 -xx
Wageningen University and Research Centre
Laboratory of Geo-Information Science and Remote Sensing

Abstract

Twitter has been recognized as a valuable information source for disaster management, due to the content in combination with the spatial component. Raw Twitter data has an unstructured nature, which consist of both relevant and irrelevant data, therefore classification and information retrieval techniques are needed to make it useful for the emergency services.

There is a lack of integration between information retrieval methods and spatial temporal analysis, which makes the practical application limited. Therefore the objective was: what critical location based information can be derived from Twitter during the response phase of a disaster to support the decision making process in disaster management.

The chemical fire at Chemi-pack was chosen as a case study to demonstrate the effectiveness of Latent Dirichlet Allocation (LDA) Topic Modelling. Self-Organizing Maps (SOM) and word clouds were used to analyze and interpret the results more in depth. The dataset consists of Tweets which have been retrieved from Twitter several days after the incident. Before topic discovery was done, additional location information was retrieved from Twitter to geocode Tweets without a location and enlarge the spatial dataset.

The topic modelling several topics which could be matched with events in reality. One of these topics was considered of interest for emergency services. This topic was related to the toxic plume. We found that half of the Tweets from this topic contained actionable information and that one third of the Tweets was correctly geocoded. Thereafter it was possible to spatially visualize the event on the granularity level of residential areas.

In conclusion, with additional location information provided by Twitter it is possible to significantly increase the size of the dataset with an acceptable accuracy. LDA topic modelling has a limited performance when used on Twitter data, but the use of SOM and word clouds makes these results easier to analyze and interpret.

The various steps in this research and the limited performance of LDA topic modelling resulted in insufficient Tweet classification, therefore a majority of the Tweets remained unused.

We recommend to extend the contextual information in Tweets by means of tweet pooling. Additionally tokenization and word stemming can be applied to improve the performance of topic modelling.

Keywords: Twitter, LDA, Topic Modeling, Disaster Management, Self-Organizing Maps, Actionable Information

Foreword

Before you is my master thesis for the

Table of Contents

List of abbreviations

API	Application Programming Interface
CSV	Comma-Separated Value
GPS	Global Positioning System
HTTP	Hypertext Transfer Protocol
JSON	JavaScript Object Notation
LDA	Latent Dirichlet Allocation
OGC	Open Geospatial Consortium
QGIS	Quantum Geographic Information System
SOM	Self-Organizing Maps
URL	Uniform Resource Locator
WKT	Well-Known Text

1. Introduction

Disasters are events that have a an disruptive effect on the community and a high impact on the surroundings (Cutter, 2003). Emergency services are responsible for the management of these events and the main goal is to prepare the community for a disaster and respond when one strikes to limit hazards for the people and property. Disasters are managed according the disaster management cycle which consists of four phases as shown in Figure 1: mitigation, preparedness, response and recovery. (Horita, Degrossi, Assis, Zipf, & De Albuquerque, 2013). Each phase includes specific activities which are either related to preparing the community for a disaster, responding to it or recovering from a disaster. The response takes place in the immediate aftermath of a disaster and is considered the most critical and complex phase. The main purpose of this phase is preventing further escalation of the emergency to limit the threats and disruptive impact on the community. This involves putting procedures into practice, mobilizing resources and prioritization of actions according to the threats.

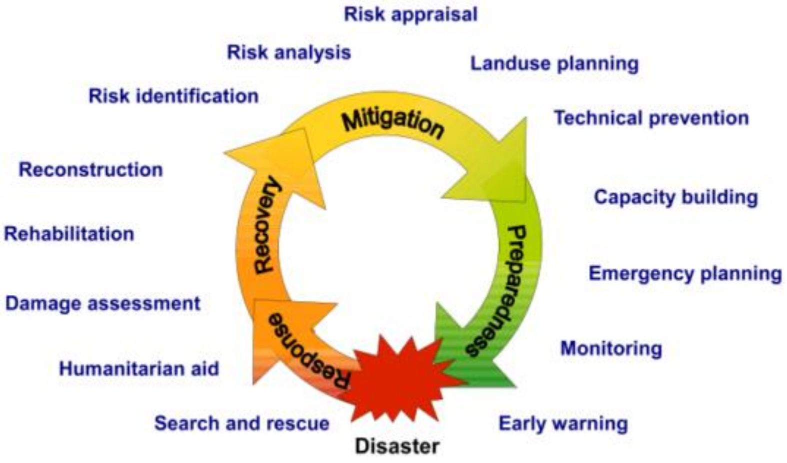


Figure 1: Disaster management cycle (Horita et al., 2013)

A critical factor for effective disaster management is obtaining situational awareness and finding the needs of the public (J. R. Harrald, 2006). One of the first definitions of situational awareness is "one's ability to remain aware of everything that is happening at the same time and to integrate that sense of awareness into what one is doing at the moment" (Haines & Flatau, 1992). Using this definition in the field of disaster management, situational awareness means creating an understanding on the severity, location and urgency of events during incidents. This requires timely, accurate and complete spatial and temporal information: what is happening where and when (Horita et al., 2013). The combination of situational awareness and accurate information enables well informed decision making on the allocation of resources and prioritization of actions. On the contrary, inadequate information and lack of awareness has the potential to hinder operations with an escalation of the crisis, loss of life and further damage to property as a result (Ostermann & Spinsanti, 2011).

Traditional information sources for obtaining situational awareness are standardized operational procedures, pre-plans and maps which are defined during the preparedness phase. These plans include information such as the vulnerabilities, hazards, resources and demographics of the area which enable first responders to create an initial situation awareness which is to be used on understanding the environment and possibly affected area with its hazards. However, these plans consist of static data that may be outdated and which does not depict the current state of the hazards, resources and vulnerabilities (Vivacqua & Borges, 2012). Effective decision making on the scene requires information on the current state, possible development of the disaster and the impact on the community. This information is often not readily available. Therefore the first priority of emergency services is to collect additional information upon arrival on the scene. This is a process during which first responders heavily rely on their experience, knowledge, training, on-site observations and cooperation with other organizations (Zhang, Shen, Zhang, Xie, & Yang, 2015). The collection of additional information is also an ongoing process as to maintain an accurate situational awareness. The actions taken to mitigate the hazards alternate the situation and the environment the first responders work in.

In recent years social media services such as the micro-blogging platform Twitter and multimedia sharing service Flickr have emerged as valuable information sources for many different applications (Stefanidis, Crooks, & Radzikowski, 2013). Typically these services have the same goal: to provide the public with means to create content, share it and interact with other users on the Internet (Kaplan & Haenlein, 2010). The popularity of these services has increased and as a result a vast amount of content is created every single day. This content contains views, opinions and experiences related to the direct surroundings of the users (Sakaki, Okazaki, & Matsuo, 2010). Parallel to this development more and more content is becoming tagged with a location as the number of location-aware mobile devices with access to the Internet has grown as well. Moreover, the data is freely accessible at any point.

In the field of disaster management, social media content has been recognized as a potential information source to obtain and maintain a situational awareness during the response phase of a disaster (Horita et al., 2013). People turn to Twitter to inform themselves about an ongoing disaster or incident, to report about their situation and discuss topics related to rapidly evolving events (Teevan, Ramage, & Morris, 2011). The content in combination with the spatial component is considered to be valuable information for emergency services to create and maintain a situational awareness. Based on the online activity it is possible to get real-time, accurate and precise information on what is happening where. Emergency services can use this content to increase their understanding of the urgency and impact of the incident and locate the people in need of help.

1.1 Problem definition

One of the main challenges of using social media as a spatial information source for disaster management is the unstructured nature of the data (Vivacqua & Borges, 2012). There is a lack of pre-defined rules on how the content and spatial information is organized. This results in datasets containing both relevant and irrelevant data. Presenting this raw data to the emergency services without any processing can lead to an information overload (Kiatpanont, Tanlamai, & Chongstitvatana, 2017). This occurs when first responders are confronted with more information than they can process in the limited time available during a disaster. The information is therefore either not used or interpreted incorrect. To make social media data readily available, classification and information retrieval techniques are required.

In the research field of social media content and disaster management, studies mostly focus either on the spatial analysis of the spatial component or on the information that can be retrieved from the content. Currently there is a lack of integration between information retrieval methods and spatial temporal analysis that utilize the wide spectrum of valuable information and meta-data of social media content (Steiger, de Albuquerque, & Zipf, 2015). As a result many studies show the potential value of social media content for aiding disaster management (Horita et al., 2013), but the practical application for the emergency services remains limited. Studies that primary focus on the spatial component lack semantics and contextual information. Moreover, a common practice is to only take into account geo-tagged content to maintain the spatial accuracy. It can be stated that a large amount of valuable content is discarded with this method considering only one to three percent of the content is geo-tagged (Tsou et al., 2017). Typically these studies show *where* something is happening, but cannot define *what* is happening on that location. On the contrary, studies that have a focus on information retrieval are able to retrieve *what* is happening, but cannot specify *where* this happens.

This study aims to overcome the above described information and spatial gap. The overall goal is to utilize available spatial information from Twitter and integrate this with information retrieval techniques to meet the actionable information needs of the emergency services during the response phase.

1.2 Objective and research questions

Regarding the overall goal, the objective of this research was to study what critical location based information can be derived from Twitter during the response phase of a disaster to support the decision making process in disaster management. The research questions are as follows:

1. To what extent can the spatial information of the dataset be increased by using location information from the Twitter meta-data?
2. Which events can be derived from the available Tweets using Latent Dirichlet Allocation topic modelling?
3. How well do the identified events and locations resemble the actual course of events?
4. What actionable information can be derived from the events for the emergency services?

1.3 The case study and study area

To answer the research questions and objective we will use a case study. On fifth of January 2011 a chemical packaging and storage facility Chemi-Pack caught fire, causing a huge disaster. A thick black plume arose from the fire and was taken by the winds towards dense populated areas around Dordrecht and Rotterdam. As a result the incident was quickly scaled up involving three safety regions, namely Midden- en West-Brabant, Rotterdam-Rijnmond and Zuid-Holland-Zuid. The extent of these regions is shown in Figure 2. The primary concern of the emergency services was the toxic substances and particles in the plume. These could cover the area and cause a serious health hazard for the dense populated area. People were warned to stay inside and close windows and doors by activating the sirens. This caused public distress which was further increased by the conflicting information provided by governmental agencies related to actual danger. The fire was heavily discussed on social media and this resulted in a large dataset which can be used for this research.

The emergency services struggled with this incident for several reasons. First, in the immediate aftermath it was unclear what the extent of the disaster was in terms of affected people. Second, there was a mismatch between the information that was provided by the governmental agencies and the information need of the public. The public asked for more information than the agencies provided. As a result the public distress increased. Third, the public concern on social media and lack of information resulted in the spread of false information on Twitter. The public distress and misinformation was further increased. Even the governmental agencies and emergency services were not trusted anymore (De Onderzoeksraad voor Veiligheid, 2011).

The severity of this disaster, the events (such as the public distress) and the activity on Twitter make this event an interesting case study for extraction actionable information for emergency services.

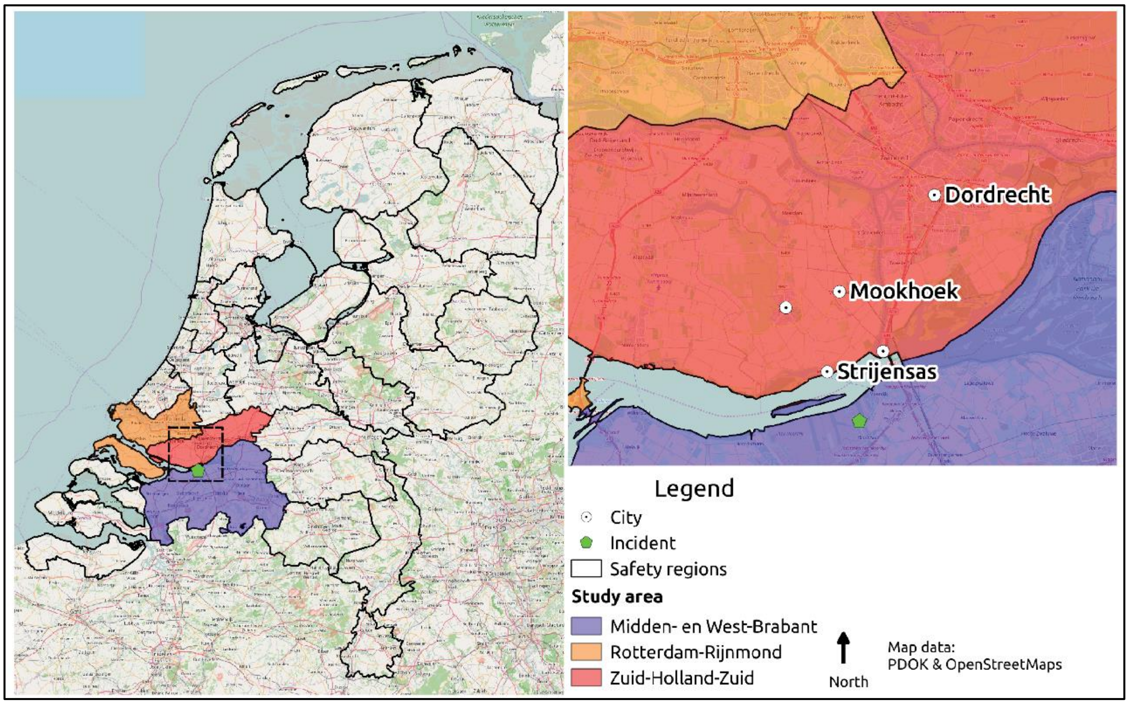


Figure 2: Overview of the study area.

Due to the limit time available for this research, we will focus on Tweets and topics related to the toxic plume from research question two onwards. This topic is of interest for the emergency services as it was not clear what effect the plume had on the people in the surrounding area (De Onderzoeksraad voor Veiligheid, 2011).

1.4 The dataset

The available dataset for this study consist of Comma-Separated Value (CSV) file containing 117,183 Tweets which have been retrieved from Twitter one week after the incident. The dataset covers the entire incident from when the fire was first reported at 14:27 on January fifth until January eight. The Tweets have been retrieved from Twitter using the Application Program Interface (API). The API allows programmatic reading and writing access to the database of Twitter. Ranter (2011) has created the dataset by querying the API for Tweets that contain at least one of the following keywords:

- Moerdijk
- Moerdijkbrand
- Brandmoerdijk
- Sirenes
- Luchtalarm
- Ramen en deuren
- Chemie-pack
- Blusdeken
- Rampenzender
- Gifwolk

2. Review

2.1 Disaster management and information needs

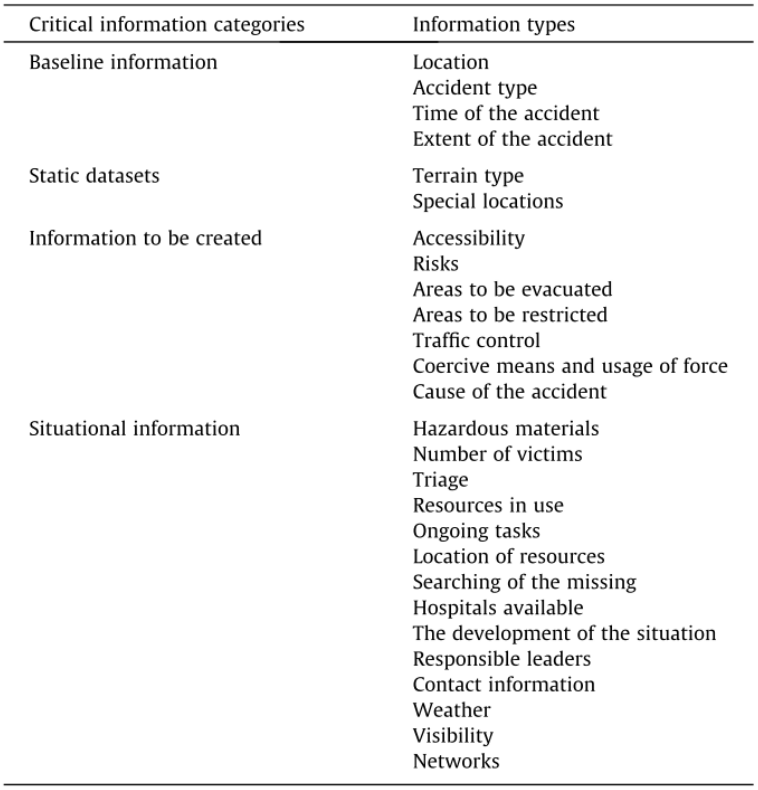
In each phase of the disaster management cycle information is needed related to the hazards, risks, potential effects on the surroundings and the available resources. The information needs and the requirements for this information vary for each phase. Therefore the characteristics of each phase should be considered when identifying the information needs of the emergency services

The quality of the information is one of the most important aspects of the information needs (Harrald & Jefferson, 2007). Emergency services heavily rely on complete and timeliness information for effective decision-making and the allocation of limited resources. The completeness aspect of the quality is described as “the level of similarity that exists between the data produced and the ’perfect’ data that should have been produced (that is, data produced without error)” (Devillers & Jeansoulin, 2010). Timeliness refers to whether the information is available when it is needed. For example, incoming information describing the presence of heavily flammable chemicals on a site is of no use when these chemicals are already burning. If this information came in earlier, than actions could have been taken to prevent the chemicals from catching fire in the first place.

The requirements for the quality should be considered in relation to the context the information is needed (Seppänen & Virrantaus, 2015). This context is described by the time critically of the activities and the desired outcomes of each phase (Vivacqua & Borges, 2012). The main activities in the mitigation and preparedness phases are focused on preparing a community for a possible disaster. Hazards and risks are identified; measures are taken to limit possible effects; plans and procedures are defined on how to act if a disaster occurs; people are trained and resources are allocated. The time critically is low as these phases can last several years. The quality requirements are high due to the available time for data collection, processing and analyzing. Moreover, a well prepared community will be able to act more effectively when a disasters strikes. The response phase is the most time critical phase as emergency services need to take control of the situation. First responders are initially overwhelmed as there is no situational awareness and the severity and extent of the disaster. They are challenged by complex conditions such as the uniqueness of the event; time pressure; uncertainty of the situation due to constant changes; lack of complete information and a large number of people under threat. Gaining control over the situation requires quick decision making. The time for processing, analyzing and validating the available information is very limited. Relying on possibly incorrect data in this context often outweighs the risk of doing nothing (Tapia, Moore, & Johnson, 2013). As control is gained over the situation and the hazards are taken away, the time critically decreases as the situation moves to the mitigation phase. People are more self-reliant and the situation is taken over by humanitarian organizations.

The information needs can be further subdivided in information categories. Seppänen, Mäkelä, Luokkala, & Virrantaus (2013) have defined four categories for critical information needs as shown in Table 1.

Table 1: Categories of information need categories.



Critical information categories	Information types
Baseline information	Location
	Accident type
	Time of the accident
	Extent of the accident
Static datasets	Terrain type
	Special locations
Information to be created	Accessibility
	Risks
	Areas to be evacuated
	Areas to be restricted
	Traffic control
	Coercive means and usage of force
	Cause of the accident
Situational information	Hazardous materials
	Number of victims
	Triage
	Resources in use
	Ongoing tasks
	Location of resources
	Searching of the missing
	Hospitals available
	The development of the situation
	Responsible leaders
	Contact information
	Weather
	Visibility
	Networks

These categories can be further divided into information that is available before an emergency occurs, and information that is not (Seppänen & Virrantaus, 2015). The mitigation and preparation phase require as much as information possible for the tasks, but will mostly rely on information from the static datasets. Some information types from the information to be created and situational information categories can also be made available during these phases. For example, daily activities of emergency services when not responding is creating accessibility maps of industrial areas; creating spatial datasets with hazardous materials; identification of risks and hazards etc.. During the response phase the information to be created and the situational information are the most important information needs as these are needed for the creation of situational awareness.

2.2 Twitter as a source of information for disaster management

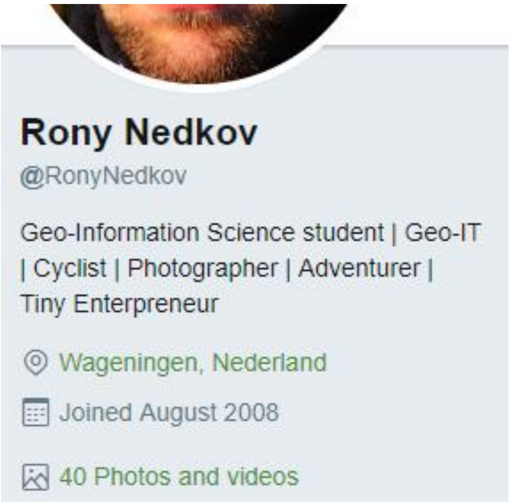
Twitter is a popular real-time micro-blogging social platform which allows users to post and read short messages - called Tweets - of 140 characters including media such as photos and videos. The content of the Tweets contain views, opinions and experiences (Bruns & Burgess, 2011) (Sakaki et al., 2010). The content is publicly available meaning that anyone with a connection to the Internet can read the Tweets at any point in time. The metadata of the Tweets often also contain a geographical component which can relate the content of the message to a specific geographical location.

Twitter is of particular interest as an information source for disaster management due to its nature of a real-time information dissemination platform. Various studies have shown that people use Twitter to communicate and keep each other informed during disasters and large scale incidents (Herfort, de Albuquerque, Schelhorn, & Zipf, 2014; Longueville & Smith, 2009; Starbird, Palen, Hughes, & Vieweg, 2010; Vieweg, Palen, Liu, Hughes, & Sutton, 2008). More specifically, Acar & Muraki (2011) found that individuals within the direct surroundings of an incident are more likely to Tweet about the unsafe situation, whereas people living further away are Tweeting about possible secondary effects such as inaccessible roads etc. Even the news agency use Twitter as an information source by embedding Tweets directly in Live Blogs during breaking news events (Thurman, 2014). These findings illustrate that Twitter potentially contains valuable information in the event of a disaster which can be used by the emergency services to create a situational awareness.

Another important characteristic of Twitter data is the spatial component of the Tweets. Users have three options for to associate a Tweet with a geographic location, namely: geographic coordinates, profile locations and predefined places.

First, the geographic coordinates are the most accurate method for associating a Tweet with a location. Users can assign geographic coordinates to a Tweet using the Global Position System (GPS) of the device which is being used for posting the Tweet. The GPS receiver has an accuracy of several meters if the signal is not obstructed by high objects such as trees or buildings. However, Burton, Tanner, Giraud-Carrier, West, & Barnes (2012) have shown that on average only 2,07% of the Tweets worldwide is geo coded using this method.

Second, profile locations are free text locations assigned by a user to his or her profile that indicate the approximate whereabouts of the user as shown in Figure 3. The profile locations are the biggest source of location information as a majority of the users has set a profile location (Twitter, 2017). The disadvantage of this method is that the user may not be at the location as specified in the profile when posting Tweets. The location may even be false or non-existing (Takhteyev, Gruzd, & Wellman, 2012). However, Burton et al. (2012) have found that profile locations match the geographic coordinates of a Tweet (if provided by the user) in 87% of the cases.



Rony Nedkov

@RonyNedkov

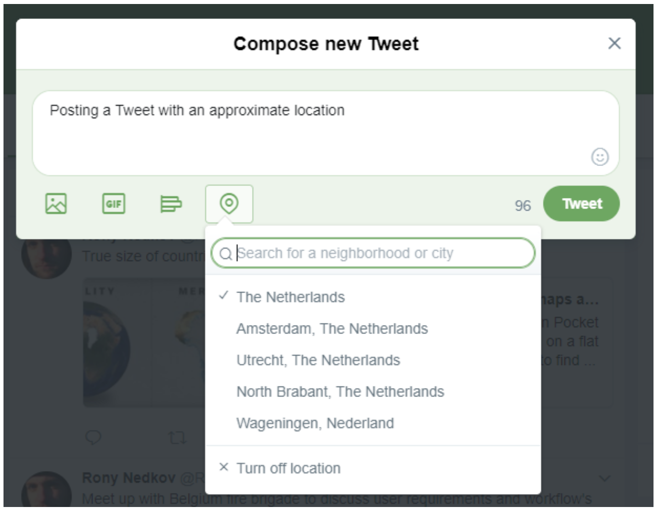
Geo-Information Science student | Geo-IT | Cyclist | Photographer | Adventurer | Tiny Enterpreneur

Wageningen, Nederland

Joined August 2008

40 Photos and videos

Figure 3: Example of a location in the user profile.



Compose new Tweet

Posting a Tweet with an approximate location

Search for a neighborhood or city

- ✓ The Netherlands
- Amsterdam, The Netherlands
- Utrecht, The Netherlands
- North Brabant, The Netherlands
- Wageningen, Nederland
- × Turn off location

Figure 4: Example of selecting a place name to geo reference a Tweet.

The third option users have to associate a Tweet with a location, is by means of predefined places. These places have geographic coordinates and are provided by Twitter. The user can choose to add a place from a list as shown in Figure 4. A disadvantage of this method is that the user can choose a location which doesn't match the actual location. This happens when people Tweet *about* a location rather than the location they are *at*.

The real-time activity of the users during incidents in combination with the spatial component enables spatio-temporal analysis of rapidly evolving events such as disasters. However, the use of social media information has several challenges. Users creating content do not purposely create geographic information as done by contributors of Volunteered Geographic Information (Goodchild, 2007). As a consequence, social media information is unstructured, not standardized and thematically divers and not readily available and useable (Croitoru, Crooks, Radzikowski, & Stefanidis, 2013). Retrieval of meaningful information for the purpose at hand, requires additional processing and analyzing. Also, there is a risk of false information and rumors being spread which require additional validation of the analyzes results (Mendoza, Poblete, & Castillo, 2010). The lack of geographic coordinates of

Tweets may lead to inaccuracy in spatial analyses due to the uncertainty of the actual locations (Burton et al., 2012).

However, many social media content contains additional meta-data which associate the content with a location such as mentioned of place names in the content. These can be used to utilize a large part of the data. A disadvantage of these place names however, is that the spatial accuracy can decrease (Burton et al., 2012). However, in the field of hazard management the risk relying on possible incorrect data outweighs the risk of doing nothing (Tapia et al., 2013). In other words, the error margin in time critical decision making is bigger.

2.3 Discovering topics in micro-blogging data

Identifying topics in micro-blogging texts requires the use of text mining techniques. Text mining is the process of extracting useful information from both structured and unstructured data (Vijayarani, Ilamathi, & Nithya, 2015). The literature describes many methods ranging from simple manual classification to more advanced text mining techniques using super- or unsupervised machine learning algorithms.

Manual classification is done by human annotators which read every text and classify it according to its relevance to the topic of interest. This method has been widely applied in various studies (Hahmann, Purves, & Burghardt, 2014), (Starbird et al., 2010), (Imran, Elbassuoni, Castillo, Diaz, & Meier, 2013) (Kiatpanont et al., 2017). The advantage of this method is that it is the most accurate classification method compared to methods that rely on computer interpretation of texts (Hahmann et al., 2014). A major disadvantage is the need for human interpreters to read, review and classify every single text. This is a very time consuming process. Moreover, technical information experts may be needed which have sufficient expertise to correctly classify the text. For example, in the field of disaster management information experts are required that are familiar with the information needs of the emergency services in relation the time criticality of these needs.

Another common method for classifying Tweets is keyword filtering. In keyword filtering, a Tweet is considered relevant when one or more manually selected keywords are present in the content of a Tweet. Squicciarini, Tapia, & Stehle (2017) have used keyword filtering to retrieve Tweets related to the Hurricane Sandy for sentiment analysis. Longueville & Smith (2009) have shown that with this method an initial and general understanding of the chronological order of the events can be created. However, the method can also introduce noise since the selected keywords can be present in texts that are not related to the topic of interest.

Methods that require less human interpretation are supervised machine learning algorithms. These algorithms rely on a training dataset which is used to train a classifier to classify new texts. The training dataset requires manual annotation by human interpreters if no dataset is available. Commonly used supervised algorithms are probabilistic algorithms such as the Naïve Bayes Classifier, Support Vector Machines and K-Nearest Neighbor (Bilski, 2011). The accuracy of the classifiers depends on the input and the training datasets. Hahmann et al. (2014) have for example shown that these classifiers have a lower accuracy when used on micro-blogging texts. The reasoning behind this conclusion is that contextual information is required for differentiating content and finding patterns in texts. Microblogging texts contain little contextual information as the messages are very short, a maximum of 140 characters. Increasing the contextual information can be done by increasing the size of the training dataset. However, the available data is often limited and it can therefore be impossible to create a larger training dataset.

To overcome the need for human interpreters, unsupervised machine learning algorithms can be applied. These algorithms do not rely on a training dataset and are able to train themselves using the entire input dataset. A popular unsupervised machine learning method is Latent Dirichlet Allocation (LDA) topic modelling algorithm (Blei et al., 2003). This algorithm can automatically discover latent (i.e. hidden) structures between words from text-documents and define topics. The method assumes every text document is a mixture of all topic, and each topic is a mixture of all words. Each topic is therefore represented by all words from the dataset, ordered by their co-occurrence probabilities. These probabilities are calculated by an iterative process in which the text documents are regenerated using word probabilities. The generated documents are compared to the input documents

2.4 The chosen methods and definitions

In this study we are interested in retrieving valuable information for emergency services when it is needed most: the response phase. During this phase information is the most critical resource for creating a situational awareness and decision making. Currently emergency services heavily rely on their experience, knowledge of the area and observations when arriving on the disaster scene to get an initial understanding of the scale and impact of the incident. (Li, Yang, Ghahramani, Becerik-Gerber, & Soibelman, 2014). The consequences of a disaster can geographically spread far beyond the source. Emergency services are unable to comprehend the severity of the situation in the surrounding areas due to lack of local information and their absence to collect it. In these areas the local public can be used as an information source. These people can potentially provide valuable and accurate information on the current conditions in their direct surroundings. In paragraph 2.2 we introduced Twitter as a real-time information dissemination platform which is widely used by the public in affected areas of disasters to report on their surroundings. Therefore in this study we will use Twitter as an additional information source.

The majority of the Tweets does not have geographic coordinates. We will therefore utilize the place names and the profile locations to increase the number of Tweets with a location. We are aware of the questionable accuracy of the profile locations, as users may Tweet from different locations than specified in their profile. However, during the response phase first responders require more aggregated information for quick decision-making. In this study we will explore to what extent the profile locations and place names meet the information requirements of the emergency services.

We consider the contextual information in Tweets to be unstructured. Presenting this data to the emergency services without processing and filtering relevant from irrelevant information, can lead to information overload. To prevent this from happening, the Twitter data needs to be processed using text-mining methods. Manual filtering as mentioned in chapter 2.3 is very time consuming as each Tweet has to manually be classified. With the keyword filtering method it is possible to filter data, but not structure it. Therefore we have chosen to implement a machine learning algorithm to structure the raw Twitter data. We will use the (unsupervised) LDA topic modelling algorithm since the supervised algorithms require a training dataset which is not available during an incident.

The interpretation difficulty of the LDA topic modelling output can be simplified by using visualization techniques as suggested by Chaney & Blei (2012). We will use Self Organizing Maps (SOM) to create heat maps of Tweets with similar probabilities for every topic. SOM is an unsupervised algorithm that brings multidimensional data back to a two dimensional map. During this process it also groups the input samples and colors these according to the similarity of their properties. By doing so, heat maps are created which show clusters of samples that are more related than others. This method will allow us to quickly and visually explore the results of the topic modelling algorithm.

3. Methodology

In this chapter we will present the methodology that has been used to answer the objective and the research questions. **Error! Reference source not found.** shows a general overview of the methodology including. In section 3.1 we will explore the dataset to gain an understanding of how the data is spatially and temporally distributed. In section 3.2 we have pre-processed the data and import it in a database for the analysis

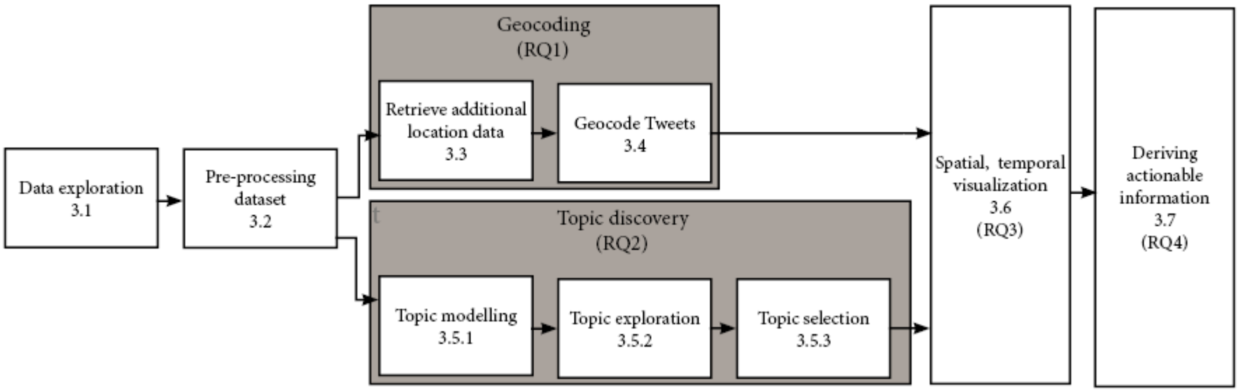


Figure 5: Overview of the research methodology and the relation to the research questions (RQ).

Sections 3.3 and 3.4 will focus on the retrieval of additional location information from Twitter to geocode Tweets that do not have location. In the section 3.5.1 we will apply standard LDA Topic Modelling to automatically identify topics in the available dataset. These results will be further explored in section 3.5.2 using Self Organizing Maps, a reference value and word clouds. Based on the results we will select the topics and Tweets of interest in section 3.5.3.

In section 3.6 we will visualize the selected topics spatially and temporally. In section 3.7 we will derive actionable information from the selected Tweets and validate both the content and the locations of the Tweets.

3.1 Preliminary exploration of the dataset

The first step in this study is to conduct a preliminary exploration of the dataset to gain insights in the temporal and spatial distribution of the dataset. We have written a Python script that takes the initial CSV with Tweets as input and writes it to a PostgreSQL database with a Postgis spatial extension. This database allows easy and efficient (spatial) querying to store, manipulate and analyze the data. All results of the remaining analyzing and processing steps will be saved to this database as new tables.

The temporal attribute of the Tweets depicts the time a Tweet has been posted. We have explored this by plotting the number of Tweets for every hour as shown in Figure 6. This figure reveals that most of the activity took place between 14:30 on fifth of January and 04:00 on sixth of January. An interesting observation is a steep increase in activity around 14:00 o'clock which reaches a maximum at 19:00. There is also a second peak around 23:00.

The spatial attribute is the geographic location of the Tweets. As described in the review chapter, the availability of a geographic locations depends on whether the user has added a GPS location when posting the Tweet. In the initial dataset approximately 1% (1250) of the Tweets has been geocoded by the users by means of geographic coordinate. We have used these Tweets to calculate the number of Tweets per municipality as shown in Figure 6Figure 7. Even though the majority of the Tweets does not have geographic coordinates, this figure shows that users from the entire country have contributed to the online activity. It can also be seen that most of the activity is very likely to have emerged from the surroundings of the incident.

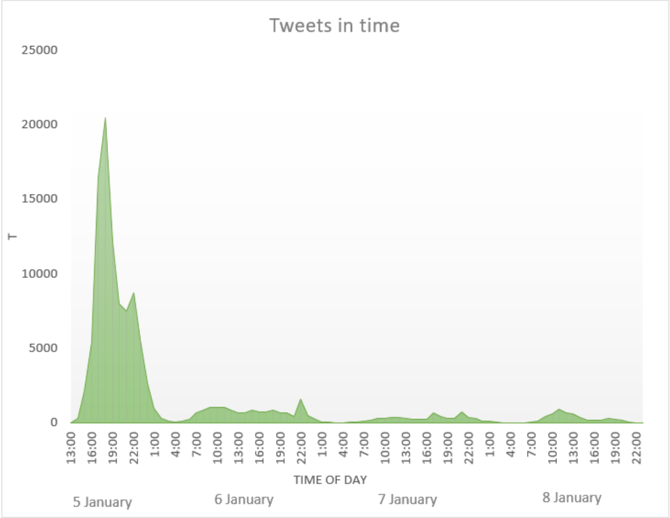


Figure 6: Temporal distribution of the Tweets as number of Tweets per hour.

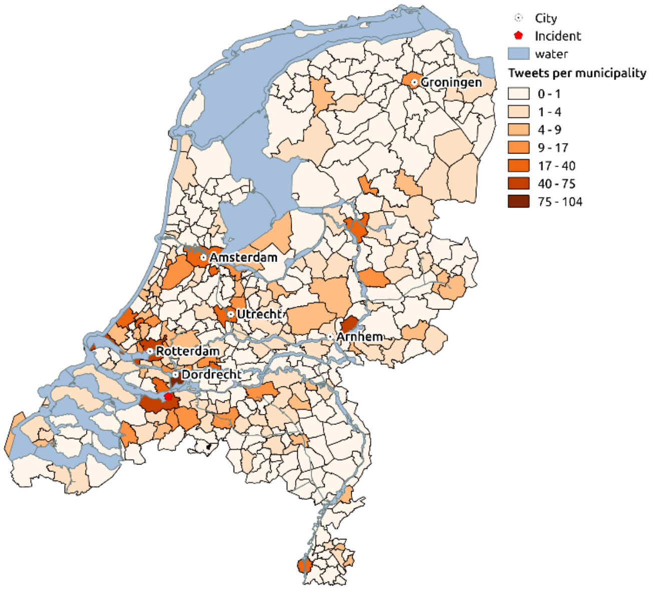


Figure 7: Number of geotagged Tweets per municipality

3.2 Pre-processing the dataset

The cleaning of the initial dataset consist of removing irrelevant columns, removing the screen names and creating geometries from the available coordinates. We consider columns to be irrelevant when these do not contain any information that is of use for the text and spatiotemporal analysis. In this particular case, the columns of the platform used by the user and the Uniform Resource Locator (URL) of the Tweets are removed. The *id* and *username* columns are considered to be useful because these can be used to retrieve additional metadata about the Tweets and the users at a later moment in this study. We have also added an additional column *geo_ref_by* which we will use to classify how a Tweet has been geo referenced for validation and accuracy purposes.

A visual inspection of the georeferenced Tweets in QGIS revealed that several Tweets have been posted outside the Netherlands. These Tweets are outside the study area and will therefore be removed. For this we have downloaded a polygon of the Dutch border from the Dutch governmental geographic data store Publieke Dienst Op

De Kaart (PDOK). This dataset has been written to the database using the database manager in Quantum Geographic Information System (QGIS). We have created a new table with all the Tweets that are inside The Netherlands using the within spatial query in Postgis. With this query we have created a new table from the selected points; the Tweets that fall inside the polygon representing the Dutch border. In the final step we have added all Tweets without a location to the output table using an insert selection query. These Tweets are included for further geo referencing steps in the following steps.

3.3 Retrieving additional location information

Exploration of the available dataset revealed that only 1250 Tweets have been geo located using the GPS receiver of the mobile devices. To enable the spatial visualization of the event's location, additional Tweets need to be geo located.

As mentioned in the review chapter, Twitter provides two additional sources of location information that can be used to associate a Tweet with a location: the profile locations and the place names.

The profile locations and the place names associated with respectively the users and the Tweets are currently not available in the dataset. These need to be retrieved from the Twitter databases. Twitter provides access to its database via the Application Programming Interface (API). The API can be used to retrieve additional meta-data and contextual information related to the Tweets, users, places and entities in the content such as media, hashtags and links (Twitter, 2015). We have created two methods to retrieve the profile locations for all users and the place names for the Tweets. The methods send a request to the API, parse the JavaScript Object Notation (JSON) which is returned and saves the found locations to the database. Both methods have been implemented in Python using the Twython package.

3.3.1 Retrieve profile locations

The profile locations can be retrieved using the users_lookup request which takes the usernames or user ids as a parameter. The access to the API is limited by time slots of 15 minutes, 180 requests per time slot and 100 usernames per request. The method takes these limitations into account by retrieving all unique usernames from the database and sending them in batches of 100 usernames per request as shown in Figure 8. The script is paused for 15 minutes when the API rate limits have been reached.

The API returns a JSON with the profile meta-data for every username that has been found in the database. A JSON is a standardized data-interchange format built up from structured key value pairs which can be easily generated and parsed with any programming language. The value of the key *location* contains the profile location for the requested username. This value can either be a string of text or empty. If it is a string, then this value is directly saved to the table. Empty values mean that the user has not set a profile location. In this case, the value *notSpecified* is saved to the database.

The API does not return usernames and profile information for profiles that have been deleted. To identify the usernames that are not existing anymore, an extra step has been included that compares the usernames that have been requested with the usernames that have been returned. If a username has not been returned, then this means that the user has been deleted. In this case, the value *userDeleted* is assigned to the *profile_location* of that user in the database.

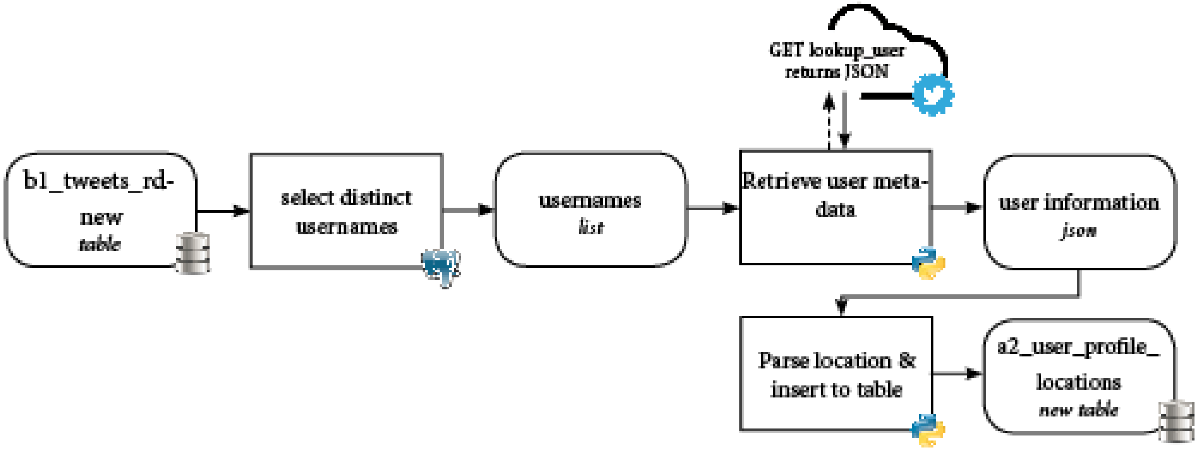


Figure 8: Flowchart of retrieving profile locations

3.3.2 Retrieving places

The places associated with the Tweets can be retrieved with the status_lookup request using the *Tweet id* as a parameter. The method for this step is similar to the method for retrieval of profile locations: query all Tweet id's from the database, request the meta-data for every 100 id's, parse the *place* key from the JSON if available and write the result to a new table as shown in Figure 9: Flowchart for retrieving, processing and writing places.. If the request does not return a result for a specific tweet id, then this Tweet has been removed and is therefore not available. In this case the value *tweetDeleted* is written to the table. An empty string value for the *place* key means that the user has not specified a place, which is written to the table as *notSpecified.*

The status_lookup request also returns a bounding box with an extent the size of the geographical boundary of the place that has been returned. These boxes are also written to a new table since these can be used to validate the geocoding steps taken in the next step of this study. We can write the boxes as Polygon geometries in Postgis by first converting the list of coordinates to a Well Known Text (WKT) object and then using the Postgis *ST_GeomFromText* method to create the geometries. WKT is an Open Geospatial Consortium (OGC) standardized text markup language for representing vector objects and transformations between spatial reference systems. The output of this step results in two tables, a3 which contains the places for every Tweet if available and a4 which contains the polygon geometries of the places.

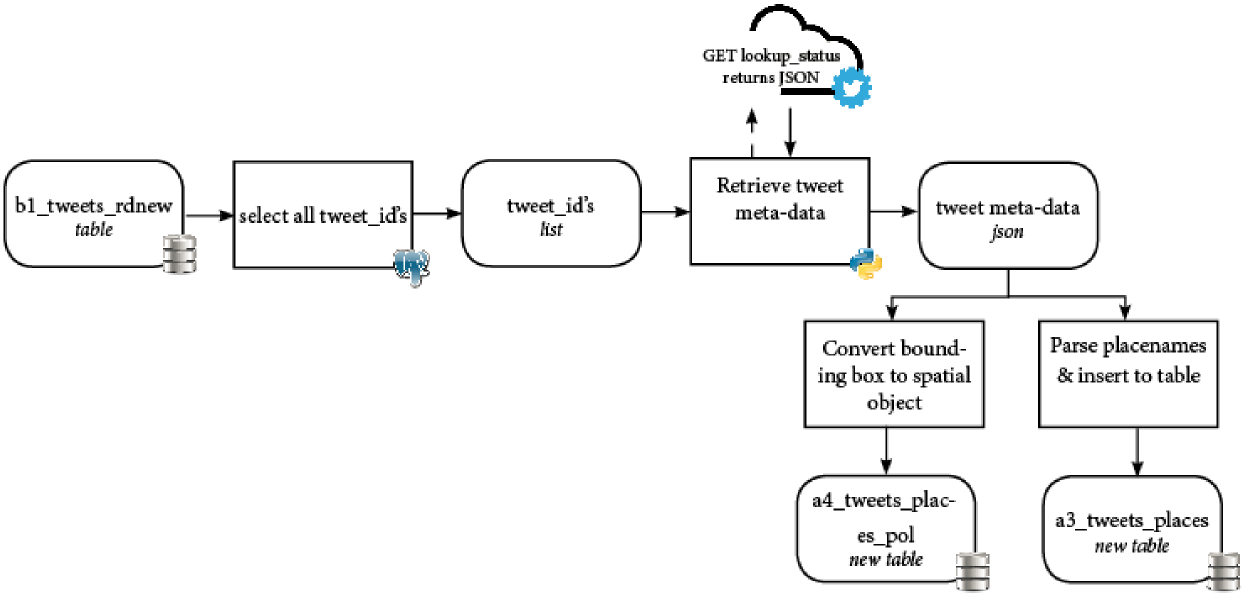


Figure 9: Flowchart for retrieving, processing and writing places.

3.4 Geocoding location information and Tweets

The retrieval of the profile locations and the place names from the Twitter API resulted in many text-based place names. To enable the spatial analysis, these place names need to be geocoded. Nowadays there are many online services available for geocoding. These services use open geographic databases with place names and corresponding exact coordinates. For this study we have chosen to use Mapzen, and more specifically the Mapzen Search API (Mapzen, 2017). The major advantage of this service is that it uses all available open geographic databases to find the best matching location. It also allows us to geographically narrow down the search by specifying a country in which the locations are to be searched and by place type. The place type range from fine granularity such as points of interests to coarser once like countries. An example of the possible place types and there spatial granularity is shown in Table 2.

Profile locations are text-based names which do not have geographic coordinates and therefore require an additional geocoding step before these can be used for spatial analyses. Geocoding is the process of translating text-based names into geographic coordinates using large geographic databases(Goldberg, 2011).

Table 2: Place types and granularities of Mapzen (Mapzen, 2017)

PLACE TYPE	GRANUALIRITY
VENUE	Points of interest, businesses
ADDRESS	Places with a street address
NEIGHBOURHOOD	Social communities
LOCALADMIN	Local administrative boundaries
LOCALITY	Towns, hamlets and cities
REGION	States and provinces
COUNTRY	Places that issue passports, nations, nation-states

3.4.1 Geocoding profile locations and place names

The steps taken to geocode the profile locations are shown in Figure 10. First, we have selected all available profile locations with a SQL query into a new table in step 3.4.1a. Profile locations that are missing are not taken into account to save processing time and limit the number of requests.

In step 3.4.1b we select all distinct profile locations and send each of the locations as a Hypertext Transfer Protocol (HTTP) request to the Mapzen Search API. The API returns a JSON with a list of features where each feature represents a result that matches the input location and parameters. The results are ordered according to a confidence level which is a value between zero and one. The confidence level is an estimation of how accurately the results match the query. It is based on parsing the input text and comparing this to the results. If the result perfectly matches the query and it is a valid place name, then the confidence level will be one. If the result does not entirely match the query, then the confidence level will be less than one. For example, searching for an address may return the geographic coordinates of the same street, but with a different house number. In this case the confidence level will be lower. We save the coordinates and place name of the result with the highest confidence level to the database for further processing.

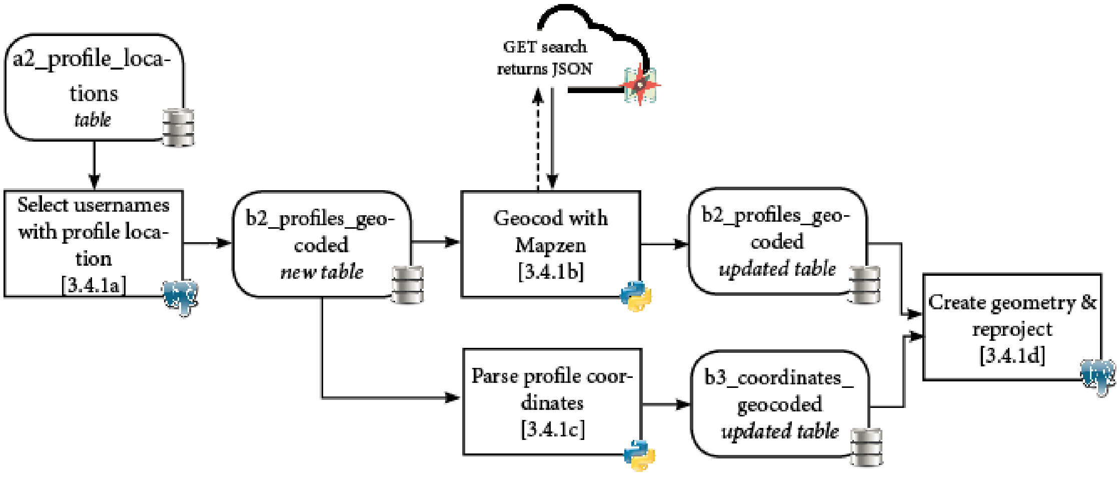

a2_profile_loca-tions *table*
GET search returns JSON
Select usernames with profile loca-tion [3.4.1a]
b2_profiles_geo-coded *new table*
Geocod with Mapzen [3.4.1b]
b2_profiles_geo-coded *updated table*
Create geometry & reproject [3.4.1d]
Parse profile coor-dinates [3.4.1c]
b3_coordinates_geocoded *updated table*

Figure 10: Method for geocoding the profile locations.

In step 3.4.1c we have defined a method that parses profile locations which contain geographic coordinates. The exploration of the profile locations revealed that many users use the GPS receiver of their mobile device to automatically update the profile location according to their actual location. Examples of this type of profile locations are shown in Table 3. The extra step to geocode these locations is required as Mapzen Search API can only geocode place names, and not geographic coordinates. The profile locations with coordinates are selected using string pattern matching SQL query. The selected profile locations are then parsed to extract the latitude and longitude coordinate which are saved to the table.

In the final step, step 3.4.1d, the point geometries are created based on the latitude and longitude coordinates and reprojected to RD_New

Table 3: Example of profile locations with coordinates.

USERNAME	PROFILE LOCATION
SNIRPJAN	iPhone: 51.842367,4.423487
CAMIELVDBERGH5	UT: 51.887812,5.192504
KAYINTVEEN	UT: 51.973705,4.483917
GERONIMO___NL	UT: 51.377659,6.12636
FRISKY_DE_GEUS	UT: 51.442891,5.475895
JORGDEGROOT	iPhone: 51.236628,5.686530
LEAL_ARAZZI	51.915296,4.438586
JPLUIMERS	Amsterdam 52.351317,4.769685
TOMVANDEVEN	iPhone: 51.503201,3.473545

The place names associated with the Tweets have been geocoded using a similar method as for the profile locations. The geographic coordinates of all found place names are send as a request to the Mapzen search API. The results are saved to a new table. Place names that did not return a geographic coordinates are saved as '*noResult*'.

3.4.2 Geocoding the Tweets by SQL queries

The final step of the geocoding step is to geocode the Tweets based on the available location information for every Tweet. This may be a profile location, profile coordinates, a place name or a GPS location. The preferred location to be used is the one with the lowest granularity and a certain temporal validity. The smaller the granularity of the location, the higher the precision of the spatial analyses related to identifying the location and spread of the plume will be. The temporal validity also needs to be taken into account since profile locations may have changed over time whereas GPS coordinates and places not. To select the best location for every Tweet, we have ordered the available location information according to their granularity and temporal validity as shown in Table 4.

Table 4: Type of locations order by granularity and temporal validity.

LOCATION TYPE	GRANULARITY	TEMPORAL VALIDITY
1. GPS	Point	Certain
2. PLACES	City	Certain
3. PROFILE COORDINATES	Point	Uncertain
4. PROFILE LOCATIONS	City	Uncertain

The GPS locations are the most certain locations as they represent the point location where the Tweet has been posted. This location has not changed over time and is therefore still certain. The places are assigned to a Tweet when the Tweet is created, the temporal validity has therefore not changed and is also certain. However, the granularity is coarser compared to GPS coordinates; in most cases the places are on city level. The profile locations are geographical coordinates and have therefore a finer granularity compared to the profile locations. However, both the coordinates and profile locations have an uncertain temporal validity since we have used current profile locations which may have been changed over time.

A model has been developed which automatically geocodes every Tweet by checking which locations are available for a Tweet. The first location that is found according to the order as specified above, is assigned to the Tweet. The output of this method is a new table with the geocoded Tweets.

3.5 Topic discovery and classification of the Tweets

To identify the topics in the dataset and classify the Tweets accordingly, we have used the Latent Dirichlet Allocation (LDA) Topic Modelling algorithm (Blei et al., 2003). This is an unsupervised probabilistic learning algorithm that can automatically discover latent (i.e. hidden) word structures (i.e. topics) in a set of text documents based on the co-occurrence of these words. The basic principle of LDA is that each document is considered to be a mixture of all identified topics. Therefore a probability is calculated for every topic *k* which denotes what proportion of document *d* is described in topic *k*. A topic is represented by all words in the corpus ordered by their relevance to the topic. The more relevant a word is to a topic, the higher it is ranked.

Based on the identified topics and the topic probability the algorithm classifies each text document accordingly. A text document is assigned (classified) to the topic with the highest probability. An example of this classification and the topic probabilities is shown in Table 5. An extensive description of the algorithm is included in Appendix A: Topic modelling explained.

Table 5: Example of topic modelling results for k=3.

document	Topic 1	Topic 2	Topic 3	Classification
1	0.25	0.60	0.15	2
2	0.01	0.3	0.69	3
3	0.13	0.33	0.54	3
…	…	…	…	…

The following steps in this section will describe the method we have developed in R to automatically apply the LDA topic modelling algorithm to our dataset. These steps are shown in Figure 11. The interpretation and understanding of the LDA output can become a complicated and time consuming task due to the large amount of numerical data as a shown in Table 5. Inspired by the suggestion of Chaney & Blei (2012) to simplify this task by using visualization methods, we have extended the LDA topic modelling with two additional steps. The first one is visualization of the results using Self Organizing Maps (SOM). This method reduces the number of dimensions to a two dimensional map and groups text documents with similar properties onto a two dimensional heat map. This method will allow us to identify cluster of Tweets with cohesive content based on the topic modelling probabilities. The second step is the use of word clouds to explore the themes that are discussed in a cluster. These methods will allow us to explore the content of the clusters which have been identified with the SOM heat maps. Based on this content we will select the Tweets which are to be used for the remainder of this study.

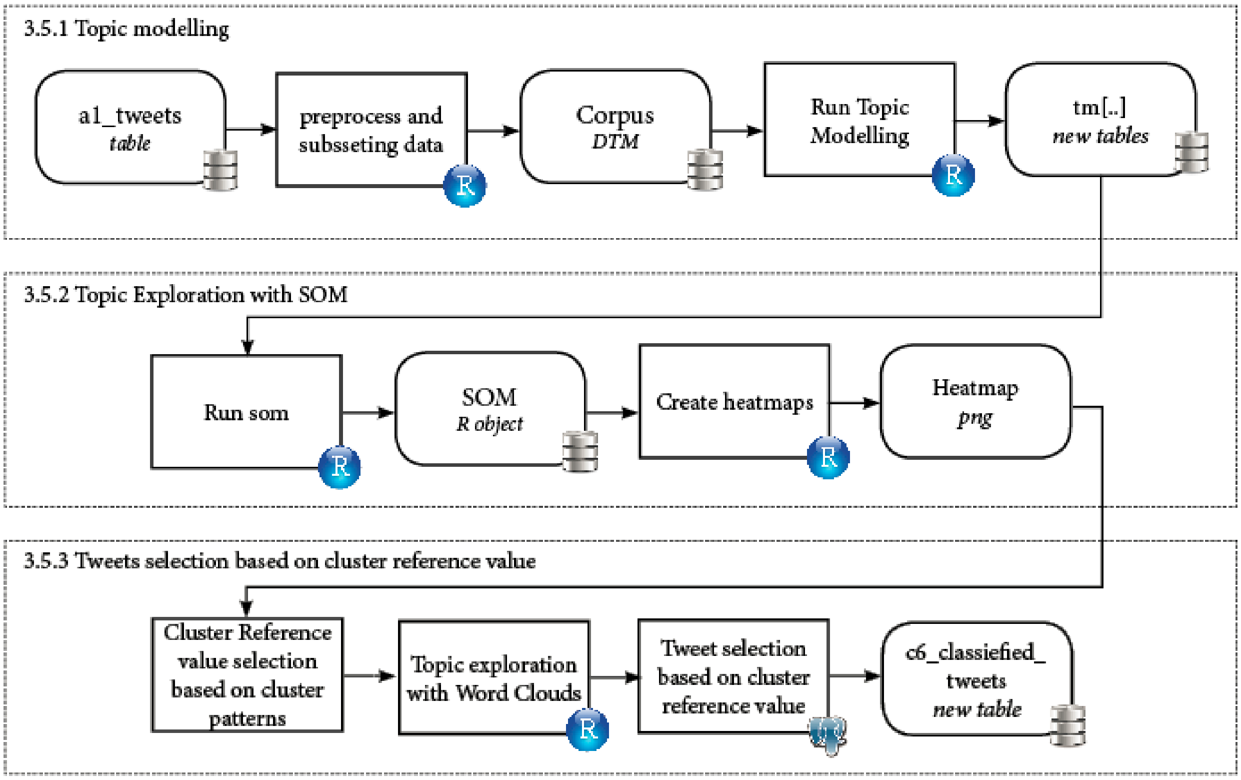

3.5.1 Topic modelling
a1_tweets
table
preprocess and subsseting data
Corpus
DTM
Run Topic Modelling
tm[..]
new tables
3.5.2 Topic Exploration with SOM
Run som
SOM
R object
Create heatmaps
Heatmap
png
3.5.3 Tweets selection based on cluster reference value
Cluster Reference value selection based on cluster patterns
Topic exploration with Word Clouds
Tweet selection based on cluster reference value
c6_classiefied_ tweets
new table

Figure 11: Flowchart for classifying the topic of Tweets.

In section 3.5.1 the preprocessing steps for the text data are described after which the topic modelling algorithm is applied. The document to topic distributions are then visualized in section 3.5.2. Finally, in section 3.5.3 the topics and corresponding Tweets will be selected by exploring cluster patterns in the SOM heat maps with word clouds.

3.5.1 Topic modelling

The topic modelling algorithm takes text documents as input. In this study a text document is one Tweet. In step 3.6.1a subsets of a two hours interval are created from the entire database starting at the first hour when the disaster was first reported at 16:27:00 on fifth of January and ends 24 hours later. The content of the Tweets in the subset is pre-processed in step 3.6.1b. Pre-processing in the field of text mining is “the process of cleaning and preparing the data for text classification” (Haddi, Liu, & Shi, 2013). This step is required as LDA topic modelling is a bag-of-words model in which relations between words are discovered based on their co-occurrence in the input data. The algorithm may identify relations between words that have no information value. Twitter data typically contains a lot of noise such as website links, emoticons, Twitter specific syntax, meaningless words etc.. As a result, noise and classification inaccuracy can be introduced to the output. To reduce these effects, a preprocessing step is required.

The preprocessing step in this study consists of removing text elements that have no information value. The cleaning of the text is done using the text mining package *tm* in R (Meyer, Hornik, & Feinerer, 2008). We have removed the following text elements from every Tweet:

- mentions '@username'
- URL's
- Annotation for retweets 'RT'
- '#brand' and '#moerdijk'
- stop words
- punctuation and capital letters

Username, website links and the 'RT" notation of the Tweets are common text elements in many Tweets. These are considered to be noisy and are therefore removed. The hashtags '#brand' and '#moerdijk' have been used by users to relate the Tweets to the incident and make these searchable in the Twitter feed. It is expected that the majority of the Tweets will contain these hashtags irrespective of the content and topic of the Tweet. The hashtags are removed as Tweets may be related to each other based on these hashtags. Moreover, the information value of making the topic searchable has been lost as the dataset has been created by filtering the twitter feed on these hashtags. The stop words are the most common words in language such as 'the', 'a' and 'that'. These words have also been removed as these are meaningless. The removal is based on the Dutch stop words list that is included in the *tm* package. Finally we have also removed the punctuation and lowered all capital letters to prevent the algorithm interpreting words such as 'brand' and 'Brand.' as two different words.

The preprocessed subsets of Tweets are used as an input for the topic modelling algorithm in step 4.6.1c. The topic models package in R which has been created by Grun & Hornik (2011) contains an implementation of the LDA algorithm. We hae used this implementation to apply topic modelling on our datasets. The algorithm requires a Document Term Matrix (DTM) as input. A DTM is a matrix in which the documents are listed as rows and the words of the entire corpus as columns. The matrix indicates how often word *w* occurred in document *d*. To limit the processing time we have removed the sparse items (i.e. words) from the matrix; words that appear in a very limited number of documents. This step and the preprocessing step may result in short documents being stripped from all words, leaving these empty. These empty documents have also been removed to further limit the processing time.

As described before, the algorithm requires setting a parameter *k* which denotes the number of topics to be identified. As there is no method for defining the optimal value for *k*, we have arbitrary chosen the value three based on the assumption that the subset of the two hours interval will significantly reduce the number of documents in the input. Too much topics may result in overlapping topics. The topic modeling algorithm is run for all subsets. The output is saved to the database for further processing.

3.5.2 Topic exploration with Self Organizing Maps

We have visualized the Topic Modeling output using Self Organizing Maps. SOM is an unsupervised machine learning algorithm that produces a two dimensional representation of a multi-dimensional input dataset, which is visualized as a heat map. The input dataset in this study is the output from the topic modeling algorithm: the tweet to topic probabilities for every Tweet. We can consider this output as a multi-dimensional problem with *k* dimensions where *k* is the number of topics that have been identified by the LDA algorithm. The heat map consists of nodes as shown in Figure 12. On each node one or more samples from the subset are mapped based on the dis- or similarity of the sample's properties for every dimension. The more similar the values of the properties are, the closer these samples are mapped to each other. The color of the node represents the value of one property. This mapping process results in clusters which are visually easily to identify and to compare for each topic in the subset.

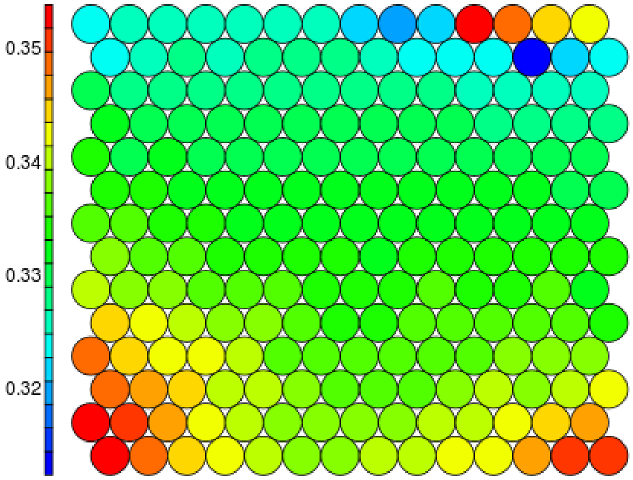


Figure 12: Example of the output of the Self organizing Map algorithm: a heat map.

The Kohonen package for the programming language R contains an implementation of SOM (Wehrens & Buydens, 2007). The SOM requires setting two parameters. The first parameter is the grid size of the heat map and the second is the number of iterations. We have used this implementation to explore the topic modelling results from the previous section. The following steps have been taken:

1. Parameter exploration of grid size *dim*
2. Parameter exploration of the number of iterations *rlen*
3. Running the model for all subsets in R

The first step is to conduct parameter exploration of the grid size *dim*. We have used a heuristics approach to explore these paramaters.The grid for the Self Organizing Maps consist of nodes on which samples are mapped. The size of the grid depends on the number of samples in the input dataset and their resemblances. Currently there is no method available for determining the optimal grid size beforehand. We have therefore conducted a parameter exploration by running the SOM algorithm using the Kohonen package on all subsets with a grid size varying from 8x8 nodes to 30x30 nodes with a step of two. The SOM implementation in R contains a *counts* plot which shows the number of samples that have been mapped to every node in the map. We have selected the optimal value for parameter *dim* by comparing all *counts* plots of one subset and picking the plot where the number of samples per node are evenly distributed over all nodes.

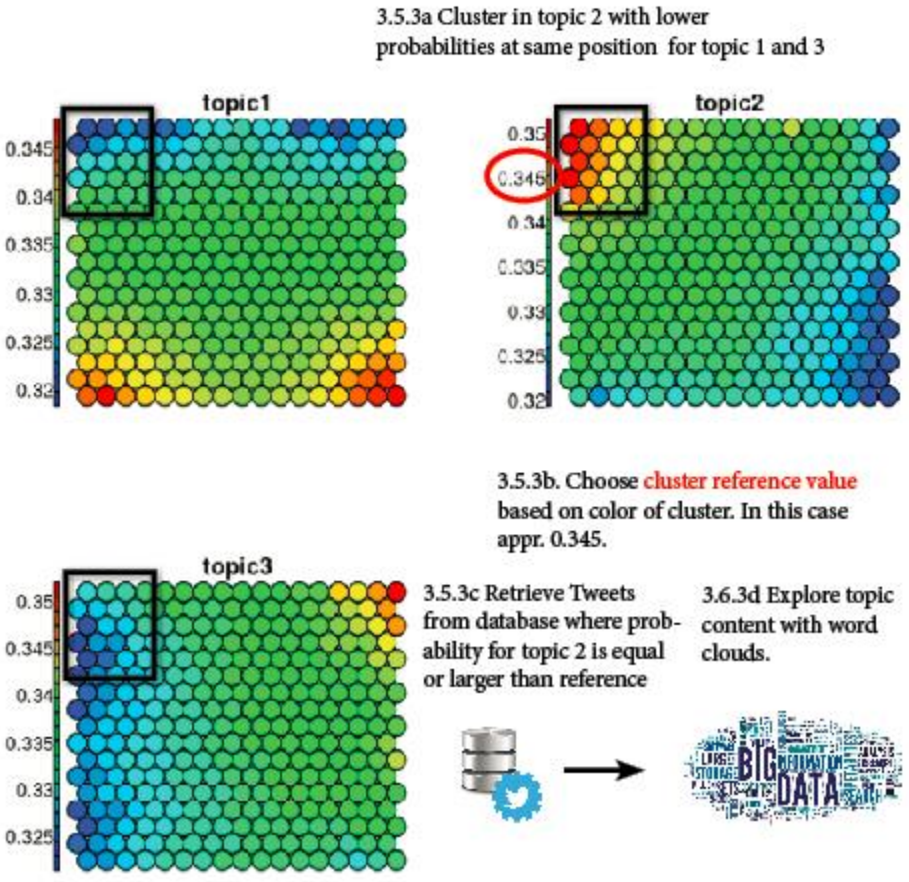
The second step is to determine the optimal value for *rlen*. This step also requires parameter exploration as the number of iterations depends on the size of the grid and the number of samples mapped on each node. We have picked the subset with the largest number of samples. We expect that the value for *rlen* will be sufficient for the other subsets as these contain less samples and have therefore smaller grids. The grid size for the runs is set to the value which we specified in the previous step. We have then run the SOM algorithm for the following *rlen* values: 100, 250, 500, 750 and 1000. For each value a mean distance graph is plotted that visualizes the training progress. The optimal number of iterations has been reached when the mean distance has reached a minimum and remains stable for an increasing number of iterations. We have chosen the optimal *rlen* value based on these plots and we will use this value for all subsets.

In the final step we have run the SOM algorithm on all subsets of our dataset using the identified parameters. The results are plotted as heat maps for each subset and each topic.

3.5.3 Tweets selection based on cluster reference value and word clouds

The SOM heat maps are a visual representation of the Tweet to topic probabilities. A tweet to topic probability denotes what proportion of the Tweet is described by each of the topics. The Tweets have been mapped based on the similarity of their Tweet to topic probabilities for every topic. If we consider the Tweet to topic probabilities as characteristics of the Tweets, then it can be stated that the SOM algorithm has mapped Tweets with similar characteristics closer together. As a result, clusters of Tweets with similar characteristics have formed on the heat maps. The heat maps allow easy visual identification of clusters and an understanding of which Tweets are related to one another considering the content of the Tweets. At this point however, it is unclear what these topics are about. Therefore the next step in this study is to explore the content of the topics using the clusters in the SOM heat maps and the Tweets which belong to these clusters. By doing so, we will be able to select the topics and Tweets of interest for this study.

Figure 13 shows a schematic overview of the four steps we have conducted to extract the meaning of the topics. We have used the SOM heat maps, a cluster reference value and word clouds.



3.5.3a Cluster in topic 2 with lower probabilities at same position for topic 1 and 3

3.5.3b. Choose cluster reference value based on color of cluster. In this case appr. 0.345.

3.5.3c Retrieve Tweets from database where probability for topic 2 is equal or larger than reference

3.6.3d Explore topic content with word clouds.

Figure 13: Overview of topic content exploration using SOM heat maps, a cluster reference value and word clouds.

In order to understand our approach, we need to consider the following aspects of the SOM heat maps:

- Each circle is a node
- On each node, one or more Tweets are mapped with similar characteristics
- Each node (and therefore each Tweet) is mapped on exactly the same position on each map
- The color represents a value for the topic probability which can be looked up in the legend on the left of the maps (this is not an y-axis)
- Every map represents a different topic probability for the same Tweet

In the first step (3.6.3a) we have visually explored the set of heat maps for each time subset. We have selected clusters that have a high probability for one topic, and lower probabilities for other topics. By doing so, we will ensure the ambiguousness of the Tweet's content is limited. The color of the nodes in the cluster represent a tweet to topic probability. We have chosen a probability value that matches the colors of the cluster in the heat map (step 3.6.3b). This probability is the reference value which we have used to retrieve the Tweets that belong to this cluster in step 3.6.3c. The Tweets have been retrieved using a database query which selects all Tweets with a probability equal to or larger than the reference value. We have applied word clouds in step 3.6.3d to analyze the content of every cluster. Wu, Provan, Wei, Liu, & Ma (2011) have shown us that word clouds can be used to create quick and simple visual summaries of text data based on word frequencies.

3.6 Spatial and temporal visualization

To visualize where the people have tweeted about the toxic plume, we have used the steps as schematically shown in Figure 14.

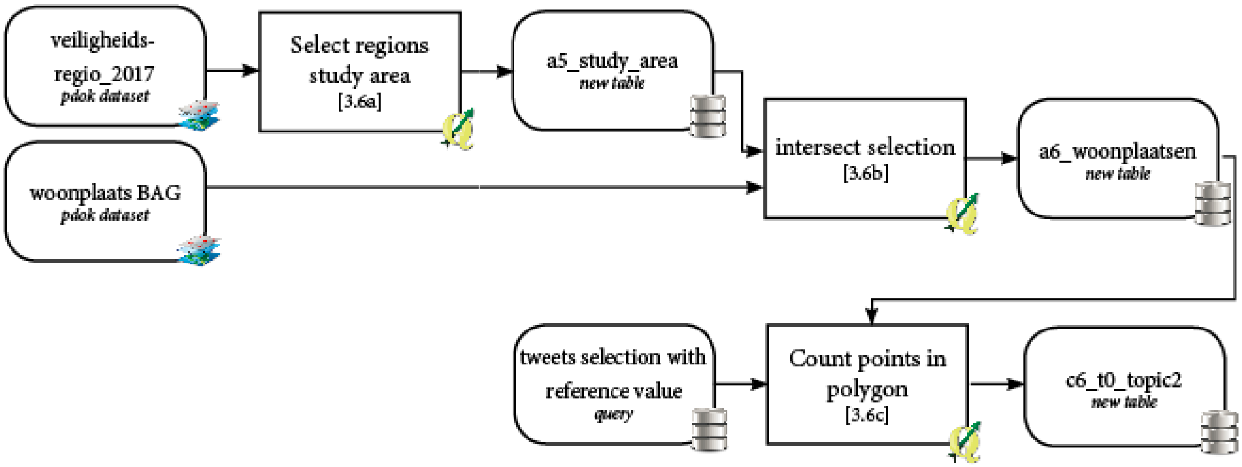


Figure 14: Overview of the methods used for the spatial analysis of the toxic plume.

We have first created a spatial extent of the study area based on the safety regions that have been involved in the incident, namely: the safety regions Midden- en West-Brabant, Zuid-Holand-Zuid and Rotterdam-Rijnmond. For this step (3.6a) we have retrieved a spatial dataset of all safety regions in The Netherlands from the Dutch open spatial database at PDOK (PDOK, 2017). We have manually selected the polygons of the three safety regions. These have been saved to our database as a new spatial dataset using QGIS.

We have visualized the Tweets related to the toxic plume using the Dutch 'woonplaatsen' dataset. 'Woonplaatsen' represent residential areas which are a geographical subdivision of municipalities. To retrieve the residential areas that fall within the study area, we have intersected this dataset with the study area in step 3.6b. The result has been saved as a new dataset in the database.

In the final step (3.6c) we have visualized the results by counting the number of Tweets for each residential area in the polygon using QGIS. The Tweets have been retrieved from our database using a query that selects all the Tweets that are larger or equal to the reference value for the topic of interest.

3.7 Deriving actionable information

Once we have obtained the Tweets for the topic related to the toxic plume and visualized these spatially, we need to determine whether the identified Tweets contain relevant information. We have defined a coding scheme inspired by the extensive work of Olteanu et al. (2015) to further categorize the content in Tweets. In this scheme we first differentiate Tweets in two categories based on their information value as shown in Table 6. We consider a Tweet to be informative if it contains information related to one of the categories as specified in Table 7. Not informative Tweets are messages which do not contain actionable information.

Table 6: Categories for differentiating Tweets based on information value for situational awareness.

	Category	Description	Specification
A	Informative	Related to the crisis and contains information for situational awareness.	Table 7: Categories of informative messages.
B	Not informative	Related to the crisis, but does not contain information for situational awareness.	Off-topic, humor, spam, rumors, irrelevant content etc.

Table 7: Categories of informative messages.

	Category	Description
A	Affected individuals	Reports about oneself, missing, causalities, harmed etc.
B	Infrastructure & utilities	Unavailability of structures, roads, services, damage to environment
C	Caution & advice	Warnings, advice, caution, tips, instructions, information source etc.
D	Information about hazard	Location of hazard, status of hazard, type of hazard, intensity etc.
E	Request for help/information	Information related to hazard, help with evacuation etc.
F	Other	Any other information that does not fit in one of the categories above.

We have further extended the coding scheme with a validation step for the location of the Tweets. Majority of the Tweets has been geocoded in section 3.4.2 based on profile locations and places. To validate whether this location is correct we have compared the location of the Tweet with the location mentioned in the content. For this step we have defined three categories as shown in Table 8.

Table 8: Categories for location verification.

	Category	**Description**
A	Location valid	Location mentioned in Tweet matches location of Tweet
B	Location not valid	Location is mentioned in the content, but it does not match location of Tweet.
C	No location mentioned	No location is mentioned in the content. Validation not possible.

If the location mentioned in the content of the Tweet matches the location of Tweet. Than we consider the Tweet to have a valid location. If a location is mentioned, but it does not match the location, than the Tweet's location is considered not to be valid. Tweets that do not mention a location cannot be validated.

This code scheme has been applied by retrieving all Tweets from the database of the topic with a location. Each Tweet has been manually annotated in three steps. First, the content has been evaluated as informative or not. Second, if a tweet is considered informative, then it is categorized according the information categories in Table 7. In the final step, the location of the Tweet is validated according to categories in Table 8.

4. Results

4.1 Retrieving location information

The dataset contains 47,812 unique users which have contributed to the online activity during the Moerdijk incident. The profile locations have been retrieved from the Twitter databases using the Twitter API and the results are shown in Table 9. 46% of the users has specified a location in their profile and 18% has not done so. 36% of the requests did not return any meta-data related to the user, meaning that these users have been deleted. This is relatively large number of users which can be clarified by the fact that the incident took place six years ago and users may have stopped using Twitter by deleting their account.

Table 9: Results of retrieving profile locations from Twitter API.

	Usernames	Percentage of total
Location specified	21,833	46%
User deleted	17,295	36%
Location not specified	8,684	18%
Total	47,812	100%

The places that have been associated with a Tweet have also been retrieved for every Tweet. The results of this step are shown in Table 1010. Only 1% of the Tweets has been associated with a place and 62% not. 37% of the Tweets have been deleted implying that these could have been associated with a place, but this information is not available anymore.

Table 10: Results of retrieving places from Twitter API.

	Tweets	Percentage of total
Place specified	1,137	1%
Tweet deleted	43,995	37%
Place not specified	72,741	62%
Total	117,873	100%

Table 111 presents the type of places that have been retrieved. Majority of the places is a city. The places on country level are Tweets which have all been associated with The Netherlands. In this research we are mostly interested in city level

Table 11: Places types of the retrieved places.

	Places	Percentage of total
City	1,084	95%
Admin	38	3%
Country	15	2%
Total	1,137	100%

4.2 Geocoding location information and Tweets

4.2.1 Geocoding location information

We have used 22,330 profile locations which have been retrieved from the Twitter API to geocode. 21,366 of these locations are place names which have been geocoded using the Mapzen Search API. The remaining 964 profile locations consisted of coordinates which have been parsed from the profile location using pattern matching. The results are summarized in Table 12.

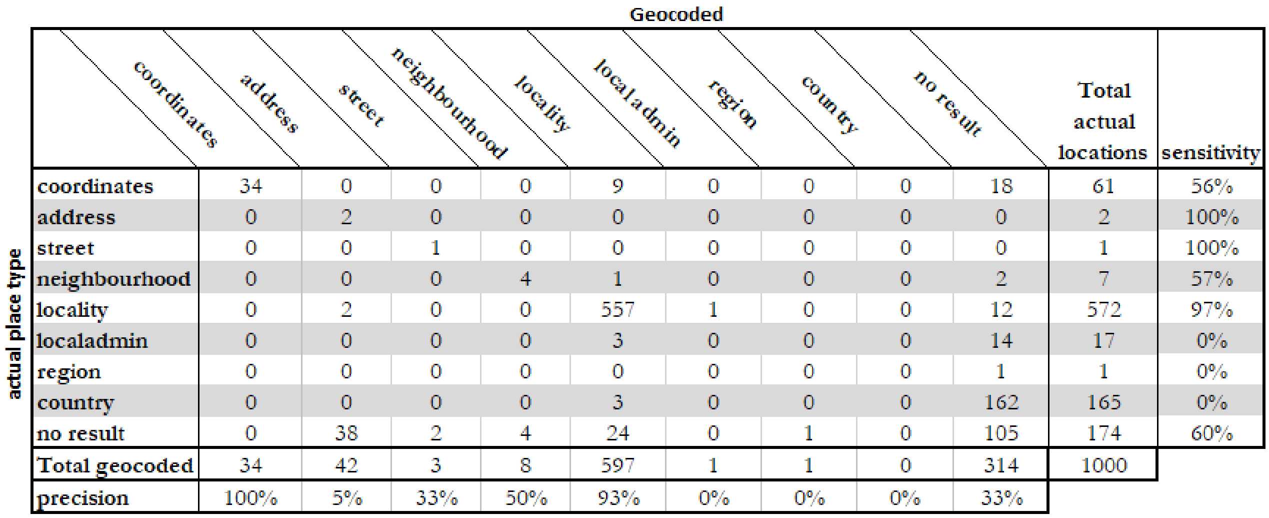
Table 12: Results of geocoded locations.

PLACE TYPE	LOCATIONS	PERCENTAGE OF TOTAL
COORDINATES	818	5.1%
ADDRESS	954	6.0%
STREET	31	0.2%
NEIGHBOURHOOD	241	1.5%
LOCALITY	13,901	87.0%
LOCALADMIN	4	0.0%
REGION	10	0.1%
COUNTRY	16	0.1%
NOT GEOCODED	6,355	28.5%
GEOCODED	15,975	71.5%
TOTAL	22,330	100%

Approximately 72 percent of the profile locations has been successfully geocoded. The remaining 28 percent of the Tweets has not been geocoded which implies that these are invalid or non-existent locations. Of the geocoded profile locations, 87 percent of the locations is a locality (e.g. city, village or a town). An interesting result is the relatively large number of coordinates and addresses. These type of locations are preferred as these have a lower granularity and allow for higher precision spatial analyses later in this study.

However, from this summary it cannot be determined how accurate the geocoding method is. Profile locations may have been geocoded incorrectly which can increase the uncertainty in the spatial analyses via error propagation. To explore the performance and accuracy of the method, we randomly selected 1000 profile locations. Every location is checked manually to determine whether it is a valid location and whether the correct place type has been assigned. The results of this validation step are summarized in a confusion matrix as shown in Table 13. The confusion matrix also shows the precision and the sensitivity for each place type.

Table 13: Confusion matrix for the geocoded profile locations.



		Geocoded										
		coordinates	address	street	neighbourhood	locality	localadmin	region	country	no result	Total actual locations	sensitivity
actual place type	coordinates	34	0	0	0	9	0	0	0	18	61	56%
	address	0	2	0	0	0	0	0	0	0	2	100%
	street	0	0	1	0	0	0	0	0	0	1	100%
	neighbourhood	0	0	0	4	1	0	0	0	2	7	57%
	locality	0	2	0	0	557	1	0	0	12	572	97%
	localadmin	0	0	0	0	3	0	0	0	14	17	0%
	region	0	0	0	0	0	0	0	0	1	1	0%
	country	0	0	0	0	3	0	0	0	162	165	0%
	no result	0	38	2	4	24	0	1	0	105	174	60%
	Total geocoded	34	42	3	8	597	1	1	0	314	1000	
	precision	100%	5%	33%	50%	93%	0%	0%	0%	33%		

The confusion matrix immediately shows that there are several place types with a very low precision. The precision states how much of the geocoded locations of a certain place type actually belong to that place type in reality. Place types with a low precision are the addresses, street, neighborhoods and the profile locations with no result. Of the 42 profile locations which have been geocoded as an address, only two locations are actual addresses, two are localities and 38 are invalid. This means that the place type 'address' is very inaccurate and recognizes invalid locations as valid once. The place type street and neighborhood also have a low precisions, but these percentages are heavily influenced by relatively low number of profile locations which have been geocoded as such. The low precision of the place type 'no result' can mostly be explained by the large number of profile locations which have not been geocode but are countries in reality. The low sensitivity implies that many profile locations which have not been geocoded are actual locations from other place types. Finally, the place type 'coordinates' has a 100% precision, meaning that if coordinates have been parsed from the profile location, these are always correct. However, the sensitivity of the coordinates place type is relatively low. Our method did not recognize 18 profile locations with coordinates as such.

4.2.2 Geocoding place names

The place names which have been associated with the Tweets have also been geocoded using the Mapzen Search API. The results of this step are shown in Table 14.

Table 14: Geocoding place names result

PLACE TYPE	LOCATIONS	PERCENTAGE OF TOTAL
LOCALITY	890	78%
LOCALADMIN	184	16%
COUNTRY	39	3%
NOT GEOCODED	24	2%
GEOCODED	1113	98%
TOTAL	1137	100%

98 percent of the place names has been geocoded using our method. The majority of the place names are localities, followed by local administrations (e.g. provinces, municipalities). To validate the results we have used the bounding boxes of each place name - which we have retrieved from the Twitter API - and intersected these with the point which is geocoded using the Mapzen Search API. The query for this intersection is shown in Code snippet 1. We consider a geocoded place name as valid if the point location is within the polygon of the place name. This validation method returned 10 locations which did not intersect with the polygon of that place name. In the remainder of this research we will only take into account the validated place names for the spatial analysis.

```
SELECT b4_places_geocoded.*, ST_Within(b4_places_geocoded.geom,
a4_places_polygon.geom) INTO c2_places_validated FROM b4_places_geocoded,
a4_places_polygon WHERE b4_places_geocoded.tweet_id = a4_places_polygon.tweet_id AND
b4_places_geocoded.place_type != 'admin' AND b4_places_geocoded.place_type !=
'country';
```

Code snippet 1: Query for validating the places.

4.2.3 Geocoding the Tweets

Table 15 shows the geocoding of the Tweets result. We have geocoded Tweets based on GPS coordinates, places associated with Tweets, coordinates mentioned in the profile location and profile locations. We were able to geocode 37,049 out of the 117,882 Tweets in the dataset. By doing so we increased the number of Tweets with a location from one percent to 31 percent.

Table 15: Results of geocoding the Tweets.

PLACE TYPE	NUMBER OF TWEETS	PERCENTAGE OF TOTAL
GPS	1250	3.4%
PLACES	663	1.8%
PROFILE COORDINATES	1762	4.8%
PROFILE LOCATION	33,374	90.1%
TOTAL	37,049	100%

The spatial distribution of the geocoded Tweet is shown in Figure 15: Spatial distribution of the geocoded Tweets.. This map shows the number of Tweets that have been posted for every municipality in The Netherlands. It can be seen that a large amount of the online activity has taken place in the surroundings of the incident.

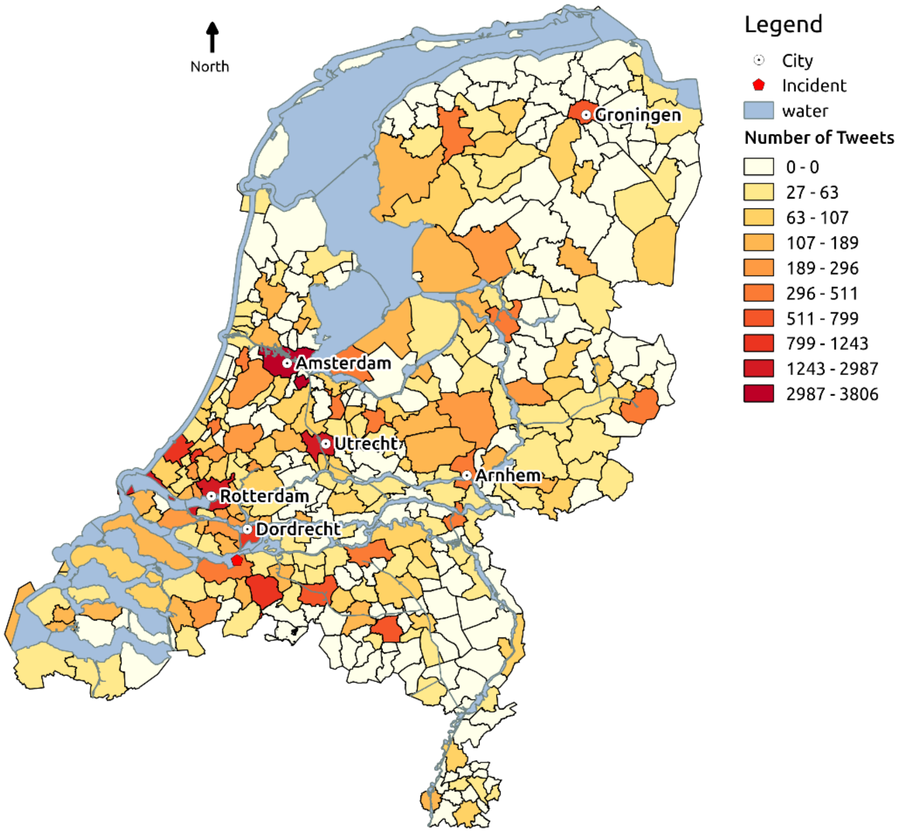


Figure 15: Spatial distribution of the geocoded Tweets.

4.3 Topic discovery and classification of tweets

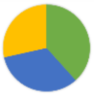
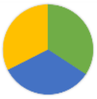
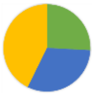
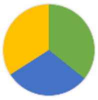
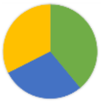
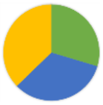
4.3.1 Results topic modelling

A summary of the results from topic modelling algorithm which we have applied on the subsets of the data are shown in Table 166. For every subset the number of Tweets, the Tweets per topics and the top six topic terms have been shown. The results have been obtained in three steps. First, subsets have been created using the initial dataset based on a two hour time interval starting at the moment the incident was first reported until the incident was mitigated. This resulted in 92,360 Tweets which have been further pre-processed in the second step. During this step the Tweets have been cleaned and stripped from words and characters that hold no information value, leaving the final input dataset for the topic modelling algorithm of 89,590 (97%) Tweets. Tweets have been lost during the pre-processing as these may have been completely stripped from the content. In the final step we applied the topic modelling algorithm on every subset for $k = 3$.

The number of Tweets show that users have mostly been active two hours after the incident first has been reported. As midnight approaches the activity gradually has decreased, reaching a minimum during the night. All topics have been equally discussed during each subset. Only topic one in subset three shows a larger amount of Tweets compared to the other topics in this subset.

The topic terms provide insight in the content of the Tweets that have been assigned to each topic. These terms are sorted by the probability that a term occurs in a topic. The probability for each word is different for every topic. The basic assumption of LDA topic modelling is that the content of each text document is a mixture of all topics. Each topic is therefore described by all terms in the corpus, but with a different word to topic probability. The lists of top terms as shown in Table 166 show a minimal differentiation. The term 'moerdijk' and 'brand' are present in almost every subset. The first topics in the first two subsets show words that are related to the toxic plume such as 'gifwolk', 'luchtalarm' and 'rookwolk'. From subset 3 onwards there is a major overlap between the terms like 'vuurbal', 'jonguh', 'attack', 'trending' etc. Ideally the topic definition consist of terms which are cohesive within one topic, and differs between topics. This is clearly not the case in these results. Based on these results in is not possible to select a topic for further processing without additional processing steps.

Table 16: LDA topic modelling results.

Subset	Time interval	Tweets	Tweets per topic	Terms topic 1	Terms topic 2	Terms topic 3
1	14:27:00 16:27:00	4,042		Moerdijk, grote, brand, luchtalarm, chemisch, zeer, bedrijf, deuren, woedt, dordrecht	Brand, moerdijk, ramen, grote, dordrecht, deuren, zeer sirenes, grip, ventilatie, bedrijf	Moerdijk, brand, grip, grote doredrecht, ramen, bedrijf, af, sluiten, rook
2	16:27:00 18:27:00	30,628		Gifwolk, grote, moerdijk, luchtalarm, vuurbal, brand, deuren, stoffen, richting, chemisch	Brand, moerdijk, grote, dordrecht, komt, rtl, ramen, bedrijf, dicht, gifwollk	Moerdijk, gifwolk, luchtalarm, brand, ramen, jonguh, vuurbal, mensen, deuren, chemiepack
3	18:27:00 20:27:00	25,532		Grote, moerdijk, vuurbal, brand, deuren, gifwolk, trending, ramen, neem, komt	Jonguh, vuurbal, moerdijk, grote, brand, trending, gifwolk, ramen, just, dutch	Moerdijk, brand, vuurbal, jonguh, rook, topic, gifwolk, buiten, trending, Rotterdam
4	20:27:00 22:27:00	15,239		Grote, jonguh, vuurbal, dutch, moerdijk, attack, look, say, think, fotos	Vuurbal, moerdijk, jonguh, grote, brand, america, dutch, people, weer, say	Grote, jonguh, vuurbal, panic, moerdijk, trending, brand, think, look, make
5	22:27:00 00:27:00	11,865		Grote, jonguh, moerdijk, vuurbal, dutch, say, trending, panic, think, terrorist	Vuurbal, grote, jonguh, brand, america, panic, look, dutch, attack, think	Vuurbal, jonguh, moerdijk, grote, dutch, look, brand, attack, terrorist, say
6	00:27:00 02:27:00	2,284		Vuurbal, grote, moerdijk, brand, jonguh, dutch, think terrorist, panic, look	Grote, jonguh, moerdijk, vuurbal, say, dutch, attack, brand, america, terrorist	Vuurbal, moerdijk, jonguh, brand, controle, brandweer, panic, say, look, trending
	Total	89,590				

4.3.2 Topic exploration with Self Organizing Maps

We visually explored the heat maps which have been created with the SOM algorithm. We identified clusters which have a relatively higher probability for one topic and lower probabilities for the other topics to ensure the cohesiveness of the topics.

Table 177 shows the references values which we have chosen for each cluster, in each topic and for every subset. The heat maps can be found in Appendix B: Results of the Self Organizing Maps.

Table 17: The selected reference value for every topic in every subset.

subset	Topic 1	Topic 2	Topic 3
1	0.342	0.34	0.335
2	0.356	0.35	0.342
3	0.35	0.365	0.356
4	0.35	0.35	0.35
5	0.347	0.347	0.347
6	0.35	0.35	0.345

4.4 Spatial and temporal visualization

The purpose of the word clouds is to get an understanding of the contextual information of the topics selected from the SOM clusters in each subset. We have selected the word clouds which match events that occurred in reality and visualized these in Figure 20: Example of Tweet with multiple location mentions.6. We have included the word clouds from several subsets above the timeline. The actual events are shown below the timeline. The word clouds that did not contain words related to the incident have been excluded from this figure. An overview of all word clouds can be found in Appendix C: Results of the Tweet selection based on a cluster Reference Value. The information for the comparison of the events on Twitter and the events in reality has been extracted from the official investigation report of the incident.

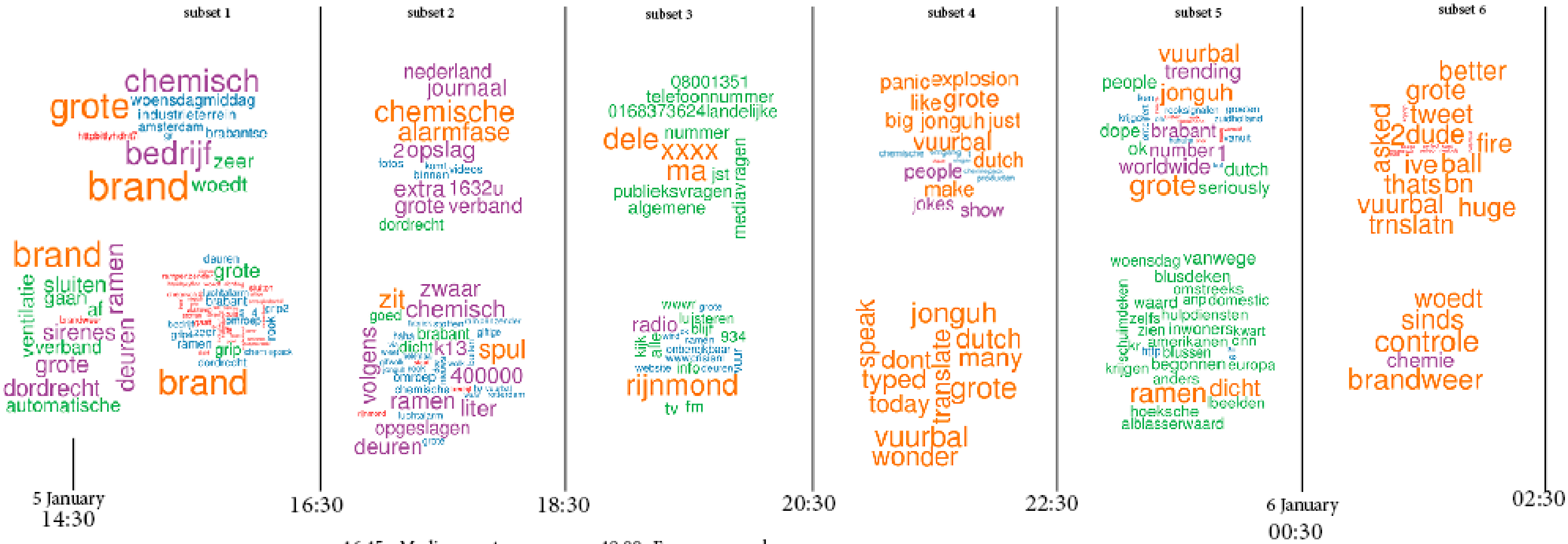
The fire was first reported approximately at half past two in the afternoon on 5th of January 2011. This event is clearly visible in the first word cloud in which a blazing fire is mentioned together with words related to industrial area and chemicals. The fire officer on the scene immediately scaled the fire up to a GRIP 2 incident since the thick black smoke was driven north in the direction of Dordrecht and Rotterdam. At 15:30 the incident is sized up to a GRIP 4. This event seems to be depicted by the third word cloud in the first subset. This word cloud contains words as ‘grip2’, ‘grip4’ and the name of the industrial site ‘chemiepack’. In the meantime the sirens were sound in the places Moerdijk, Mookhoek, Strijen, Strijensas and Willemsdorp as these were covered by the plume. The local population was urged to stay inside and to close all windows and doors to limit their exposure to the plume. This event is present in the second (and partly in the third) word cloud. In these word couds words related to the windows, doors, ventilation and sirenes are included.

In the second subset (16:30 – 18:30) the media spread false information on the presence of large amounts of a highly toxic chemical substance, k13. The resulted in public distress which was reinforced by conflicting information provided by various governmental officials. This event is also widely discussed on Twitter as can be seen in the word cloud for this subset. The terms 'k13', 'chemisch', '400,000' an 'liter' are describing this event. Around 17:00 o'clock the intensity of the fire increases and various explosions and fire balls can be seen and heard from a distance. This events has also been picked up by the users on Twitter, but the effect is not visible before subset 4. The topic is heavily discussed in subset 4 and becomes worldwide trending can be seen in the word clouds in Appendix C: Results of the Tweet selection based on a cluster Reference Value. The words 'grote', 'vuurbal' and 'jonguh' refer to a scene from a Dutch action comedy movie.

At 19:00 the governmental agencies initiate communication efforts to inform the population on the incident and the possible threats of the plume. A radio station and a website are appointed as the main sources of information for people affected by the incident. Shortly after this announcement the website was offline due to the large amount of visitors. The word cloud in subset three contains keywords related to this event.

At 23:00 the emergency services on the scene decide to attack and extinguish the fire with a so called "foam-blanket". It is believed that this will generate significantly more toxic particles which will be spread by the plume and therefore the public was warned again via the media and the sirens around 22:37. The second word cloud in in subset four has captured this event.

Shortly after the foam blanket was applied, the fire services gained control over the fire and incident was mitigated. This can also be seen in the word cloud in subset 6. The investigation report does not mention any activity related to subset 7 and 8. The word clouds from these subset only contain keywords related to scene from the Dutch movie. Also, the activity on Twitter has decreased substantially during the night.



14:27 - First report of incident

15.18 - Sirenes activitated in Mookhoek, Strijen, Strijensas en Willemsdorp.

16:45 - Media reports on presence of 400.000 liter of toxic materials.

17:00 - Fire grows rapidly. Various explosions and fire balls can be seen.

19:00- Emergency website and information phonenumber offline.

22:37 - Sirenes activitated due to fire extuinghishing practices

00:30 - Fire is extuinghished

Figure 16: Comparison of the word clouds with the events that occurred in reality.

In the results above only one word cloud contained keywords which are possibly related to the toxic plume. That is the word cloud in the first subset containing the words 'ventilatie', 'brand', 'ramen' and 'deuren'. Although no terms are present in the word cloud related to the toxic plume, we assume that this word cloud contains Tweets which are directly related to plume. Figure 17 shows the spatial distribution of the Tweets related to the word cloud which have been geotagged. In this Figure it can be seen that a large amount of the Tweets is originating from surroundings of Dordrecht and Rotterdam. But the activity is not limited to these areas. South from the incident also shows activity. However, this area has not been affected by the incident. Further exploration is required to discover whether these Tweets contain actionable information for the emergency services.

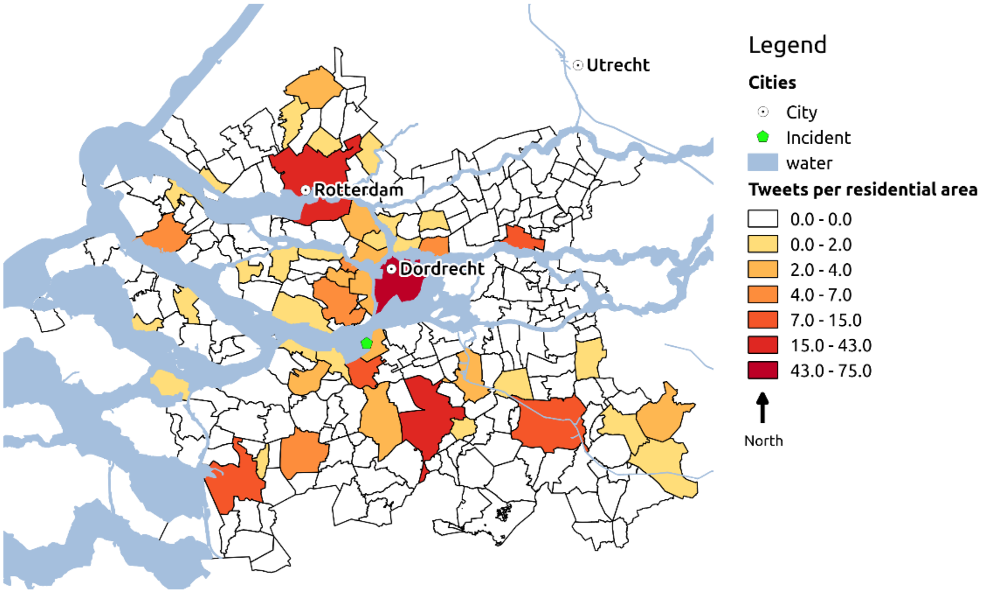


Figure 17: Spatial distribution of the Tweets related to the toxic plume.

4.5 Derived actionable information

We have derived the actionable information from the topic related to the toxic plume. The results are shown in
Table 188. In total 207 Tweets were available with a location related to the toxic plume. 59% of these Tweets contained information which have been annotated as informative for the emergency services.

Table 18: Information value of Tweets.

Information value	number of tweets	% of total
Informative	123	59%
Not informative	84	41%
Total	207	100%

The information categories and the number of Tweets per category are shown in **Error! Reference source not found.**9. 42% of the Tweets contained information related to the hazard. Many Tweets mentioned the presence of the toxic plume in their surroundings. The second largest category is people warning others to close the windows and doors and the reason.

Table 19: Tweets per information category.

Information category	# tweets	% total	Example
Affected individuals	17	14%	"ik ga nu mijn ramen sluiten. Grote brand op industrie terrein Moerdijk"
Infrastructures & utilities	5	4%	"A17 is afgesloten in beide richtingen, in het gebied waar rook kan komen #Moerdijk"
Caution & advice	33	27%	"Sirenes gaan af in Dordrecht in verband met de grote brand in #Moerdijk. Ramen en deuren sluiten, automatische ventilatie uitschakelen aub."
Hazard information	52	42%	"Hier in Wielwijk, Dordrecht nog geen luchtalarm gehoord...zie wel de rookwolken overdrijven....#moerdijk"
Help & information request	12	10%	"Ruikt iemand al iets van chemische lucht afkomstig van brand op moerdijk?"
Other	4	3%	"in de mookhoek komen stukjes as naar beneden brandmoerdijk"
Total	123	100%	

We have also validated whether the location of Tweet is correct based on the content of the Tweet. These results are shown in

Location	Number of Tweets	% of total
Tweet matches user location	67	33%
Tweet does not match user location	32	15%
No location	108	52%
Total	207	100%

Table 20. In 33% of the cases the location mentioned in the content matched the location which was assigned to the Tweet based on a user profile or a place (see 3.4.2). We consider these locations therefore to be valid. In 15% of the cases a location was mentioned in the Tweet, but this did not correspond with the assigned location. 52% of the Tweets did not mention a location in the content. The location of these Tweets could therefore not be validated.

Table 20: verification of the Tweet locations

Location	Number of Tweets	% of total
Tweet matches user location	67	33%
Tweet does not match user location	32	15%
No location	108	52%
Total	207	100%

5 Conclusions

5.1 Conclusion

The objective of this research was to study what critical location based information can be derived from Twitter during the response phase with the aim to support the decision making process in disaster management. We retrieved meta-data of the user profiles and Tweets from Twitter to geocode Tweets that did not have a location. We used a geocoding API to geocode the place names in the meta-data to geographical coordinates. We discovered that the majority of the meta-data contained location information on city level. We also found that although the geocoding API can geocode various place types ranging from countries to addresses, only the geographical coordinates of the cities is sufficiently accurate. The number of Tweets which has been associated with a place is limited. Nevertheless, with the combination of the place names and the profile locations we could increase the number of geotagged Tweets from 1.2% to 31%. We conclude that with the additional location information provided by Twitter it is possible to significantly increase the size of the spatial dataset. This compared to the initial dataset which only contained a limited number of Tweets geotagged by the user.

We applied LDA Topic Modelling to automatically derive events from the available dataset. The initial results did not reveal any cohesive topics. We introduced a novice approach to study the results in more depth by using Self Organizing Maps and word clouds. With this method we simplified the results from mutli-dimensional numerical data to two dimensional data. By doing so we managed to create heat maps and expose clusters of Tweets with cohesive topics. The word clouds of these clusters revealed that topics were identified that matched events that occurred in reality. However, the number of Tweets belonging to these topics was limited. The majority of the Tweets had an ambiguous topic classification. We therefore conclude that LDA topic modelling has a limited performance when used on Twitter data. We can also conclude that the use of SOM and word clouds makes the LDA results easier to analyze and interpret than without these extra methods.

We found one topic in the LDA results which we considered to be of interest for the emergency services. This topic was related to the toxic plume. We have used the Tweets related to this topic to validate the locations of the geocoded Tweets and to assess whether the Tweets contained critical information for decision making purposes. We found that half of the Tweets contained actionable information and that one third of the Tweets was correctly geocoded using the places and profile locations. Furthermore, as the majority of the location information had a city level granularity, it was only possible to spatially visualize the event on the granularity level of residential areas. The emergency services cannot depict the exact location of the toxic plume on this granularity level, but it does show the areas of potentially affected people. Based on this information the emergency services can for example decide whether evacuations are necessary.

5.2 Discussion

5.2.1 The dataset

The dataset of 117,183 was collected shortly after the incident in 2011 using the freely accessible Twitter API. While the data may contain valuable information for disaster management, questions arise considering the completeness of the data (Harrald & Jefferson, 2007). The free API limits the accessibility to the data by only returning 1% of all Tweets that have been posted. The entire dataset can only be accessed via Firehouse, the commercial service. Morstatter, Pfeffer, Liu, & Carley (2013) studied whether the sampled data from the free API covers the actual activity on Twitter as a whole. They found that it only covers 50% to 60% of the available data compared to the Firehouse service. This means that in reality, more content has been posted, possibly more topics have been discussed and more people have been involved on Twitter than actually available in the dataset. The geotagged content on the other hand is complete. Morstatter et al. (2013) found that 98% of the geotagged Tweets is returned.

When considering the completeness of the dataset, one should also consider to what extent the population is represented. Although we did not find any statistics for The Netherlands, in general it can be said that Twitter is most favored amongst people living in urban areas and are at the age of 18 to 30 years (Croitoru et al., 2013). The visualization of the Tweets related to the toxic plume (Figure 188) confirms this trend: the majority of the online activity seems to be originating from the cities such as Rotterdam, Dordrecht, Breda and Bergen op Zoom. However, a comparison with statistics of the population per 500 square meters in The Netherlands (CBS, 2017) reveals that the activity follows the population density which is shown in Figure 199.

Throughout the study we found that a large part of the dataset contained Tweets related to the keywords ‘grote’, ‘vuurbal’, ‘jonguh’ and ‘attack’. This can for example be seen in the topic modelling results shown in Table 16: LDA topic modelling results.6 (section 4.3.1). From the second subset onwards only topics have been identified related to the keywords mentioned above. A simple search in the database on the keywords returned 35,000 Tweets. These Tweets had become a worldwide trending topic on Twitter during the incident. The Tweets refer to a scene in a comedy action movie which was associated with the explosions and fireballs seen at the incident. This trending topic introduced a lot of noise in the dataset. Considering the fact that the free API only returns 50%-60% of the data from Twitter as a whole, valuable data may have not been present in the dataset.

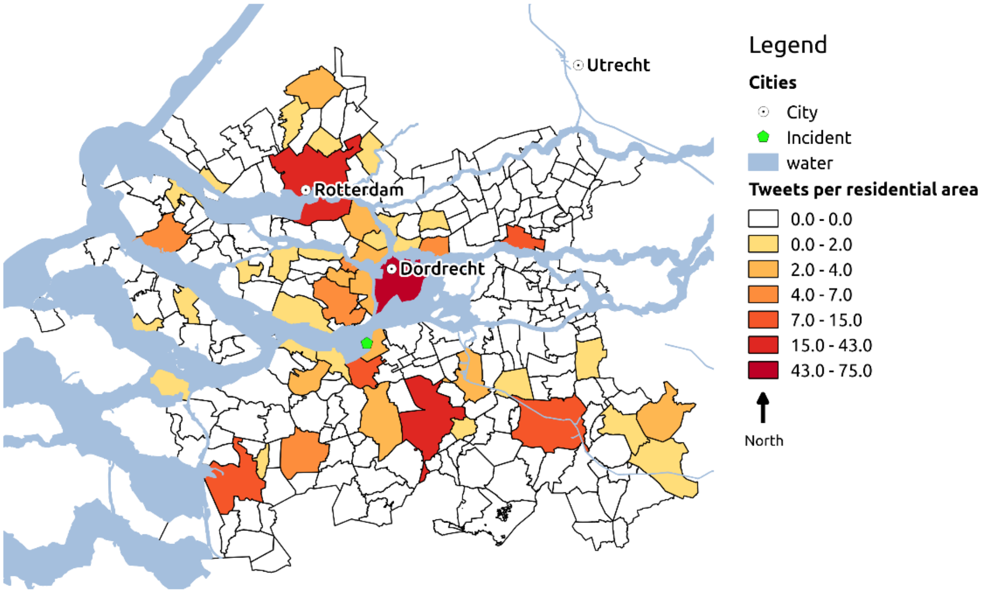


Figure 18: Spatial visualization of Tweets related to the toxic plume.

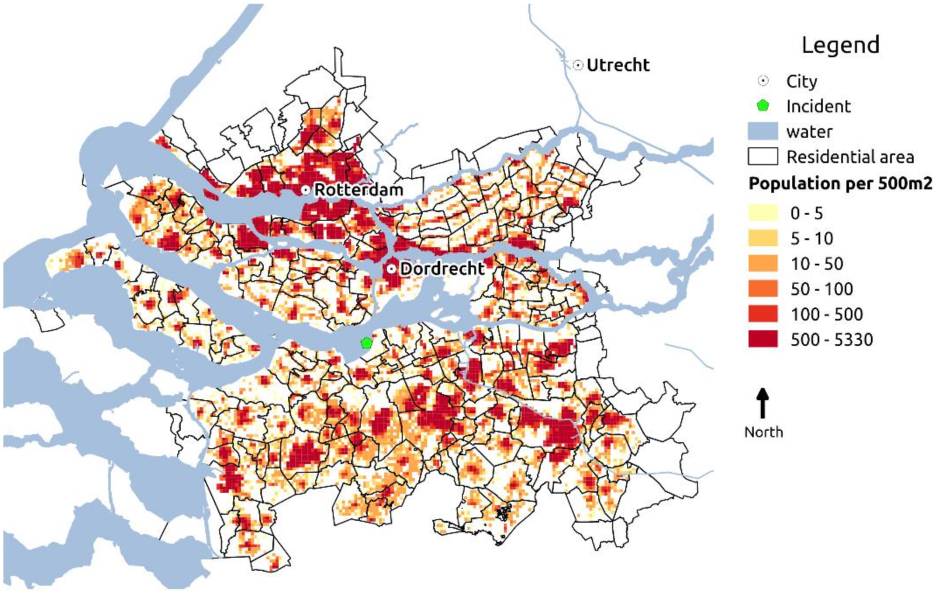


Figure 19: Population density per 500 square meters (CBS, 2017).

5.2.2 Spatial accuracy

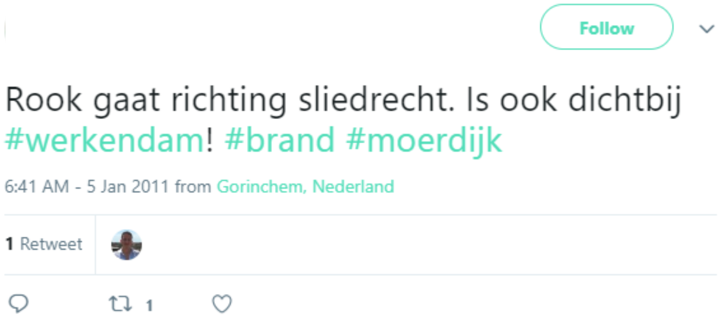
The initial dataset available for this study contained 1250 geotagged Tweets, which is approximately 1.2% of the entire dataset. We have extended this amount by retrieving additional meta-data from Twitter. This meta-data contained both profile locations of the users as well as places with which Tweets may have been associated. With this step we geotagged another 35,799 tweets. The majority of the Tweets has been geocoded based on the profile locations. 90% of these profile location had a city level granularity which resulted in many Tweets having the exact same location as these referred to the same city. For this reason it was only possibly to spatially visualize the Tweets related to the toxic plume on residential level.

The accuracy of profile locations is questionable for two reasons. First, a user may not be at the location where the Tweet has been posted. Second, even if the user posted from the profile location, the time difference between the moment the dataset was created (2011) and the moment the profile locations were retrieved (2017), may introduce an additional accuracy error. The users may have changed their profile location during this time period. Nevertheless, our method to validate the profile locations by comparing these with the location as mentioned in the content, revealed promising results. 33% of the geotagged Tweets using the profile locations matched the location mentioned in the Tweet. A similar method for validating profile locations has been applied by Burton et al. (2012). In this study they compared the profile location of geotagged Tweets to evaluate the accuracy of geotagging Tweets based on profile locations. They found that in 87% of the cases the GPS coordinates matched the profile location. With this method they were able to use 17.13% of the user profiles to geotag Tweets. A combination of both methods (comparing both GPS coordinates and location mentions in the Tweets with profile location) may further increase the number of geotagged Tweets and the spatial accuracy.

Although the number of Tweets available for the spatial analysis in this study was very limited, the manual annotation of the Tweets revealed that many Tweets were geotagged correctly. Locations that could not be validated seemed highly plausible considering the content. This may be explained based on two reasons. First, the incident occurred on a normal weekday in the late afternoon around dinner time. Typically many people are home during this time of the day. The probability of the profile location being correct is therefore relatively high if we assume the user has set the profile location to the home location. Second, content closer to the source is generally more related to the subject at hand than distant content (Herfort et al., 2014; Peters & Porto De Albuquerque, 2015; Imran et al., 2013).

Furthermore, the amount of meta-data for geocoding Tweets can be further increased by retrieving the meta-data timely when creating the dataset. In this study 36% of the profiles and 37% of the Tweets appeared to be deleted when retrieving this information from the Twitter API. As a consequence, profile locations and places could not be retrieved for these users and Tweets. These percentages should be zero if the meta-data is retrieved together with the Tweets.

The manual comparison of the profile locations with locations mentioned in the Tweet can be automated by using Named Entity Recognition (NER). With this method places and place names can automatically be extracted from text documents for geocoding purposes (Imran et al., 2013). In this study location based information was extracted form Tweets with a 91% precision. This method will also enable to geocode Tweets which do not have additional location information such as a profile location or a place name. However, the manual validation of profile locations revealed that various Tweets contain multiple locations as shown in Figure 20. This may introduce errors when Tweets are geotagged using named entity recognition.



Rook gaat richting sliedrecht. Is ook dichtbij #werkendam! #brand #moerdijk

6:41 AM - 5 Jan 2011 from Gorinchem, Nederland

1 Retweet

Figure 20: Example of Tweet with multiple location mentions.

We used Mapzen API to geocode the profile locations of the users. A major drawback of using unprocessed textual data is that it may be misspelled or incorrect. As a result, the geocoding services can return incorrect locations. This issue can be overcome by introducing an additional pre-processing step to clean and correct misspelled textual location data. A possible method is using the Levenshtein word distance that checks the number of required character edits between words. Based on this distance, words can be corrected. Ugon et al. (2015) used this method on 9818 entries of textual location data. They managed to correct 70.5% of the entries.

5.2.3 Classification of Tweet topics

We used unsupervised LDA topic modelling algorithm to identify topics in our dataset with no prior knowledge of the content. A short recap: the basic principal of the algorithm is grouping text documents (tweets in this study) to *k* number of topics based on the words in the documents. The algorithm assigns all words to the *k* number of topics with random probabilities. These probabilities denote how likely the words will appear together in the documents. The probabilities are calculated based on an iterative process during which the algorithm samples words from the topics and regenerates the input documents. By comparing word occurrences of the generated documents with the input documents, the probabilities can be recalculated. New sampling iterations follow until an optimum is reached. The result of this process is the definition of *k* number of topics containing all words of the input dataset, but with different probabilities per topic. The algorithm than assigns each text document to all topics with a different probability based on the words in the document and the word probabilities. This should be interpreted as a Tweet primarily being assigned to one topic (the topic with the highest probability), but with a relation to the other topics to some extent.

In our study we chose the value three for the number of topics parameter *k*. The reasoning behind this choice is that we would have a limited number of Tweets available for the algorithm as we subdivided the data based on two hour subsets. We also assumed no more than three topics will be discussed during each time period. We did not study the influence of this choice in our research and it is therefore unclear what the effect is. A recent study has defined a heuristic approach to estimate the optimal value for *k* based on analysis of variation of statistical perplexity and cross validation (W. Zhao et al., 2015). Ghosh & Guha (2013) applied the perplexity analysis to determine value for *k* in their study on identify and locate obesity health issues in the US based on Twitter data. The found three cohesive and clear topics with *k=50*. Based on these results it can be said that increasing the number of topics may have an effect on the output.

In our study we found that the differentiation between the three topics was very limited. The top six topic terms as summarized in Table 166 show many overlapping terms. Based on these terms it is not possible to identify clear and cohesive topics. This observation is further confirmed by the Tweet to topic probabilities performed by the algorithm. Ideally the difference between the highest probability and second highest probability should by relatively high. The closer the probabilities are, the more ambiguous the Tweet will be. An example of the Tweet to topic probabilities from our study is shown in Table 213. It can be seen that the majority of the probabilities are more or less equal. It can be interpreted as that a Tweet is equally related to each of the topics. Differentiation between topics and Tweets is therefore nearly impossible.

Table 21: Tweet to topic probabilities.

Tweet	topic1	topic2	topic3	classification
1	0.334724	0.332858	0.332418	1
2	0.336204	0.331901	0.331895	1
3	0.336204	0.331901	0.331895	1
4	0.333891	0.333321	0.332788	1
5	0.333089	0.33436	0.332551	2
6	0.33156	0.330462	0.337978	3
7	0.331259	0.33186	0.336881	3
8	0.33356	0.332298	0.334143	3
9	0.333012	0.337293	0.329695	2

10	0.336204	0.331901	0.331895	1
11	0.335959	0.33292	0.331121	1
12	0.336204	0.331901	0.331895	1
13	0.329904	0.330147	0.339949	3
14	0.332029	0.333348	0.334623	3
15	0.332239	0.33451	0.33325	2
…	…	…	…	…

Naturally, the next step, when using LDA Topic Modelling, is to pick a topic of interest and retrieve all text documents which have been assigned that topic. However, due to the lack of differentiation in Tweet to topic probabilities, it was not possibly to trust the topic assignment of the LDA algorithm. The limited performance of the LDA algorithm is caused by the limited number of words in the Tweets. Tweets can have a maximum length of 140 characters, but are often much shorter. As a result, context is missing to identify cohesive topics (W. X. Zhao et al., 2011). Selecting all Tweets that have been assigned to a topic, will result in a subset of the data rather than a selection of a theme. Manually searching for topic assignments with large differentiation is not possible due to the large amount of data.

To overcome this issue, we developed an additional analyzing step based on Self Organizing Maps. With this method we created heat maps on which Tweets with similar characteristics are grouped together. The characteristics are the three topic probabilities assigned to every Tweet. In other words, all Tweets with similar topic probabilities for all three topics have been grouped together. An example is shown in Figure 211. To understand these heat maps we need to consider the following aspects:

- Each circle is a node
- On each node, one or more Tweets are mapped with similar characteristics
- Each node (and therefore each Tweet) is mapped on exactly the same position on each map
- The color represents a value for the topic probability which can be looked up in the legend on the left of the maps (this is not an y-axis)
- Every map represents a different topic probability for the same Tweet

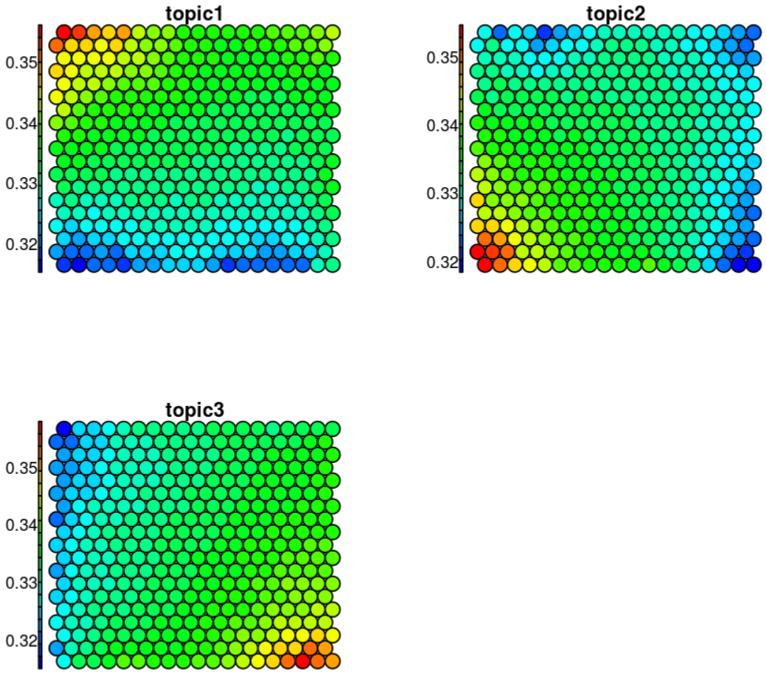


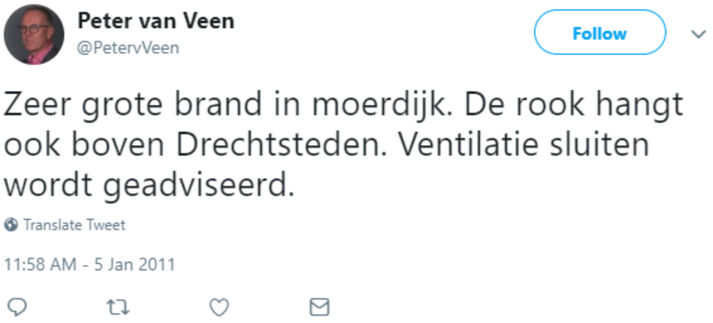
Figure 21: Example of the SOM heat maps of topic probabilities.

What we can observe is that indeed the majority of the Tweets lacks differentiation in topic probabilities. The green color (approximately 0.33) covers the majority of the maps. The Tweets that have been mapped on these nodes have a probability of 0.33 for each topic. We can also observe clusters which consist of Tweets with a relatively high topic probability (yellow) for one topic, and lower probabilities for the other topics (blue). These clusters can be interpreted as Tweets with a coherent topic. In our study we selected the Tweets from a cluster by manually picking a reference value. We have done this by comparing the color of the cluster with the value in the legend. This approach has a large influence of the results. The choice of reference value will ultimately decide how much Tweets are retrieved for further processing and the accuracy of the theme. For example, we chose our value mostly somewhere between the transition from yellow to red color. Depending on the hour of the day and the activity on Twitter, we retrieved somewhere between 400 and 1100 hundred Tweets using this method. For one topic (topic 2 in subset 2) we chose a value equal to the red color. This resulted in only 57 Tweets. This examples shows that choosing a lower probability will result in more Tweets, but potentially less coherent Tweets. Whereas a high reference value will retrieve less Tweets, but with higher coherence as the topic probability is higher. A random error may be introduce by manually choosing the reference value. A better approach would be to apply K-means clustering in the SOM heat maps to mathematically calculate the clusters (Riga, Stocker, Rönkkö, & Kolehmainen, 2015). Nevertheless, the selection of the Tweets based on the clusters in the SOM heat maps and the exploration of the content in these Tweets with word clouds, revealed promising result. Comparing the new word clouds with the topic terms of LDA topic modelling, we could identify more coherent themes.

Unfortunately, due to the many steps in this research a majority of the Tweets remained unused. The main cause is the limited performance of the LDA topic modelling on the Twitter data and the limited amount of spatial information. We were able to retrieve 452 Tweets from the topic modelling results. 417 (82%) of these Tweets had a location. After selecting only the geotagged Tweets within the study area, only 207 Tweets remained for spatial visualization of the toxic plume. Nevertheless, 59% of the Tweets contained actionable information for the emergency service. This is a relatively high percentage considering the use of unsupervised algorithms for classifications is the least accurate method (Hahmann et al., 2014).

5.2.4 Deriving actionable information needs

We derived actionable information from the selected topic manually. We assigned each Tweet to only one information type. We found a significant amount However, we found several Tweets which relate to more than one information type. An example is shown in Figure 222. This Tweets does not only refer to the location of the hazard, but also contains advice on what to do. It can be stated that the information categories are not complete. Valuable information may be missed due to this method. In future work the completeness of the information categories can be increased by assigning multiple information categories to one Tweet (Kiatpanont et al., 2017).



Zeer grote brand in moerdijk. De rook hangt ook boven Drechtsteden. Ventilatie sluiten wordt geadviseerd.

11:58 AM - 5 Jan 2011

Figure 22: Example of tweet with to information types.

Furthermore, we categorized the Tweets based on a theoretical framework of information types. We did not address the question how useful this information is. It is possible that the same information was available for the emergency services via other sources (Harrald & Jefferson, 2007). A better approach would be to validate the information categorization by professionals which have been involved with the incident. Li et al. (2014) applied a similar methodology to identify information needs of fire fighters to support situational awareness for building emergencies.

5.3 Recommendations

Throughout this study many opportunities for improving the proposed methods have been identified. In this section several of these opportunities are recommended for future work.

- Include word stemming and tokenization: the performance of the LDA topic modeling heavily depends on the pre-processing and text-mining steps. Word stemming can be applied to ensure morphological variations in words are eliminated. For example, the words ‘run’, ‘running’ and ‘runner’ are related and one can say that these belong to the same topic. However, LDA will discard the relation between these words and interpret them as three different words. Word stemming will convert the words to ‘run’ retaining the relation when LDA is applied. Additionally, tokenization can be applied to increase the context of the words.

- Use URL in Tweets to increase the contextual information of a Tweet: many Tweets contain URL’s to news articles or background information related to the topic mentioned in the Tweet. These articles can be used to increase the contextual information of the Tweet before applying LDA topic modelling. This will yield in better topic classification results.

- Explore LDA topic modelling results with semantic-preserving word clouds to explore the topics: in this study the differentiation of the topic terms as defined by the LDA algorithm was insufficient to interpreted topics. This issue can be overcome by the use of semantic-persevering word clouds. These word clouds do not only visualize the word counts, but also preserve the relation between words. These word clouds may reveal the meaning of a topic.

- Use Tweet pooling to improve the LDA Topic Modelling results: applying LDA Topic Modelling on the limited content in Tweets results in topics that are not coherent due to the lack of contextual information. This can be improved by using pooling schemes. The basic idea is that Tweets are aggregated according to a certain strategy before processing them with the LDA algorithm. Tweets can be aggregated by author or burst score (sudden increase in word frequencies).

- Use Labeled LDA with the crisisLex Lexicon: the topic discovery with LDA topic modeling can be further improved with labeled LDA. In this method, every topic is labeled with several terms before the algorithm is applied on the dataset. The text documents in which the labels are present are then more likely to be assigned to the labeled topic. The application of this method is of interest when there is prior knowledge on what topics need to be identified. In the field of disaster management, emergency services are aware of what information they need. These needs can be translated in labels by selecting terms from the crisisLex. This is a lexicon containing words that are frequently used in Twitter during the disasters.

- Use k-means clustering to automatically identify clusters in the heat maps of the SOM: the manually chosen reference value is inaccurate as it is visually not possible to estimate to optimal value of a cluster. With k-means clustering clusters in the heat maps can automatically be detected.

- Use named entity recognition to automatically detect locations in content: NER is the process of locating and classifying named entities in texts using predefined classes such as location, organizations, quantities etc. Tweets that have not been geotagged often contain mentions of locations in the content which can be detected with NER.

- Explore effect of increased maximum number of characters in a Tweet: Twitter has recently increased the maximum number of characters per Tweet. The maximum number of characters has been increased with 58% to 240 characters. Tweets are very messy as users tend to shorten words and use abbreviations to fit their message within the size limits. These grammatically incorrect content is human readable, but not machine readable. The increase of the maximum number of characters may result in less messy messages and improve both the text-mining and topic discovery results.

- Apply the LDA topic modelling algorithm only on the geocoded content: in our study the majority of the geocoded content remained unutilized since we applied the topic modelling algorithm on the entire dataset. The selection of the relevant topics contained many Tweets which have not been geocoded. By first geocoding the data and applying LDA topic modelling only on geocoded Tweets, one can ensure the final Tweet selection contains only geocoded Tweets.

References

Acar, A., & Muraki, Y. (2011). Twitter for crisis communication: lessons learned from Japan's tsunami disaster. *International Journal of Web Based Communities*, *7*(3), 392. http://doi.org/10.1504/IJWBC.2011.041206

Bilski, A. (2011). A review of artificial intelligence algorithms in document classification. *International Journal of Electronics and Telecommunications*, *57*(3), 263–270. http://doi.org/10.2478/v10177-011-0035-6

Blei, D. M., Edu, B. B., Ng, A. Y., Edu, A. S., Jordan, M. I., & Edu, J. B. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*(4), 993–1022. http://doi.org/10.1162/jmlr.2003.3.4-5.993

Bruns, A., & Burgess, J. E. (2011). #Ausvotes: How Twitter Covered the 2010 Australian Federal Election. *Communication, Politics and Culture*, *44*(July), 37–56.

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. *Journal of Medical Internet Research*, *14*(6). http://doi.org/10.2196/jmir.2121

CBS. (2017). Kaart van 500 meter bij 500 meter met statistieken. Retrieved from https://www.cbs.nl/nl-nl/dossier/nederland-regionaal/geografische data/kaart-van-500-meter-bij-500-meter-met-statistieken

Chaney, A., & Blei, D. (2012). Visualizing Topic Models. In *ICWSM 2012 - Proceedings of the 6th International AAAI Conference on Weblogs and Social Media* (pp. 419–422).

Croitoru, A., Crooks, A., Radzikowski, J., & Stefanidis, A. (2013). Geosocial gauge: a system prototype for knowledge discovery from social media. *International Journal of Geographical Information Science*, *27*(12), 2483–2508. http://doi.org/10.1080/13658816.2013.825724

Cutter, S. L. (2003). GI Science, Disasters, and Emergency Management. *Transactions in GIS*, *7*(4), 439–445. http://doi.org/doi:10.1111/1467-9671.00157

De Onderzoeksraad voor Veiligheid. (2011). *Brand bij Chemie-Pack , Moerdijk 5 januari 2011.*

Devillers, R. ., & Jeansoulin, R. . (2010). *Fundamentals of Spatial Data Quality*. Wiley-ISTE.

Ghosh, D., & Guha, R. (2013). What are we "tweeting" about obesity? Mapping tweets with topic modeling and Geographic Information System. *Cartography and Geographic Information Science*, *40*(2), 90–102. http://doi.org/10.1080/15230406.2013.776210

Goldberg, D. W. (2011). Advances in geocoding research and practice. *Transactions in*

GIS, *15*(6), 727–733. http://doi.org/10.1111/j.1467-9671.2011.01298.x

Goodchild, M. F. (2007). Citizens as sensors: the world of volunteered geography. *GeoJournal*, *69*(4), 211–221. http://doi.org/10.1007/s10708-007-9111-y

Grun, B., & Hornik, K. (2011). topicmodels : An R Package for Fitting Topic Models. *Journal of Statistical Software*, *40*(13), 1–30.

Haddi, E., Liu, X., & Shi, Y. (2013). The role of text pre-processing in sentiment analysis. *Procedia Computer Science*, *17*, 26–32. http://doi.org/10.1016/j.procs.2013.05.005

Hahmann, S., Purves, R., & Burghardt, D. (2014). Twitter location (sometimes) matters: Exploring the relationship between georeferenced tweet content and nearby feature classes. *Journal of Spatial Information Science*, *9*(2014), 1–36. http://doi.org/10.5311/JOSIS.2014.9.185

Haines, R., & Flatau, C. (1992). *Night Flying*. Tab Books.

Harrald, J., & Jefferson, T. (2007). Shared Situational Awareness in Emergency Management Mitigation and Response. In *Proceedings of the Annual Hawaii International Conference on System Sciences.*

Harrald, J. R. (2006). Agility and Discipline: Critical Success Factors for Disaster Response. *The ANNALS of the American Academy of Political and Social Science*, *604*(1), 256–272. http://doi.org/10.1177/0002716205285404

Herfort, B., de Albuquerque, J. P., Schelhorn, S.-J., & Zipf, A. (2014). Exploring the Geographical Relations Between Social Media and Flood Phenomena to Improve Situational Awareness A Study About the River Elbe Flood in June 2013. In *17th AGILE Conference on Geographic Information Science, AGILE 2014* (pp. 55–71).

Horita, F. E. A., Degrossi, L. C., Assis, L. F. F. G., Zipf, A., & De Albuquerque, J. (2013). The use of Volunteered Geographic Information and Crowdsourcing in Disaster Management: a Systematic Literature Review. *19th Americas Conference on Information Systems, AMCIS 2013 - Hyperconnected World: Anything, Anywhere, Anytime*, *5*(1), The use of Volunteered Geographic Information and.

Imran, M., Elbassuoni, S., Castillo, C., Diaz, F., & Meier, P. (2013). Extracting information nuggets from disaster- Related messages in social media. In *ISCRAM 2013 Conference Proceedings* (pp. 791–801). Baden-Baden, Germany.

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons*, *53*(1), 59–68. http://doi.org/10.1016/j.bushor.2009.09.003

Kiatpanont, R., Tanlamai, U., & Chongstitvatana, P. . (2017). Extraction of actionable information from crowdsourced disaster data. *Journal of Emergency Management*, *14*(6), 377–390. http://doi.org/10.5055/jem.2016.0302

Li, N., Yang, Z., Ghahramani, A., Becerik-Gerber, B., & Soibelman, L. (2014). Situational awareness for supporting building fire emergency response: Information needs, information sources, and implementation requirements. *Fire Safety Journal*, *63*, 17–28.

Longueville, B. De, & Smith, R. S. (2009). " OMG, from here , I can see the flames !": a use case of mining Location Based Social Networks to acquire spatio- temporal data on forest fires. In *GIS: Proceedings of the ACM International Symposium on Advances in Geographic Information Systems* (pp. 73–80). http://doi.org/10.1145/1629890.1629907

Mapzen. (2017). Mapzen Seach API. Retrieved from https://mapzen.com/documentation/search/

Mendoza, M., Poblete, B., & Castillo, C. (2010). Twitter Under Crisis: Can we trust what we RT? In *Proceedings of the 1st Workshop on Social Media Analytics* (pp. 71–79). http://doi.org/10.1145/1964858.1964869

Meyer, D., Hornik, K., & Feinerer, I. (2008). Text mining infrastructure in R. *Journal of Statistical Software*, *25*(5), 1–54.

Morstatter, F., Pfeffer, J., Liu, H., & Carley, K. M. (2013). Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose. In *Proceedings of the 7th International Conference on Weblogs and Social Media, ICWSM 2013* (pp. 400–408). http://doi.org/10.1007/978-3-319-05579-4_10

Olteanu, A., Vieweg, S., & Castillo, C. (2015). What to Expect When the Unexpected Happens: Social Media Communications Across Crises. *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing - CSCW '15*, 994–1009. http://doi.org/10.1145/2675133.2675242

Ostermann, F. O., & Spinsanti, L. (2011). A Conceptual Workflow For Automatically Assessing The Quality Of Volunteered Geographic Information For Crisis Management. *Agile 2011*, 1–6.

PDOK. (2017). Publieke Dienst Op de Kaart. Retrieved August 10, 2017, from https://www.pdok.nl

Peters, R., & Porto De Albuquerque, J. (2015). Investigating images as indicators for relevant social media messages in disaster management. In *ISCRAM 2015 Conference Proceedings - 12th International Conference on Information Systems for Crisis Response and Management.*

Ranter, H. (2011). Database met alle tweets over moerdijk. Retrieved from http://crisiswerkplaats.nl/blog/2011/01/13/database-met-alle-tweets-over-moerdijk/

Riga, M., Stocker, M., Rönkkö, M., & Kolehmainen, M. (2015). Atmospheric environment and quality of life information extraction from twitter with the use of self-organizing maps. *Journal of Environmental Informatics*, *26*(1), 27–40. http://doi.org/10.3808/jei.201500311

Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors. In *Proceedings of the 19th international conference on World wide web* (pp. 851–860). ACM.

Seppänen, H., Mäkelä, J., Luokkala, P., & Virrantaus, K. (2013). Developing shared situational awareness for emergency management. *Safety Science*, *55*, 1–9. http://doi.org/10.1016/j.ssci.2012.12.009

Seppänen, H., & Virrantaus, K. (2015). Shared situational awareness and information quality in disaster management. *Safety Science*, *77*, 112–122. http://doi.org/10.1016/j.ssci.2015.03.018

Squicciarini, A., Tapia, A., & Stehle, S. (2017). Sentiment analysis during Hurricane Sandy in emergency response. *International Journal of Disaster Risk Reduction*, *21*(December 2016), 213–222. http://doi.org/10.1016/j.ijdrr.2016.12.011

Starbird, K., Palen, L., Hughes, A. L., & Vieweg, S. (2010). Chatter on the red: what hazards threat reveals about the social life of microblogged information. In *CSCW '10 Proceedings of the 2010 ACM conference on Computer supported cooperative work* (pp. 241–250). http://doi.org/10.1145/1718918.1718965

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

Steiger, E., de Albuquerque, J. P., & Zipf, A. (2015). An Advanced Systematic Literature Review on Spatiotemporal Analyses of Twitter Data. *Transactions in GIS*, *19*(6), 809–834. http://doi.org/10.1111/tgis.12132

Takhteyev, Y., Gruzd, A., & Wellman, B. (2012). Geography of Twitter networks. *Social Networks*, *34*(1), 73–81.

Tapia, A. H., Moore, K. a, & Johnson, N. (2013). Beyond the Trustworthy Tweet: A Deeper Understanding of Microblogged Data Use by Disaster Response and Humanitarian Relief Organizations. *Proceedings of the 10th International ISCRAM Conference*, (May), 770–779. Retrieved from http://star-tides.net/sites/default/files/documents/files/Beyond the Trustworthy Tweet - A

Deeper Understanding of Microblogged Data Use by Disaster Response and Humanitarian Relief Organizations.pdf

Teevan, J., Ramage, D., & Morris, M. R. (2011). #TwitterSearch: A comparison of microblog search and web search. In *Proceedings of the 4th ACM International Conference on Web Search and Data Mining, WSDM 2011* (pp. 35–44).

Thurman, N. (2014). Real-time online reporting: Best practices for live blogging. In *Ethics for Digital Journalists: Emerging Best Practices* (pp. 103–114). Taylor and Francis. http://doi.org/10.1111/j.1471-0528.2004.00303.x

Tsou, M.-H., Jung, C.-T., Allen, C., Yang, J.-A., Han, S. Y., Spitzberg, B. H., & Dozier, J. (2017). Building a real-time geo-targeted event observation (Geo) viewer for disaster management and situation awareness. In *Lecture Notes in Geoinformation and Cartography* (pp. 85–98).

Twitter. (2015). Twitter for developers. Retrieved December 11, 2015, from https://dev.twitter.com

Twitter. (2017). Twitter Geographical Metadata. Retrieved from http://support.gnip.com/articles/geo-intro.html

Ugon, A., Nicolas, T., Richard, M., Guerin, P., Chansard, P., Demoor, C., & Toubiana, L. (2015). A new approach for cleansing geographical dataset using Levenshtein distance, prior knowledge and contextual information. In *26th Medical Informatics in Europe Conference* (Vol. 210, pp. 227–229). http://doi.org/10.3233/978-1-61499-512-8-227

Vieweg, S., Palen, L., Liu, S. B., Hughes, A. L., & Sutton, J. (2008). Collective intelligence in disaster: Examination of the phenomenon in the aftermath of the 2007 Virginia Tech Shooting. In *Proceedings of ISCRAM 2008 - 5th International Conference on Information Systems for Crisis Response and Management* (pp. 44–54).

Vijayarani, S., Ilamathi, M. J., & Nithya, M. (2015). Preprocessing Techniques for Text Mining - An Overview. *International Journal of Computer Science & Communication Networks*, *5*(1), 7–16. Retrieved from http://www.ijcscn.com/Documents/Volumes/vol5issue1/ijcscn2015050102.pdf

Vivacqua, A. S., & Borges, M. R. S. (2012). Taking advantage of collective knowledge in emergency response systems. *Journal of Network and Computer Applications*, *35*(1), 189–198. http://doi.org/10.1016/j.jnca.2011.03.002

Wehrens, R., & Buydens, L. M. C. (2007). Self- and super-organizing maps in R: The kohonen package. *Journal of Statistical Software*, *21*(5), 1–19. http://doi.org/10.18637/jss.v021.i05

Wu, Y., Provan, T., Wei, F., Liu, S., & Ma, K. L. (2011). Semantic-preserving Word Clouds by Seam Carving. *Computer Graphics Forum*, *30*(3), 741–750.

Zhang, M., Shen, F., Zhang, H., Xie, N., & Yang, W. (2015). Advances in Multimedia Information Processing -- PCM 2015. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *9315*, 447–455. http://doi.org/10.1007/978-3-319-24078-7

Zhao, W., Chen, J. J., Perkins, R., Liu, Z., Ge, W., Ding, Y., & Zou, W. (2015). A heuristic approach to determine an appropriate number of topics in topic modeling. *BMC Bioinformatics*, *16*(13), S8. http://doi.org/10.1186/1471-2105-16-S13-S8

Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E., Yan, H., & Li, X. (2011). Comparing Twitter and Traditional Media using Topic Models. *Proceedings of the 33rd European Conference on Advances in Information Retrieval (ECIR'11)*, *6611*, 338–349. http://doi.org/10.1007/978-3-642-20161-5_34

Appendix A: Topic modelling explained

LDA topic modelling

Topic models is an implementation of Latent Dirichlet Allocation in R, an unsupervised algorithm for automatic identification of themes in documents. The basic assumption is that each of the documents in a collection consist of a mixture of collection-wide topics. In practice words in documents are observed and not topics. Topics are identified by finding the hidden (*latent*) structure given the documents and the words in these documents.

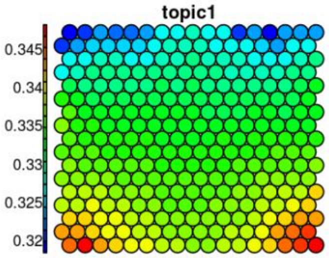
LDA identifies topics by applying an iterative algorithm that aims at recreating the documents by adjusting the importance of topics in documents and words in topics

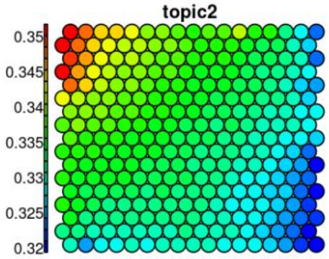
The algorithm requires manually setting the number of topics to be identified. It works as follows:

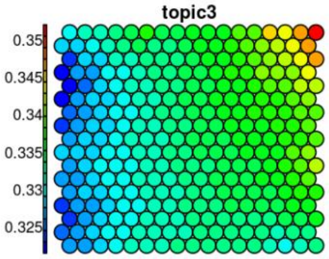
1. Run through each document and randomly assign each word in the document to a topic *t*
2. The assignment gives a topic representation of the documents and word distributions of all topics
3. Improve the topic representations and word distributions by
   a. Run through each word *w* in document *d*
   b. For each topic *t* calculate:
      i. Proportion of words in document *d* that are currently assigned to topic *t*
      ii. The proportion of assignments to topic *t* over all documents that come from word *w*
   c. Reassign *w* a new topic, where topic *t* is chosen with probability p(topic t | document d) * p(word w | topic t). This is the chance topic *t* generated word *w*. In this step it is assumed all topic assignments except for the current word are correct.
   d. Run step three many times until a stable state is reached
4. Use the topics assignments to estimate topic mixtures of each document.

Appendix B: Results of the Self Organizing Maps

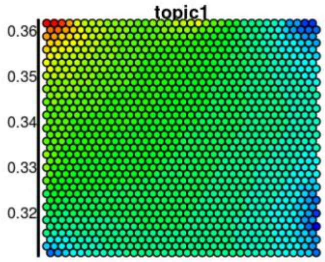
Subset: 1

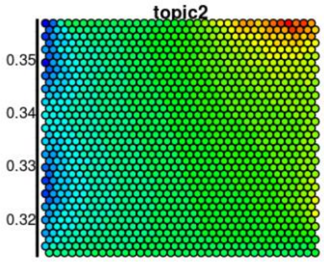


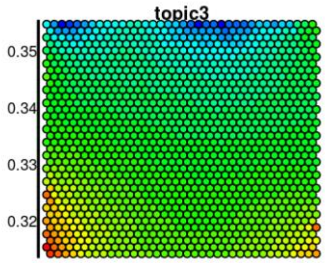




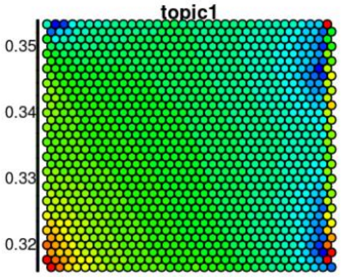
Subset: 2

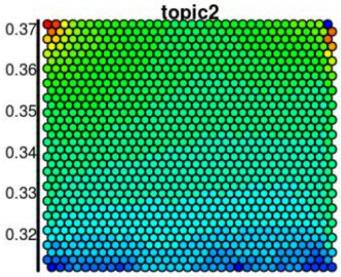


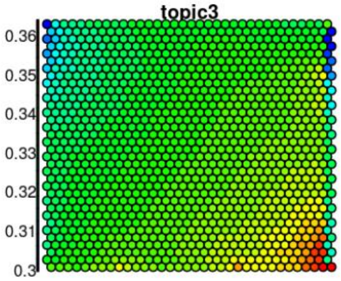




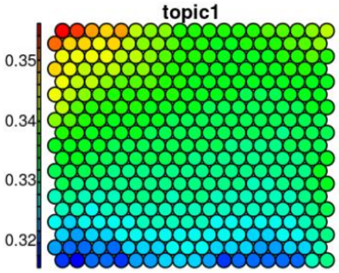
Subset: 3

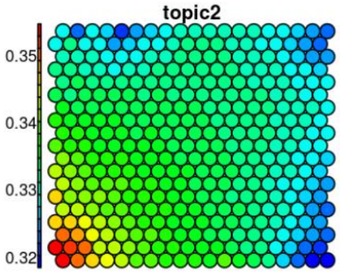


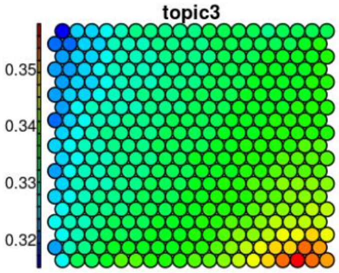




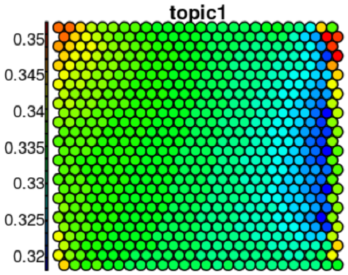
Subset: 4

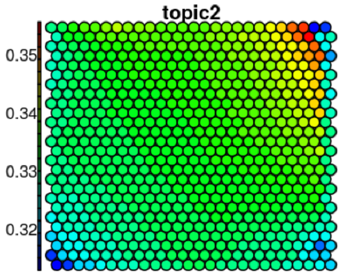


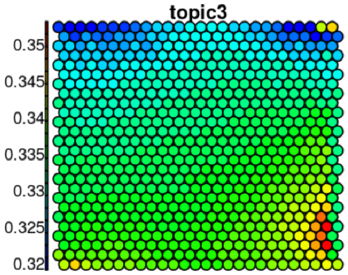




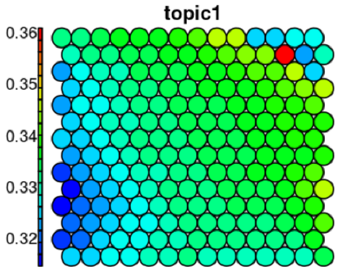
Subset: 5

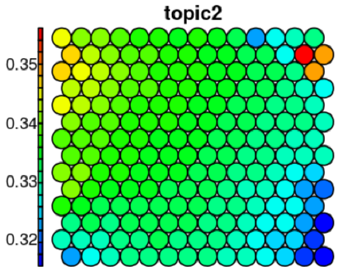


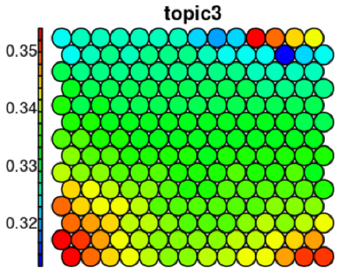




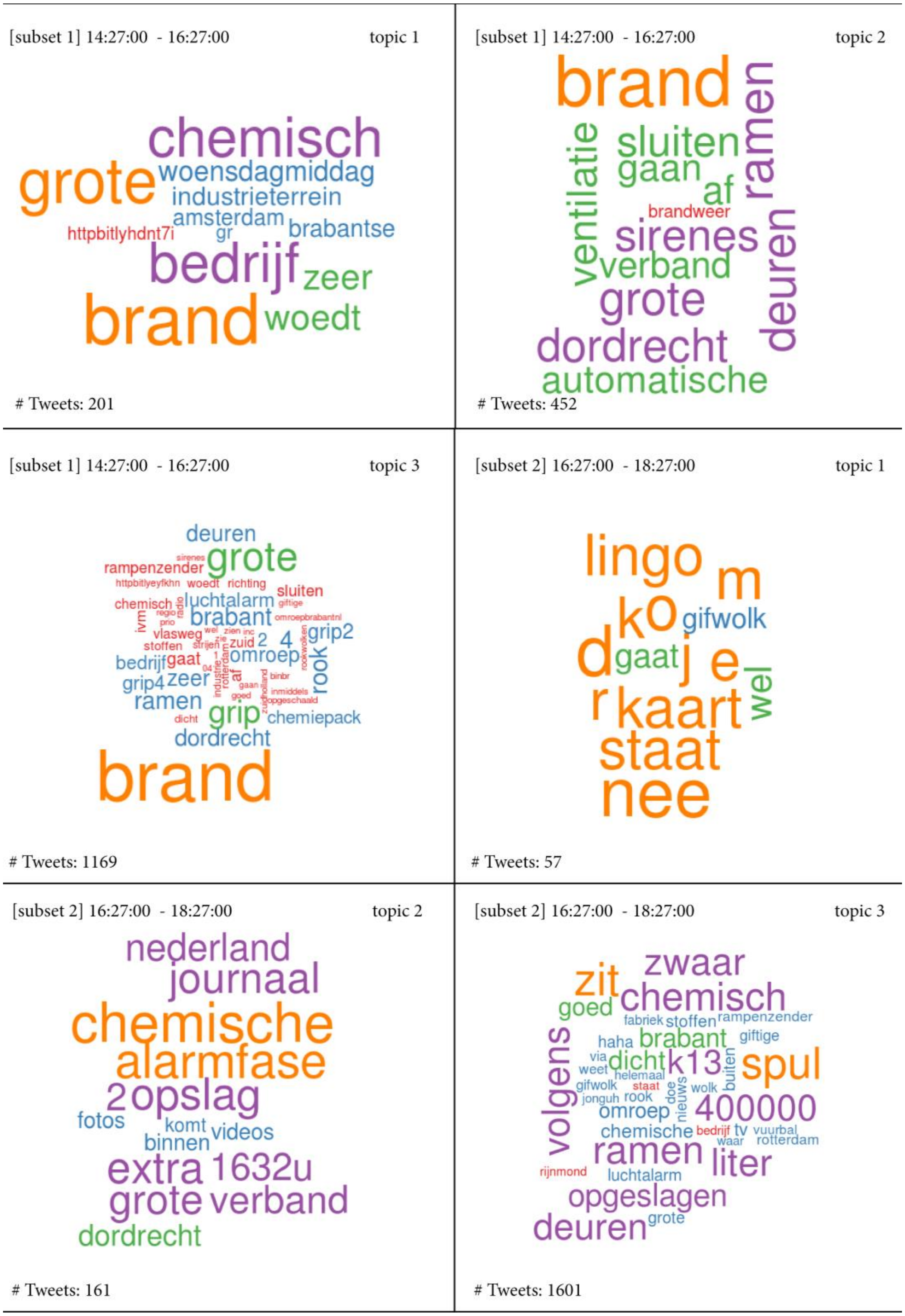
Subset: 6



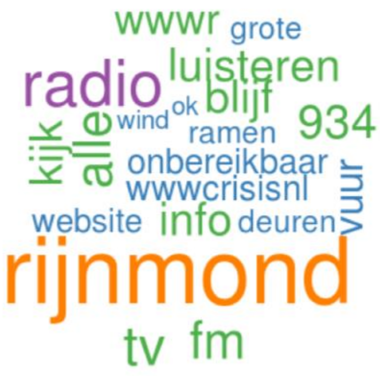




Appendix C: Results of the Tweet selection based on a cluster Reference Value

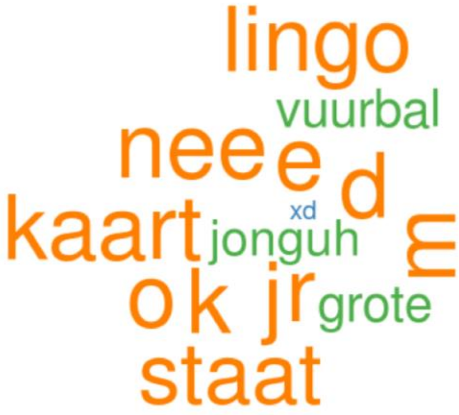


[subset 3] 18:27:00 - 20:27:00 topic 1



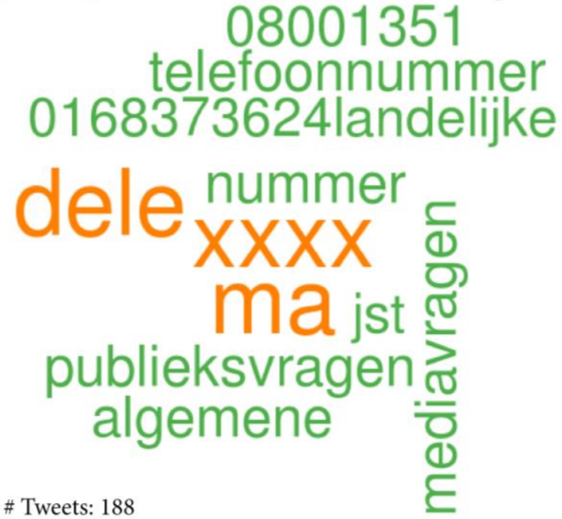
Tweets: 64

[subset 3] 18:27:00 - 20:27:00 topic 2



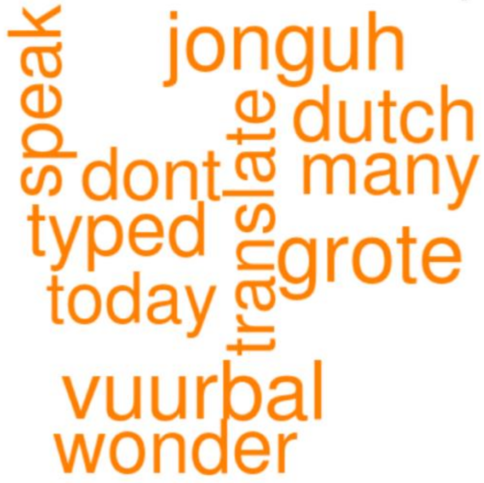
Tweets: 57

[subset 3] 18:27:00 - 20:27:00 topic 3



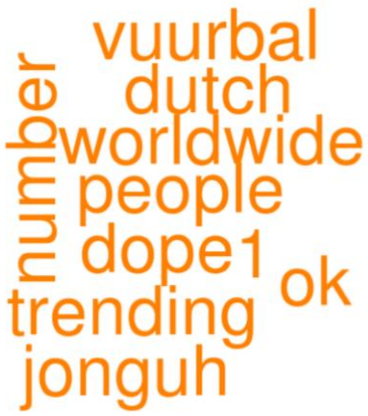
Tweets: 188

[subset 4] 20:27:00 - 22:27:00 topic 1



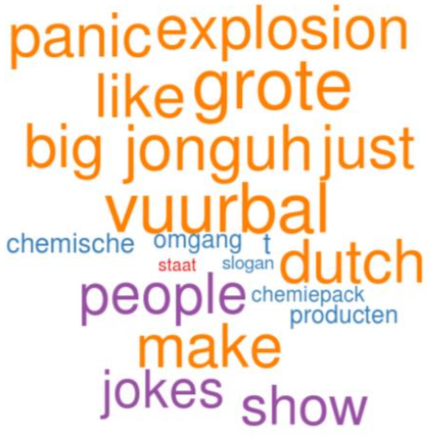
Tweets: 171

[subset 4] 20:27:00 - 22:27:00 topic 2



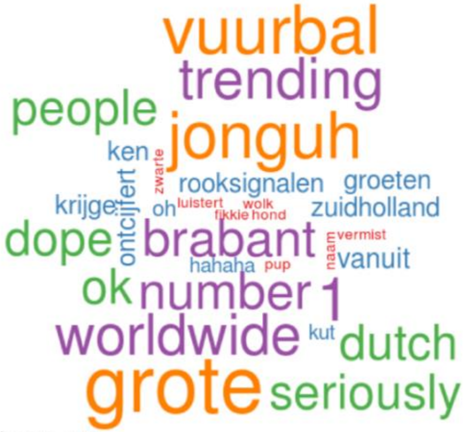
Tweets: 305

[subset 4] 20:27:00 - 22:27:00 topic 3



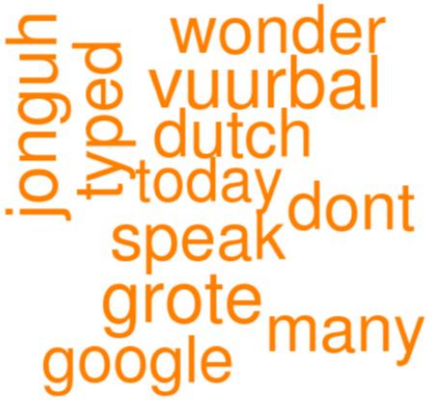
Tweets: 176

[subset 5] 22:27:00 - 00:27:00 topic 1



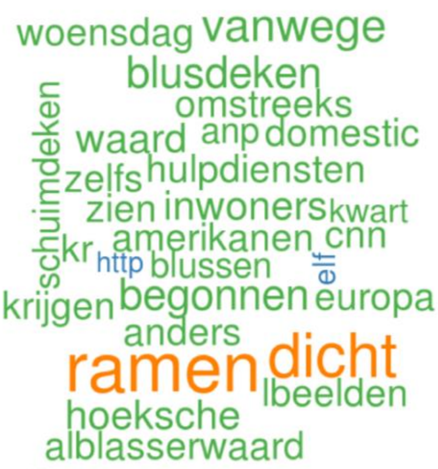
Tweets: 112

[subset 5] 22:27:00 - 00:27:00 topic 2



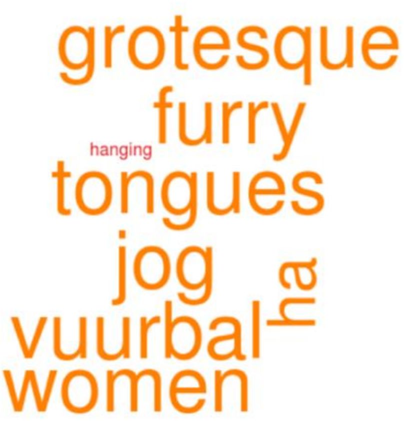
Tweets: 151

[subset 5] 22:27:00 - 00:27:00 topic 3



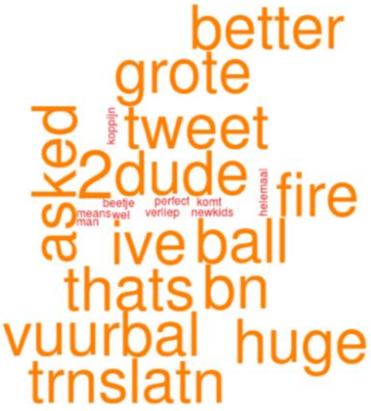
Tweets: 57

[subset 6] 00:27:00 - 02:27:00 topic 1



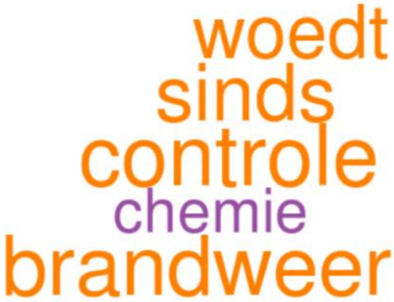
Tweets: 6

[subset 6] 00:27:00 - 02:27:00 topic 2



Tweets: 9

[subset 6] 00:27:00 - 02:27:00 topic 3



Tweets: 138