University of Bergen
Department of Information Science and
Media Studies

How can Big Data from Social Media be used in Emergency Management?

A case study of Twitter during the Paris attacks.

Author: Mariah Varum
Supervisor: Andreas L. Opdahl

June, 2018

Contents

List of Figures

Abstract

Over the past years, social media have impacted emergency management and disaster response in numerous ways. The access to live, continuous updates from the public brings new opportunities when it comes to detecing, coordinating and aiding in an emergency situation.

The thesis present a research of social media during an emergency situation. The goal of the study is to discover how data from social media can be used in emergency management and determine if existing analysis services can be proven useful for the same occasion. To achieve the goal, a dataset from Twitter during the Paris attacks 2015 was collected. The dataset was analyzed using three different analysis tools; IBM Watson Discovery service, Microsoft Azure Text Analytics and an own developed Keyword Frequency Script.

The results indicate that data from social media can be used for emergency management, in form of detecting and providing important information. Additional testing with larger datasets is needed to fully demonstrate the usefulness, in addition to interviews with emergency responders and social media users.

Acknowledgements

First, I would like to thank my thesis advisor Professor Andreas Lothe Opdahl, for sharing knowledge, providing guidance and feedback.

I will also like to thank my colleagues at the Norwegian Police ICT-Services, for inspiring me and giving me relevant experience and time to complete my masters degree.

Secondly, I would like to thank all my friends at the studyhall at the department of informatics at UiB, for providing insight and good academic discussions. My time as a student would not have been the same without you.

Huge thanks also goes to Henrik, my family and friends for their endless support and motivation.

Chapter 1

Introduction

This chapter presents a short introduction of the background of the project and the research problem. Furthermore, the chapter presents the research method, scope and limitations and target group, in addition to my personal motivation for this research.

1.1 Social Media in Emergency Situations

Today, around 40% of the world population has an internet connection. The number of users on the Internet has increased drastically since the 90s - the decade internet became accessible to "everyone". Today, there is over 3.9 billion internet users in the world (Stats, 2017).

In recent years, internet has largely consisted of "social media", a term embracing social networks, blogs, microblogs, forums, collaboration sites for creation and sharing information and documents, as well as the file sharing of audio, images and video. All the social media services hold huge amounts of data produced by their users and according to a survey done in 2017, we upload approximately 136,000 images every minute, publish 293,000 status updates and 510,000 comments on Facebook (Monappa, 2015). The availability of large amounts of data, often referred to as big data, has opened the possibility of improving our understanding of society and human behavior.

Social media has drastically changed the way people handle and perceive different situations that occur in daily life, especially when it comes to distressing events such as emergencies and disasters. An investigation shows that one in five people will try to contact each other via social media in an emergency or disaster, 35 percent will immediately post information on social media about the incident, while 58 percent will use social media to offer or assist with immediate relief (Hochmuth, 2015).

So far it is unclear how to get use of the full benefits from all the big data of information published on social media and similar technologies to increase the information level and spread of valuable information, such as warnings and real-time alerts. Social media, which has become a major part of our lives, can in the future get even greater space in our lives and in the daily service of society, as there are no other places or services where a such amount of people are gathered and reachable within a few seconds.

This research will present a case study of social media during an emergency event, with the purpose of examining how data from social media can be used during an emergency situation. The study includes data collection, preparation and data analysis, in addition to a discussion of how data could be used in an ongoing emergency situation, exemplified by the analysis.

The data from Twitter during the terrorism attack in Paris 2015 is selected as the data source for this study. The motivation for choosing the Paris attacks as the emergency situation in this thesis is that there is little scientific research done in English on the subject. Another reason for choosing the Paris attacks is the huge use of Twitter during the event. Twitter has been used in several other disasters, but this is one of the first times that it has been possible to follow the course of events on social media to such an extent, with detailed descriptions, images, audio and videos. (Twitter, 2015). Due to the lack of previous work on the Paris attacks, there are no gold standard for the data. In addition, this is an event happening in a French-speaking country, which means parts of the content are in French. I have chosen, for reasons related to limitations and understanding, to look at English tweets only.

There are several other studies on how social media have been used in emergencies, but mostly during natural disasters, such as floods, tsunamis and earthquakes. These events are less likely to happen here in Norway, and therefore I chose to study an emergency situation that we can relate and feel close to.

1.2 Research Background

This research is inspired by the research cooperation between Western Norway Research Institute, the University of Bergen and six other departments from USA, Hong Kong and Japan, which recently has established a research group for Big Data and Emergency Management (BDEM).

The utility and potential of big data for emergency management is growing but integration of big data into existing workflows and practices is far from seamless. To fulfill the potential benefits of big data for emergency management, the BDEM project will share best practices among the project partners in order to build and strengthen research and training that leverages big data and data analytics to transform emergency management for citizens and for society at- large (Lunde and Akerkar, 2017).

1.3 Research Questions

RQ1: How can big data from social media be used in emergency management?

The research question aims to explore how social media was used during an emergency situation, to identify the possibilities of using data from social media to detect, coordinate and respond.

Further, it will be interesting to discuss how early an ongoing emergency situation can be detected and if the results can provide important information for emergency management and rescuers? The possibility to detect, in addition to the availability of first-hand information is crucial in an emergency situation.

The reason for choosing this research question is my interest in exploring how the large amount of data from social media can be useful for society purposes, in this context, an emergency situation.

RQ2: How can existing analysis services be useful for emergency management?

The research question aims to investigate different analysis services to determine if they can be useful in emergency situations, to detect, coordinate and provide assistance. How early can an ongoing emergency situation be detected by an analysis tool and can the analysis tool provide functionality for rescuing and managing in the emergency situation.

Further interesting to discuss for the research question is which preparatory processes are needed for handling the data, in addition to which parts of the analysis are appropriate for social media analysis in context with emergency situations.

The reason for choosing this research question is to explore if already existing tools can be used for the purpose of handling large amounts of data from social media for emergency management.

1.4 Research Method

The research method in this project is a single exploratory case, aiming to answer the research questions of how big data from social media be used in emergency management. By using archival data to retrace an emergency event, this case study analyses, the event by collecting, storing and preprossesses the data. Further, a discussion of the analysis and the results will be presented.

1.5 Scope and Limitations

The research is regarding collection, preparation and analysis of an emergency situation. The research questions will be investigated and exemplified by a data analysis of the terrorism attacks in Paris, 2015.

Data Collection

Complete datasets from specific events are often very expensive and hard to get hold of. When the decision of which event and dataset to analyze had to be taken, it was based on the availability of a free dataset.

The dataset chosen for this analysis, The Paris attacks 2015, is published by Nick Ruest from The Scholars Portal Dataverse, which is a repository primarily for research data collected by researchers and organizations affiliated with Ontario universities (Ruest, 2017). The dataset was published with a CC BY 2.0 CA License (Commons, 2018). As Twitter's

Terms of Service (Twitter, 2017) does not allow full datasets of tweets to be distributed to third parties, the dataset included only tweet-ids (usernames) to post published during the event.

Due to the lack of previous work on the Paris attacks, there are no gold standard for the data. In addition, this is an event happening in a French-speaking country, which means parts of the content are in French. I have chosen, for reasons related to limitations and understanding, to look at English tweets only.

Analysis

This research is using the analysis services IBM Watson and Microsoft Azure. The reason for concentrating on the chosen applications are the functionalities and costs. Data analysis services are often expensive and requires a pricing plan over several months. In addition, the services require good knowledge in different programming languages and database queries, which make the learning curve for each application steep.

IBM Watson was chosen based on the cooperation agreement between University of Bergen and IBM. This agreement included a pricing plan, which made it possible to use the application for free. Unfortunately, this collaboration was not up and running before late spring 2018, which limited the possibilities of getting knowledge of and using the application. The reason for choosing Microsoft Azure as an analysis service was the Dreamspark cooperation agreement between the University of Bergen and Microsoft. Unfortunately, this collaboration ended in 2017 and didn't come up and running again before late February 2018, which limited the possibilities of learning and using the application.

Because of the master thesis limited time frame, I was unable to acquire more knowledge and expertise about other data analysis tools.

1.6 Target Group

The research is aimed at people interested in big data, data analysis and social media, in addition to emergency management.

The reader is not required to have any prior knowledge about the field, as the relevant topics are explained in the thesis.

1.7 Personal Motivation

The use of big data, social media and data analysis is a relevant phenomenon for my future career in information science. Big data has opened the possibility of improving our understanding of society and human behavior and have proven to be useful in several contexts. Social media has changed the way we interact with each other and it is very exciting to look at ways to utilize the great information flow.

The reason of choosing the area within emergency management is because of my interest in social science. I think it's exciting to see how the new era of information sharing in combination with technology can help us towards a better and safer society.

1.8 Outline

This thesis is structured into seven chapters. The following is an outline of each chapter.

Chapter 1: Introduction

This chapter presents a short introduction of the background of the project and the research problem. Furthermore, the chapter presents the research method, scope and limitations and target group, in addition to my personal motivation for this research.

Chapter 2: Theory

This chapter will present the theoretical topics that were relevant for the research, such as big data, social media, social media and emergency management and social media analysis. Further, the tools that were used and related work will be presented.

Chapter 3: Method

This chapter will present the methods used in this research concerning planning, design, data collection, analysis and results.

Chapter 4: Data analysis of Twitter

This chapter will present the preparatory work concerning the collection, storing and filtering. Further, the result of the data analysis will be presented.

Chapter 5: Discussion

This chapter will present a discussion of the research question, exemplified by the data analysis of the Paris attacks, in addition to the analysis services, techniques and results. Further, the different working methods and methodologies that were utilized in the research will be discussed.

Chapter 6: Conclusion and Further work

This chapter will present the conclusion of the research and the further work.

Chapter 2

Theory

This chapter will present the theoretical topics that were relevant for the research, such as Big Data, social media, social media and emergency management and social media analysis. Further, the tools that were used and related work will be presented.

2.1 Big Data

Big Data is an expression that has become a buzzword in recent years, and like many terms used to refer to the rapidly evolving use of technologies and practices, there are no agreed or industrial definition of Big Data (Kitchin, 2014). Below are two different definitions of Big data.

Big data: an accumulation of data that is too large and complex for processing by traditional database management tools (Webster, 2018c).

Big data refers to a process that is used when traditional data mining and handling techniques cannot uncover the insights and meaning of the underlying data. Data that is unstructured or time sensitive or simply very large cannot be processed by relational database engines. This type of data requires a different processing approach called big data, which uses massive parallelism on readily-available hardware (Techopedia, 2018).

Doug Laney's article from 2001 makes the most common reference to Big Data, the three V's: Volume, Variety, and Velocity (Laney, 2001). Volume refers to the magnitude of data, big data sizes can be reported in terabytes and petabytes. Variety refers to the structural heterogeneity in a dataset and Velocity refers to the rate at which data are generated and the speed at which it should be analyzed and acted upon (Gandomi and Haider, 2015). Kitchin on the other hand, argues that big data have seven essential characteristics: volume, velocity, variety, exhaustively, resolution/indexicality, relationality and flexibility/scalability that distinguish them from small data (Kitchin, 2014).

Big data can occur from several different sources, but the main sources are directly and automatically collected data, data from digital devices, volunteered collected data and open source data. Examples of main sources are public registrations, network monitoring, transactions and technology use, as well as social media for user-generated content (Kitchin, 2014).

2.2 Social Media

Social media is a phenomenon that has transformed the interaction and communication of individuals throughout the world. (Edosomwan et al., 2011) The term "social media" is embracing everything from social networks, blogs, microblogs, forums, collaboration sites for creation and sharing information and documents, as well as the file sharing of audio, images and video.

Social media refers to the means of interactions among people in which they create, share, exchange and comment contents among themselves in virtual communities and networks (Shahjahan and Chisty, 2014).

Figure 2.1 – The most popular social media's worldwide

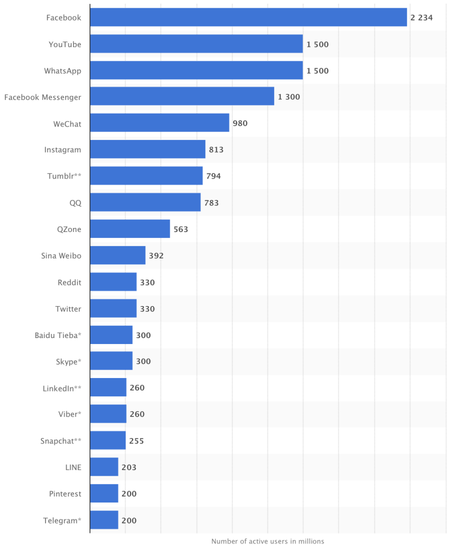


Figure 2.1 shows the most popular social media's worldwide as of April 2018. The leading social media, the social network Facebook, has over 2.2 billion users (Statista, 2017).

Social media has grown from being essentially a communication tool, to a platform for everyday life. *"While social media originally started out as a way to share information among friends, it is evident that it has evolved to serve other functions, such as a prevalent soruce for news, advertising and entertainment"* (Statement of Subcommitee Chariman Susan W. Brooks, 2013).

2.2.1 Twitter

Twitter is an online micro-blogginh service, where user interacts with messages, also called tweets. Microblogging is a form of blogging that allows users to send brief text updates or micromedia such as photographs or audio clips. An important common characteristic among microblogging services is its real-time nature. Twitter is frequently used during different events, where each event often has its own hashtag. Hashtags are used to clearly show that the content of a message is specifically related to an intended or established topic (Twitter, 2018d). Tweets can be published directly from computers, smartphones or mobile devices. Therefore, Twitter supports real-time information to large group of users, this makes Twitter an ideal tool to both access and spread information.

Microblogging is defined as "a form of blogging that lets you write brief text updates (usually less than 200 characters) about your life on the go and send them to friends and interested observers via text messaging, instant messaging (IM), email or the web." (Allen, 1983).

Twitter was launched in 2006 (Java et al., 2007), and currently has over 330 million unique users (Statista, 2017).

2.3 Social Media and Emergency Management

As social media has grown from an information sharing platform among friends and family, to a prevalent source for news, the platforms can provide important information about emergencies beyond the mass media. Emergency management deals with a wide range of events that are unexpected and may affect many people, for example natural disasters and intentional man-made events. Relevant to these events are the communication with the public. When unexpected events occur, there is high demand of information from the public that may be affected or are observing the event, in addition to the reporting mass media.

The subcommittee of Emergency Preparedness, Response, and Communications on Homeland Security has posted a statement on "How social media and New Tech are Transforming Preparedness, Response and Recovery" after a hearing in 2013 (Statement of Subcommitee Chariman Susan W. Brooks, 2013). *"We have heard numerous stories from Hurricane Sandy and the Boston Bombings of how citizens used Facebook, Twitter and Instagram to relay information to first responders, communicate with loved ones, and request assistance when cell phone services was unavailable.. We have also seen how response organizations are using social media to quickly share public safety information and maintain direct communication with disaster survivors during and after an incident."*

With the emergence of the Web 2.0, social media became a key platform that allowed people to interact and share information. Unlike traditional internet media, the Web 2.0 platform facilities not only the user's ability to access information; but also, their ability to comment on information already existing in the web sphere, and to publish or republish information. Over the last few years, users of social media have played an increasing role in the dissemination of emergency and disaster information (Kongthon et al., 2014). The information currency of disaster response is increasingly text messages, images, short videos, blog posts, and web links — all encapsulated knowledge chunks. Social media's strengths are in supporting ad-hoc network formation bringing together various players with different expertise and contexts, and providing some level of common ground between them (Yates and Paquette, 2011).

Already, social media has played an increasing role as a center for information related to emergencies and disasters, such as hurricanes (Muralidharan et al., 2011), earthquakes(Doan et al., 2012, Earle et al., 2010, Sakaki et al., 2010) and floods (Denis et al., 2014, Kongthon et al., 2014).

2.3.1 Emergency Management

Emergency management is a wide and large term that can be used in several contexts. In this research, the focus is on emergency management in conjunction with disasters, both natural and man-made.

Federal Emergency Management Agency (FEMA) is an agency within the US Department of Security which is responsible for coordinating disasters. FEMA has the following definition and principles for emergency management: (Fema, 2017).

Definition

Emergency management is the managerial function charged with creating the framework within which communities reduce vulnerability to hazards and cope with disasters.

Vision

Emergency management seeks to promote safer, less vulnerable communities with the capacity to cope with hazards and disasters.

Mission

Emergency management protects communities by coordinating and integrating all activities necessary to build, sustain, and improve the capability to mitigate against, prepare for, respond to, and recover from threatened or actual natural disasters, acts of terrorism, or other man-made disasters.

2.4 Social Media Analysis

In recent years, academic and public interest in the possibility of using social media to analyze data from e.g. public opinions, business value, growing trends and crisis communication increased. (Anstead and O'Loughlin, He et al., 2013, Muralidharan et al., 2011)

An analysis refers to breaking a whole into its separate components for individual examination. Data analysis is a process for obtaining raw data and converting it into information useful for decision-making by users (Tukey, 1992).

Data analytics services are provided by several large data companies. In this research, services from IBM Watson and Microsoft Azure have been used.

2.4.1 IBM Watson

IBM Watson is a cognitive computing technology with IBM (International Business Machines Corporation). The combination of the following capabilities makes it possible to move from reliance on structured, local data to unlock the world of global, unstructured data (High, 2012).

- **Natural language processing** helping to understand the complexities of unstructured data.

- **Hypothesis generation and evaluation** applying advanced analytics to weigh and evaluate a panel of responses based on only relevant evidence.

- **Dynamic Learning** helping to improve learning based on outcomes to get smarter with each iteration and interaction.

IBM Watson Discovery

IBM Watson Discovery Service is a cognitive search and content analytics engine, which can be added to applications to identify patterns and trends (Watson, 2018). IBM Watson Discovery have abilities to organize document and specific facts to identify correlations in data, locations and geospatial coordinates. The service works with both structured and unstructured data (IBM, 2018).

IBM Watson Discovery gives an insight of the following concepts from the enriched data:

- **Top Entities** Extracts people, companies, organizations, cities, and more. Containing the values of `entities.text`, `entities.type`, `entities.relevance`, `entities.count` and `entities.sentiment.score`.

- **General Sentiments** Identifies the overall positive or negative sentiment of each document. Containing the values of `sentiment.document.score` and `sentiment.document.label`

- **Related Concepts** Identifies general concepts that aren't necessarily referenced in the data. Containing the values `concepts.text`, `concepts.relevance` and `concepts.dbpedia.resource` (linking the concept to dbpedia.com)

- **Content hierarchy** Classifies the data into a hierarchy of categories up to 5 levels deep. Containing the values `categories.label` and `categories.score` (Range: 0.0-1.0).

IBM Watson Discovery also connects entities and concepts to other DBpedia, which is a crowd-sourced community effort to extract structured content from the information created in various Wikimedia projects (DBpedia, 2018).

2.4.2 Microsoft Azure

Microsoft Azure is a set of cloud services within Microsoft for building, managing, and deploying applications (Microsoft, 2018b).

Microsoft Azure Text Analytics

Microsoft Azure Text Analytics is a service within Microsoft, which provides APIs to detect languages, analyze sentiments, extract key phrases and identify linked entities. The Text Analytics API provides text analytics web services built with Microsoft machine learning algorithms (Microsoft, 2018a).

Microsoft Azure Text Analytics API provides four types of analysis (Microsoft, 2018e).

- **Language Detection** Detect which language the input text is written in and report a single language code for every document submitted on the request. The language code is paired with a score indicating the strength of the score.

- **Sentiment Analysis** This API returns a sentiment score between 0 and 1 for each document, where 1 is the most positive.

- **Key Phrase Extraction** Automatically extract key phrases to quickly identify the main points.

- **Entity linking** Identify well-known entities in the text and links it to more information on the web.

2.4.3 Keyword Frequecy Script

The Keyword Frequecy Script Keyword is an own developed script for keyword frequency. The script is programmed using JavaScript and Node.js, see section 2.5.4. The script reads the given input file, using a function to split each word and put the words into an array (a list), and then count each word that appears in the file. The predicates for the Keyword Frequency Script is $<$ *three letters long* and *occurring over 300 times.*

The reason for developing a Keyword Frequency Script is that keyword filtering has shown in several other cases to be a simple, but effective way to filter tweets for discovering the relevant topics (Doan et al., 2012, Lampos and Cristianini, OConnor and Smith, 2010).

2.5 Tools and Technologies

This section will present the different tools and technologies used in the research.

2.5.1 Hydrator

Hydrator is an open source application for hydrating datasets (the process of filling an object with data) of tweet-ids (Hydrator, 2018).

2.5.2 MongoDB

MongoDB is an document database that stores data in flexible JSON-like documents. It's free and open-source, published under the GNU Affero General Public License (MongoDB, 2018).

2.5.3 Trello

Trello is a tool for web-based project management, developed in 2011 by Fog Creek Software. The tool has several uses, such as real estate management and software project management. Trello uses the Kanban model for managing projects (Trello, 2018).

2.5.4 Technologies

JavaScript

JavaScript (often shortened to JS) is a lightweight, interpreted, object-oriented language with first-class functions. JavaScript can be used as scripting language for web pages and for developing web application (Mozilla, 2018).

Node.js

Node.js is an open-source, cross-platform JavaScript run-time environment that executes JavaScript code server-side. Node is designed to build scalable network applications (Node.js, 2018).

JavaScript Object Notation (JSON):

JSON is a syntax for storing and exchanging data, written with JavaScript object notation (W3Schools, 2018).

2.6 Social Media in Emergency Situations

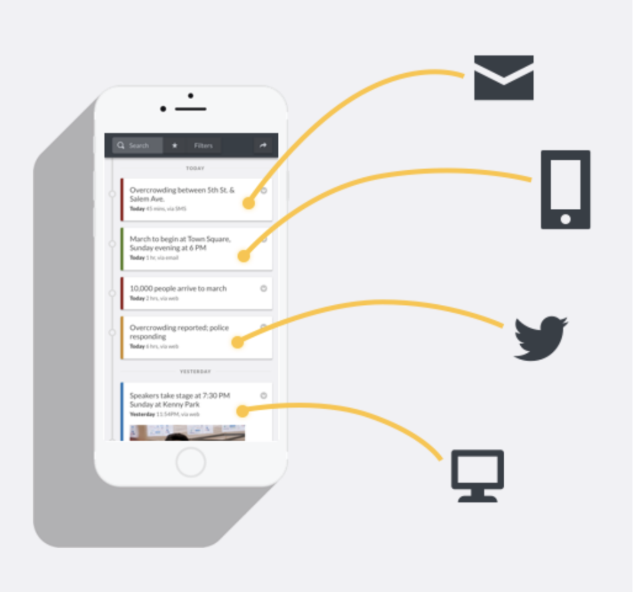
This section will present a selection of related work that illustrate the different possibilities related to social media, big data and emergency management.

The first section presents a type of solution, for the same purpose, but without using data harvesting and analysis. The last sections present theoretical studies in the same field as this research.

2.6.1 Ushahidi

Ushahidi is a non-profit company that creates open-source software for gathering information. Ushahidi was developed to map reports of violence in Kenya after the post-election violence in 2008 (Ushahidi, 2018).

Figure 2.2 – Visualization of Ushahidi



Overcrowding between 5th St. & Salem Ave.
Today 45 mins, via SMS

March to begin at Town Square, Sunday evening at 6 PM
Today 1 hr, via email

10,000 people arrive to march
Today 2 hrs, via web

Overcrowding reported; police responding
Today 6 hrs, via web

Speakers take stage at 7:30 PM Sunday at Kenny Park
Yesterday 11:54PM, via web

Ushahidi uses data from social media, offering products for Crisis Response, that collect reports from victims on the ground and field staff via SMS, email, web app, and Twitter (Ushahidi, 2018).

2.6.2 Thai Flood

The article "The Role of Social Media During a Natural Disaster: A Case Study of the 2011 Thai Flood" (Kongthon et al., 2014) presents a case study exploring how Thai people used social media such as Twitter to response to one of the country's worst disasters in recent history: The 2011 Thai Flood. This article gives a proof of the value of social media during an emergency.

The goal of the study was to determine whether analysis of the content of Twitter messages and characteristics of Twitter-users who tweeted during a crisis may yield information that can then be used to improve disaster preparedness and response. By analyzing user-generated messages it may be possible to assist local communities in obtaining up-to-date information; emergency rescuers in providing assistance according the needs of the populace in a timely manner or government agencies in analyzing and developing methods to use similar information to better centralize, coordinate, manage and plan disaster relief both during and after the event (Kongthon et al., 2014).

Findings: Since the flood reached part of the Bangkok Metropolitan area beginning in October 2011, the number of Twitter messages in Thailand increased significantly (Kongthon et al., 2014).

Figure 2.3 – Thailand Twitter Messages Year 2011

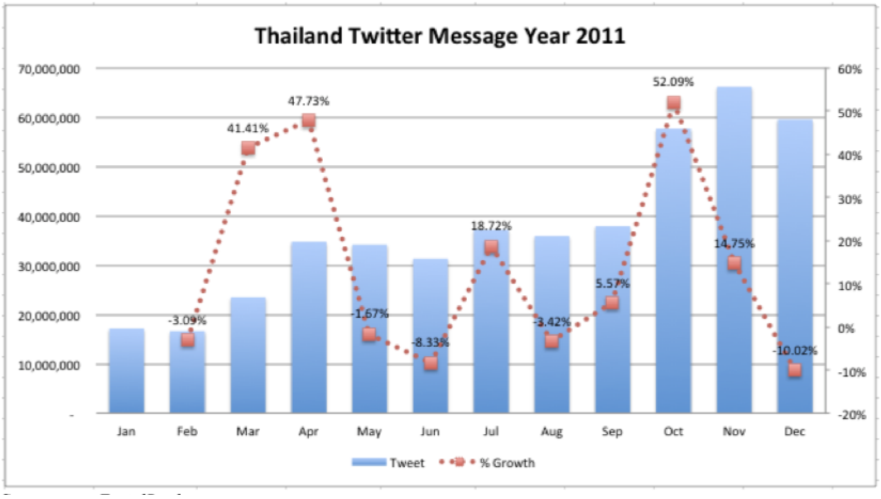


Figure 2.3 shows the number of Thai Twitter messages during the year 2011. From September to October 2011, the number of Twitter messages increased by 52%. The number of messages continued to grow until November 2011 where it reached the maximum. This may demonstrate that Thais were using Twitter to search for real time and practical information that traditional media could not provide during the natural disaster period (Kongthon et al., 2014).

To understand what type of information was disseminated in the Twitter network during the 2011 Thai Flood, 175 551 Tweets (from 23.10.2011 - 17.12.2011) using the keyword #thaiflood (Kongthon et al., 2014).

The Twitter messages were analyzed to determine what type of information was disseminated in the network. By using additional keyword analysis and rule based approach 64,852 tweets were automatically classified into 5 different categories: (Kongthon et al., 2014).
1. Situational announcements and alerts
2: Support announcements
3: Requests for assistance
4: Requests for Information
5: Other

Figure 2.4 – Thaiflood 2011, The Distribution of Five Tweet Categories

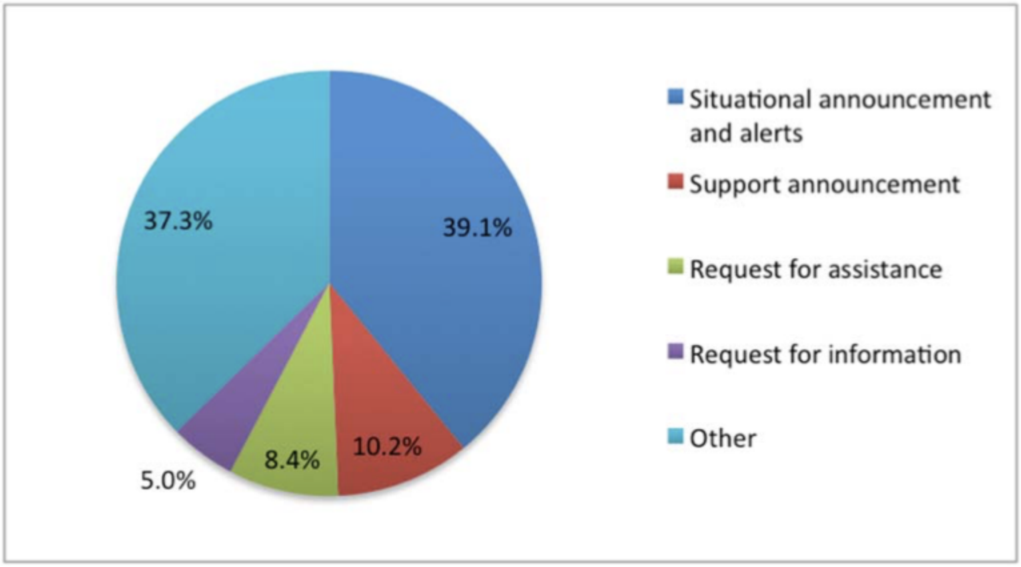


Figure 2.4 shows that the majority of the Tweets during the 2011 Thai Flood involved situational announcements and alerts.

By retrieving up-to-date information, related government agencies could use it in combination with requests for assistance information to provide help to citizens in a timely manner

For example, in one instance a Twitter user posted a message reporting the current water level in a certain area. By searching our results, we located another message posted by another Twitter user in the same area requesting medical supplies. With these two pieces of information, the flood relief agency could assess whether watercraft would be the most effective way to deliver the medical supplies to the people in need. Citizens could also monitor alerts and provide more detailed or accurate information to assist authorized agencies during an emergency incident (Kongthon et al., 2014).

During the several flooding situations that occurred in Thailand in 2011, social media such as Twitter has shown potential to be an effective tool for Thai citizens to obtain and disseminate up-to-the-minute information. With its real-time enabled platform, Twitter allows traditional journalists as well as citizen reporters to provide instant situation reports. The result can clearly be useful in coordinating resources and efforts in preparing and planning for disaster relief in the future. (Kongthon et al., 2014)

2.6.3 Tohoku Earthquake

The article "An Analysis of Twitter Messages in the 2011 Tohoku Earthquake" (Doan et al., 2012) presents an analysis of 1.5 million Twitter messages (tweets) from the period 09.03.2011 - 31.05.2011, in order to track awareness and anxiety levels in the Tokyo metropolitan district to the 2011 Tohoku Earthquake and subsequent tsunami and nuclear emergencies, happening between 11.03.2011 - 12.03.2011.

Within the stream of Twitter messages, three indicators of public response were studied:
1) Earthquake and tsunami
2) Radiation caused by the Fukushima Daiichi plant's meltdown
3) Public anxiety.
The first two types of indicators are aimed at showing people's awareness of the earthquake, tsunami and radiation and the last indicators looks at how people in Tokyo are anxious about these events (Doan et al., 2012).

Figure 2.5 – Tweet numbers by dates in English and in Japanese

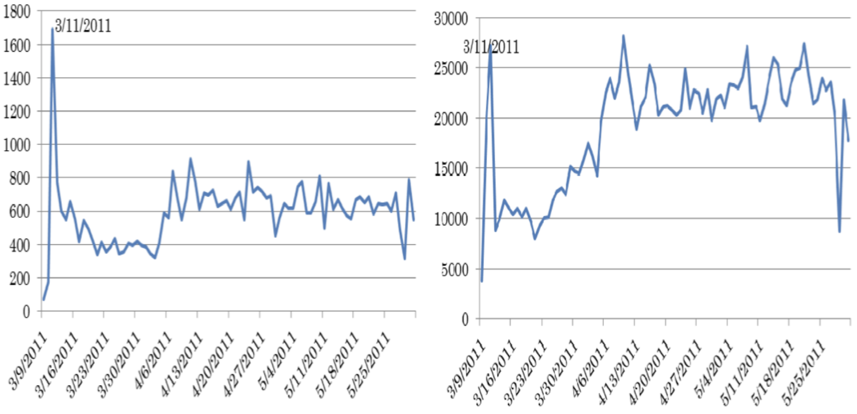
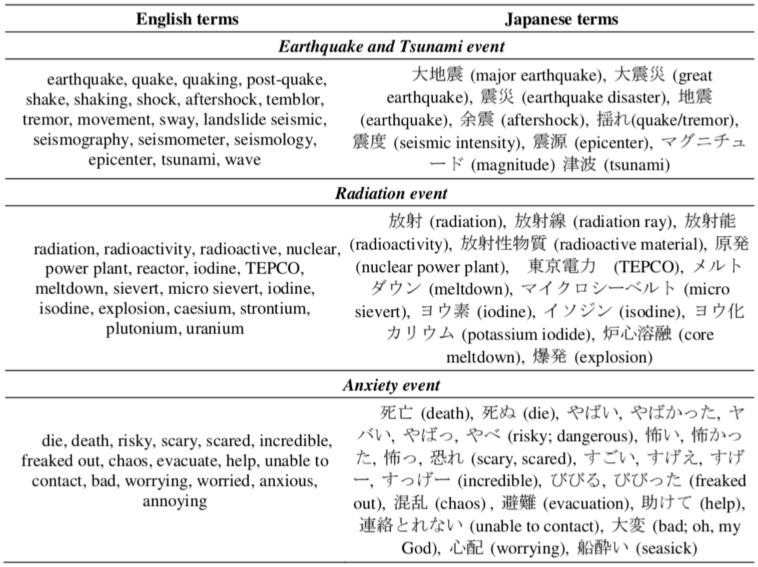


Figure 2.5 shows the number by dates during the event. The data indicates that Twitter users would like to broadcast their experience immediately (Doan et al., 2012).

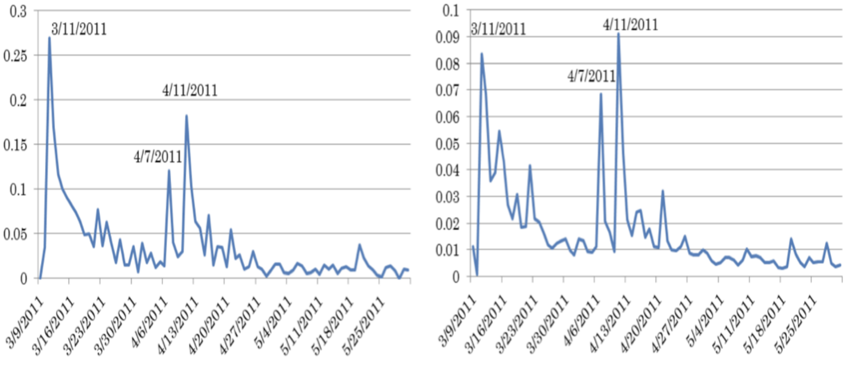
In the study, tweets were filtered by event keywords shown in figure 2.6.

Figure 2.6 – List of relevant keywords for the Earthquake and Tsunami, Radiation, and Anxiety events



English terms	Japanese terms
Earthquake and Tsunami event	
earthquake, quake, quaking, post-quake, shake, shaking, shock, aftershock, temblor, tremor, movement, sway, landslide seismic, seismography, seismometer, seismology, epicenter, tsunami, wave	大地震 (major earthquake), 大震災 (great earthquake), 震災 (earthquake disaster), 地震 (earthquake), 余震 (aftershock), 揺れ(quake/tremor), 震度 (seismic intensity), 震源 (epicenter), マグニチュード (magnitude) 津波 (tsunami)
Radiation event	
radiation, radioactivity, radioactive, nuclear, power plant, reactor, iodine, TEPCO, meltdown, sievert, micro sievert, iodine, isodine, explosion, caesium, strontium, plutonium, uranium	放射 (radiation), 放射線 (radiation ray), 放射能 (radioactivity), 放射性物質 (radioactive material), 原発 (nuclear power plant), 東京電力 (TEPCO), メルトダウン (meltdown), マイクロシーベルト (micro sievert), ヨウ素 (iodine), イソジン (isodine), ヨウ化カリウム (potassium iodide), 炉心溶融 (core meltdown), 爆発 (explosion)
Anxiety event	
die, death, risky, scary, scared, incredible, freaked out, chaos, evacuate, help, unable to contact, bad, worrying, worried, anxious, annoying	死亡 (death), 死ぬ (die), やばい, やばかった, ヤバい, やばっ, やべ (risky; dangerous), 怖い, 怖かった, 怖っ, 恐れ (scary, scared), すごい, すげえ, すげー, すっげー (incredible), びびる, びびった (freaked out), 混乱 (chaos) , 避難 (evacuation), 助けて (help), 連絡とれない (unable to contact), 大変 (bad; oh, my God), 心配 (worrying), 船酔い (seasick)

Figure 2.7 – Keyword frequencies for the earthquake event over time for English and Japanese tweets



The study shows that there is a sharp and sudden increase in the number of tweets immediately during the events. It is unknown when the first public report about the earthquake was in Tokyo, but the first tweet on the topic originating in Tokyo occurred at 05:48:08 UTC, 1 minute and 25 second right after the earthquake happened at the epicenter (Doan et al., 2012).

In the study, tweets were filtered by event keywords (see figure 2.6). As an example, the earthquake and tsunami keyword frequencies for both English and Japanese, are shown in figure 2.7.

The study has shown high correlations between aggregated tweets and disaster during the disaster. It appears that there is strong to potential for tracking both public information and anxiety in resident populations affected by the disaster. Furthermore, the study shows that Twitter data can be a useful resource in early warning surveillance systems as well as a tool for analyzing public anxiety and needs during times of disaster.

2.6.4 Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors

The article "Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors" presents an investigation of the real-time nature of Twitter and proposes an event notification system that monitors tweets and delivers notification promptly.

To detect target events and useful information, a classifier of tweets based on features such as the keywords in the tweet, the number of words and their context, applied by standard stop-words elimination and stemming. Subsequently, a probabilistic spatiotemporal model for targeting the event and find the center and the trajectory of the event location were produced. Twitter users were considered as sensors and location estimation methods such as Kalman filtering and particle filtering was used to estimate the locations of events (Sakaki et al., 2010).

As a proof of concept, an earthquake reporting system application was constructed. The earthquake reporting system, detecting earthquakes in Japan, proved high probability. 96% of earthquakes of Japan Meteorological Agency (JMA) seismic intensity scale 3 or more was detected merely by monitoring tweets. The reporting system detected earthquakes promptly and sends an e-mail to registered users. These notification is proven to delivered much faster than the announcements that are broadcast by the JMA (Sakaki et al., 2010).

Chapter 3

Research method

This chapter will present the method used in this research that concerns data gathering, analysis and results.

As presented in 1.4 this study is conducted as a case study.

3.1 Case Study

Case studies focus on one instance of something that is to be investigated. The case is comprehensively studied, typically using data generation methods such as interviews, observation, document analysis, and questionnaires (Oates, 2006).

In general, case studies are the preferred strategy when how or why questions are being posed, when the investigator has little control over events and when the focus is on a contemporary phenomenon within some real-life context (Yin, 2008).

Case studies aim to gather rich and detailed information about the specific case, its process, and relationships. By gathering rich data about the specific case, the researcher can explain how and why certain outcomes occur in given situations. This allows the researcher to retain a holistic and real-world perspective of specific case in both small group behaviour and on organizational level (Yin, 2008).

The reason for conducting a case study is due to the lack of a clear theory of the research.

The study is an explanatory case study, which will be further described in this chapter. The case study will attempt to answer the research questions presented in 1.3.

3.1.1 The Paris Attacks, 2015

For this single-case study, the terrorism attack in Paris 2015 was analyzed. The attacks were the deadliest in the European Union since the Madrid train bombings in 2005 (Jon Henley, 2015). The analysis was aimed at understanding the social media capabilities; how Twitter was used during the attacks and if data from social media can be used further. By doing this study, I could be able to identify how data from social media can be used in emergency management.

The Paris attacks, November 2015

Friday, November 13, 2015. A series of coordinated terrorist attacks occurred in Paris, France. Beginning at 21:30, three suicide bombers struck outside the Stade de France, during a football match. This was followed by several mass shootings and a suicide bombing, at cafés and restaurants. The terrorists took hostages and carried out another mass shooting at an Eagles of Death Metal concert in Bataclan Theatre. The attackers were shot or blew themselves up when the police raided the theatre. The attackers killed 130 people, including 89 at the Bataclan theatre. Another 413 people were injured, almost 100 seriously (BBC, 2015c, Hirsch, 2015).

As news of multiple attacks in Paris broke on Friday night, social media was the place where millions of people around the world first heard about it. Eyewitnesses logged onto their social networks to warn others about what was happening. It was an instinctive human reaction to tell others about the violence. Each word, image and video posted to sites like Twitter, Facebook and Instagram tell their own story (BBC, 2015b).

3.1.2 Designing a Case Study

There are four types of design for case studies: (Yin, 2008)

1. Single-case (holistic) design
2. Single-case (embedded) design
3. Multiple-case (holistic) design
4. Multiple-case (embedded) design

Figure 3.1 – Basic Types of Designs for Case Studies

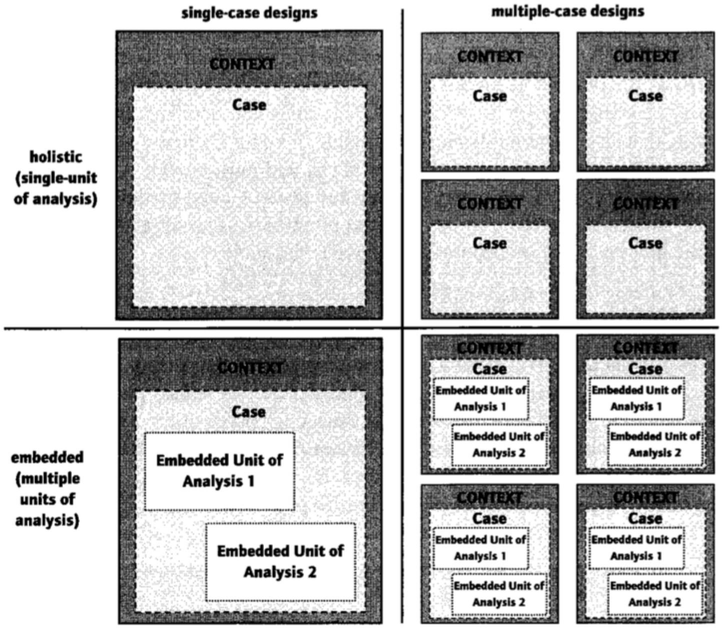

single-case designs
multiple-case designs
holistic (single-unit of analysis)
embedded (multiple units of analysis)
CONTEXT
Case
Embedded Unit of Analysis 1
Embedded Unit of Analysis 2

Figure 3.1 from (Yin, 2008) - COSMOS Corporation.

Holistic case study design The case study examined only the global nature of an organization or a program. The holistic design is advantageous when no logical subunits can be identified or when the relevant theory underlying the case study is itself of a holistic nature (Yin, 2008).

Embedded case study design The same study may involve more than one unit of analysis. This occurs when, within a single case, attention is also given to a subunit or subunits (Yin, 2008).

Single-case vs. Multiple-case design A single-case study is analogous to a single experiment, and many of the same conditions that justify a single experiment also justify a single-case study. On the other hand, multiple-case design should serve in a manner similar to multiple experiments, with similar results. Any use of a multiple-case design should follow a replication.

In addition, this research is an exploratory case study. “The exploratory case study investigates distinct phenomena characterized by a lack of detailed preliminary research, especially formulated hypotheses that can be tested, and/or by a specific research environment that limits the choice of methodology” (Streb, 2010). According to Yin, the exploratory case studies is used to define the necessary questions and hypotheses for developing consecutive studies (Yin, 2008).

Research design: A single exploratory case (holistic) design

This research is a single exploratory case study, with a holistic case study design. The study examines only the global nature of one event, including one analyze unit.

The exploratory case study is relevant for this thesis because of the limitations in terms of data access and the restrictive research environment in terms of the analyzed phenomenon.

According to Yin (Yin, 2008), single cases are common design for doing case studies and are eminently justifiable under certain conditions - where the case represents the following:
(a) a critical test of existing theory,
(b) a rare or unique circumstances,
(c) a representative or typical case,
(d) where a case serves a revelatory, or
(e) where a case serves a longitudinal purpose.

The research can be identified with several of the conditions above and are therefore justified as a case study.

A critical test of existing theory The theory has specified a clear set of proportions as well as the circumstances within which the proportions are believed to be true.

This research can be used to determine whether a theory’s proportions are correct or whether some alternative set of explanations might be more relevant, in this case, if it’s possible to use data from social media to detect, coordinate and provide information in emergency management.

A rare or a unique case The case represent an extreme case or a unique case.

The analyzed data in the case is rare, in terms of a unique event. The case on the other hand, analyzing data from social media, is not a rare or unique case.

A representative or typical case The objective is to capture the circumstances and conditions of an everyday or commonplace situation. The case study may represent a typical project among many different projects, a manufacturing firm believed to be typical of many other manufacturing firms in the same industry.

The research represents the circumstances and conditions of an everyday situation - human behaviour on social media. In addition, it is a project among many different projects on social media analysis. The lessons learned from this research is assumed to be informative about the experiences of the average person.

A revelatory case This situation exists when an investigator has an opportunity to observe and analyze a phenomenon previously inaccessible to social science inquiry.

The research is observing and analyzing a phenomenon which previously may have been inaccessible to social science inquiry, in terms of the data collection, harvesting and analysis.

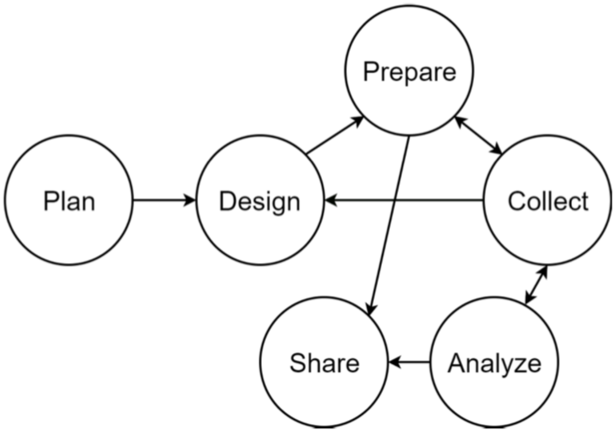
A longitudinal case Studying the same single case at two or more different points in time. The theory of interest would likely specify how certain conditions change over time, and the desired time intervals would presumably reflect the anticipated stages at which the changes should reveal themselves.

This research will not be studied at more different points in time and do therefore not meet this condition now, but nothing prevents it from happening at a later time.

3.2 Research Approach

The case study research has been done according to Yin's model (figure 3.2) of how to conduct the case study method (Yin, 2008).

Figure 3.2 – Yin's Case Study Method



3.2.1 Plan

The planning phase was spent identifying research questions and other rationales for doing the case study, in addition to plan for conducting the design, collecting data and the analysis.

At an early phase the decision to use the case study method compared to other methods was made. Case studies are the preferred method when *(a) how or why questions are being posed, (b) the researcher has little control over events, and (c) the focus is on contemporary phenomenon within a real-life context* (Yin, 2008).

The research has the following outcomes to the above points:
(a) RQ1: How can big data from social media be used in emergency management? RQ2: How can existing analysis services be useful for emergency management?
(b) The data gathered for this analysis is open data from social media during an emergency situation. The study is involving data from millions of people and there is no control over the data.
(c) A contemporary phenomenon, using Twitter as a communication tool in and for emergency management, in a real-life context.

3.2.2 Design

The research design is an action plan for getting from here to there, where here may be defined as the initial set of questions to be answered, and there is some set of conclusions (answers) about these questions. Another way of thinking about research design is as a blueprint of research, dealing with at least four problems: *what questions to study*, *what data are relevant*, *what data to collect* and *how to analyze the results* (Yin, 2008).

According to Yin(Yin, 2008), there are five components of a research design that are especially important:

1. a study's questions,
2. its propositions, if any,
3. its unit(s) of analysis,
4. the logic linking the data to the propositions; and
5. the criteria for interpreting the findings.

Study Questions The study questions provides an important clue regarding the most relevant research method to be used. To answer the study's question, one should choose a method that will answer the right questions, often in terms of who, what, where, how and why (Yin, 2008), such as the research questions described in the introduction, section 1.2.

Study Propositions Each proposition in the study directs attention to something that is to be examined within the scope of the study. However, why and how questions may not sufficiently point to what the study is about (Yin, 2008). The thesis proposes ideas and possibilities for how to use data from social media during emergency management.

Unit of analysis Defining what the actual case is by studying the questions and proportions to identify the relevant information to be collected. Yin points out the importance of determining the scope of the data collection and, in particular, how you will distinguish data about the subject of your case study (the phenomenon) from data external to the case (the context) (Yin, 2008).

The case examined in this research is a single-case study of social media during an emergency situation. The study examines only one event from social media during one specific emergency situation - the Twitter feed during the Paris attacks. Other social media or emergency situations are not included in the scope of the research. Furthermore, the research only addresses the people who used the specific hashtags chosen for this analysis.

Linking the data to the propositions The process of linking the data from the case is depending on how the data is linked to the propositions in the way of pattern matching, explanation building, time-series analysis, logic models and cross-case synthesis. According to Yin, the analysis requires the investigator to combine and assemble the data as a direct reflection of the study propositions (Yin, 2008).

The data collected underlines the study propositions in form of ideas and possibilities. As this is a big data research the amount of data analyzed may be deciding if this is a full-fledged analysis.

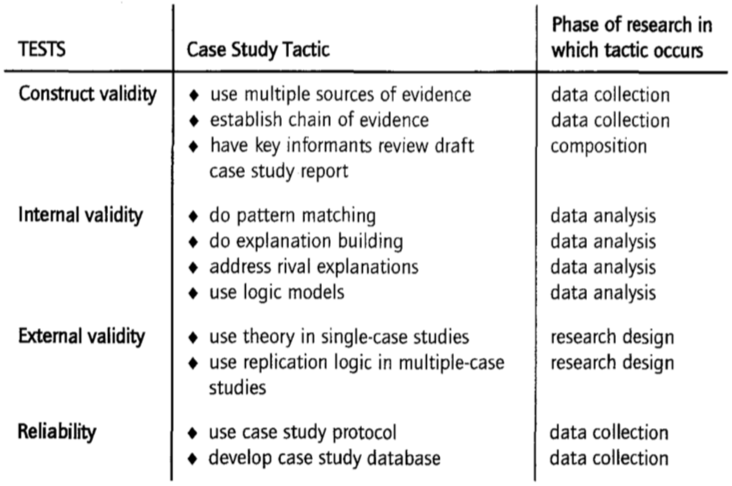
Criteria for interpreting the findings A major and important alternative strategy is to identify and address rival explanations for your findings. At the design stage, the challenge is to anticipate and enumerate the important rivals, so the investigator will include information about them as a part of the data collection. It is important to think of this before the data collection has been completed, so it becomes a part of the study's results, and not a part of a further study (Yin, 2008).

In the initial phases of this research, the plan and design were more flexible. This allowed me to investigate and explore areas that were not part of the initial research problem, but seemed interesting to examine.

Quality of Research Design

A research design is supposed to represent a logical set of statements and therefore, the quality of any given design can be judged according to certain logical tests. Four tests have been commonly used to establish the quality of any empirical social research: (Yin, 2008).
(a) construct validity,
(b) internal validity
(c) external validity
(d) reliability.

Figure 3.3 – Yin's Case Study Tactics for Four Design Tests



TESTS	Case Study Tactic	Phase of research in which tactic occurs
Construct validity	♦ use multiple sources of evidence ♦ establish chain of evidence ♦ have key informants review draft case study report	data collection data collection composition
Internal validity	♦ do pattern matching ♦ do explanation building ♦ address rival explanations ♦ use logic models	data analysis data analysis data analysis data analysis
External validity	♦ use theory in single-case studies ♦ use replication logic in multiple-case studies	research design research design
Reliability	♦ use case study protocol ♦ develop case study database	data collection data collection

Figure 3.3 from (Yin, 2008).

Construct Validity refers to identifying the correct operational measures for the concepts being studied. The researcher needs to cover two steps, in order to meet the test of construct validity (Yin, 2008).

1. Define neighbourhood change in terms of specific concepts (and relate them to the original objectives of the study)
2. Identify operational measures that match the concepts (preferably by citing published studies that make the same matches)

Internal Validity is mainly a concern for explanatory case studies, when an investigator is trying to explain how and why event x led to event y. If the investigator incorrectly concludes that there is a casual relationship between x and y without knowing that some third factor z may actually have caused y, the research design has failed to deal with some threat to internal validity. Second, the concern over internal validity extends to the problem of making inferences. A case study involves with an inference every time an event cannot be directly observed. The research design needs to consider rival explanations, and analyze the evidence's convergence and degree of truth in order to explain the accuracy of the interference (Yin, 2008).

External Validity deals with defining the domian to which a study's findings are generalizable beyond the immediate case study. Critics typically state that single cases offer a poor basis for generalizing. According to Yin, a theory must be tested by replicating the findings in a second or even a third case, where the theory has specified that the same results should occur. Once such direct replications have been made, the results might be accepted as providing strong support for that theory (Yin, 2008).

Reliability demonstrating that the operations of a study - such as the data collection procedures - can be repeated, with the same results. This means, if a later investigator would follow the same strategy, and conduct the same study, the later investigator should arrive at the same findings and conclusions. According to Yin, one prerequisite for allowing this other investigator to repeat an earlier case study is the need to document the procedures followed in the earlier case (Yin, 2008).

The quality of the study's research design will be discussed in section 5.4.1.

3.2.3 Prepare

The preparation of a case study is a complex task which considers challenge such as, gaining approval for the study. The following steps should be included in a formal part of any case study preparation (Yin, 2008).

1. Desired skills
2. Training for a specific case
3. Develop a protocol for the case study
4. Screening candidate cases
5. Conduct a pilot case study

Desired skills Case study research is a demanding task that requires a large set of skills. Yin presents a list of commonly required skills representing a good investigator: *be able to ask good questions*, *be a good listener*, *be adaptive and flexible*, *have a firm grasp of the issues being studied*, and *be unbiased by preconceived notions*.

The skills required to conduct the study are representing the knowledge I have obtained through several years of study - in conclusion with a master's degree. In preparation to this study, reading the book *Case Study Research: Design and Methods Applied Social Research*

Methods by Yin Roberts, (Yin, 2008), was helpful to obtain insight and knowledge of the case study as a research method.

Training for a specific case The goal of training for a specific case is to understand: 1. Why the study is being done, 2. What evidence is being sought, 3. What variations can be anticipated (and what should be done if such variations occur), and 4. What would constitute supportive or contrary evidence for any given proposition (Yin, 2008).

By using analysis services, I have had the opportunity to perform several tests before the final analysis, which has given me lots of training and learning for the purpose. Furthermore, the plan was to do a big data analysis, but due to the limitations of technological tool, the dataset had to be minimized. By using well-developed computer software, in form of analysis services, there are few variations occurring.

Develop a protocol for the case study A case study protocol is a document describing the case which is to be studied and should include the following sections:

1. An overview of the case study project (project objectives, case study issues, and relevant readings about the topic being investigated)
2. Field procedures (presentation of credentials, access to the case study sites, language pertaining to the protection of human subjects, sources of data, and procedural reminders)
3. Case study questions (the specific questions that the case study investigator must keep in mind collection data, table shells for specific arrays for data, and the potential sources of information answering these questions)
4. A guide for the case study report (outline, format for the data, use and presentation of other documents, and bibliographical information) (Yin, 2008).

The case study protocol used for this research was a project which was a part of a subject in the master's degree. The project was gradually transferred and rephrased into this thesis.

Screening candidate cases The goal of the screening procedure is to be sure that all the final cases are identified, prior to the formal data collection (Yin, 2008).

In the exploratory initial phase, lots of observation and research gave good indications of which type of event and dataset that would fit and was desired in this research. Further, the screening procedure included searching for a suitable, available and free dataset from this specific type of event.

Conduct a pilot case study A pilot case study is a test of the case study, which helps the investigator refine the data collection plans with respect to both the conducted data and the procedures to be followed (Yin, 2008).

A pilot case study was not conducted because of the limited time frame available for this research. On the other hand, this research concerns a data analysis where the dataset is stored in such a way that it allows changes along the way. In addition, the analysis services used makes it possible to do both analysis and testing several times before the final case study.

3.2.4 Collect

The data was collected from Scholars Portal Dataverse (Ruest, 2017) and published with a CC BY 2.0 CA License (Commons, 2018). The dataset contains user-ids (also called tweet-id) to posts published on Twitter during the terrorist attack in Paris 2015, with the following hashtags: #Paris, #Bataclan, #Parisattacks and #Porteuverte.

The data was collected in September 2017.

3.2.5 Analyze

There is no recipe for analyzing data derived for case studies. The analysis of case study evidence is one of the least developed aspects of doing case studies (Yin, 2008). Yin suggests four general strategies for analyzing a case study:

- Relying on theoretical proportions
- Working the data from the ground up
- Developing a case with description
- Examining plausible rival explanations

The study proportions formed the case study and helped lay the theoretical basis for the case study analysis. According to Yin, this strategy reflects the case's research questions, reviews of the literature, and new hypothesis or propositions (Yin, 2008).

Working with the data from the ground up is an inductive strategy which can occur by playing with the data. This strategy is useful for investigating and exploring the data. According to Yin, this strategy can be the start of an analytic path, leading the researcher towards possible unexplored areas of the data (Yin, 2008).

Developing a case description aims to organize the case study according to some descriptive framework.

Examining plausible rival explanations is a strategy that can be combined with the three previously mentioned strategies. The typical hypothesis is an evaluation that the observed outcomes are the result of a planned intervention. The simple or direct rival explanation would be that observed outcomes were in fact the result of some other influence besides the planned intervention and that the investment of resources into the intervention may not actually have been needed (Yin, 2008).

The data analysis in this research is a proof of concept of an ongoing emergency, exemplified by the data from Twitter during the Paris attacks in 2015.

The Process of Data Analysis

In a data analysis, there are several phases that can be distinguished. The phases are iterative, in that feedback from later phases may result in additional work in earlier phases (Schutt and O'Neil, 2013). The phases used in this data analysis will now be described.

Data Requirements Data is necessary as inputs to the analysis. Which data is specified based upon the requirements of those directing the analysis (Schutt and O'Neil, 2013).

Data Collection The data is collected from the choosen source(s) (Schutt and O'Neil, 2013).

Data Processing Data initially obtained must be processed or organised before the analysis (Schutt and O'Neil, 2013).

Data Cleaning Once the data is processed and organised, the data may be incomplete, contain duplicates, or contain errors. Data cleaning is the process of preventing and correcting these errors. (Schutt and O'Neil, 2013).

Exploratory Data Analysis Once the data is cleaned, it can be analyzed. Several techniques can be used to understand the messages in the data. The process of exploration may result in additional data cleaning or additional requests for data, so these activities may be iterative (Schutt and O'Neil, 2013).

Iterations

The analysis was split up into iterations, as both the collected data and the analysis services required a preprocessing phase, see section 4.4. The web-based project management tool Trello was used for the used to visualize the work flow (Trello, 2018).

3.2.6 Share

The case study is written in a linear-analytic structure, meaning the sequence of subtopics starts with the issue being studied, followed by literature, methods, data analysis, discussion and findings, ending with a conclusion (Yin, 2008).

Potential Audience

Case studies have more potential audiences than other types of research. Each audience has a different need, and no report will satisfy all audiences to the full extent (Yin, 2008). This thesis is divided into several sections, with a different degree of explanations and theoretical levels, to serve the different audiences.

3.3 Research Ethics

The research in this thesis is including open data from Twitter.

Twitter is public and all tweets are immediately viewable and searchable. Twitter's privacy policy declares that the responsibility for tweets and other information provided on

Twitter lies with the user (Twitter, 2018c). Twitter has both a Developer Policy and a Developer Agreement, which have been used as guidelines in this (Twitter, 2018a).

The data presented in the thesis is seperated from the Twitter user accounts to such an extent that no persons can be identified by the information presented in the thesis. The separation process was done in parallel with the structuring of the data.

Chapter 4

Data analysis of Twitter

This chapter will present the preparatory work concerning the collection, storing and filtering, in addition to the result of the data analysis.

The data analysis in this research is a proof of concept of an ongoing emergency, exemplified by the data from Twitter during the Paris attacks in 2015.

4.1 Preparation

Dataset

The collected data, see section 3.2.4, contained tweet-IDs to posts published on Twitter during the terrorist attack in Paris 2015. Twitter's Terms of Service (Twitter, 2017) does not allow full datasets of tweets to be distributed to third parties.

If you provide Content to third parties, including downloadable datasets of Content or an API that returns Content, you will only distribute or allow download of Tweet IDs and/or User IDs (Twitter, 2017).

As the dataset contained raw data obtained directly from Twitter, the gathered tweets were written in many different languages. In this research, it is decided to focus on English tweets only because of the language limitations in the services and the possibility of understanding the analysis results.

Services

The services used in this analysis, IBM Watson Discovery and Microsoft Azure Text Analytics are services which requires a paying user account. The accounts used in this analysis is provided by the University of Bergen and is a IBM Watson Lite account and Microsoft Azure Student account.

Overview of IBM Cloud Lite

- Up to 2,000 concurrent documents per month.
- 200 MB.
- Up to 2 Collections.
- Up to 1 Custom Model.
- 500 Element Classification pages per month
- 500 query expansions with 1,000 total terms

In addition, IBM Watson Discovery requires one language only to be specified prior to the analysis. **Overview of Microsoft Azure Student**

- Maximum size of a single document: 5,000 characters
- 1 MB
- Maximum number of documents in a request: 1000 documents

In addition, Microsoft Azure only supports a selection of languages (Microsoft, 2018d).

4.2 Collection

A hydration tool named Hydrator was used to collect the full database, see section 2.5.1. Hydration means getting the complete details (i.e. fields) of a tweet, using the status/lookup REST API call. Twitter limits users to 900 API requests every 15 minutes (360,000 tweets / hour) (Twitter, 2018b).

```
 900 requests * 100 tweets = 90 000 tweets/15 minutes
                           = 360 000 tweets/hour
```

There are a few limitations using a hydration tool: (Twitter, 2018b).

- The order of Tweet IDs may not match the order of Tweets in the returned array
- You must be following a protected user to be able to see their most recent Tweets. If you don't follow a protected user their status will be removed.
- If a requested Tweet is unknown or deleted, then that Tweet will not be returned in the results list, unless the map parameter is set to true, in which case it will be returned with a value of null

The data collection was done in the period 29. September 2017 - 10. October 2017. The full hydrated data collection resulted in the file `parisattacks.json`, which is a 61GB file of line-oriented JSON.

The dataset contained approximately 15 million tweet-IDs, but due to data loss, the hydrated version included in total 10,854,988 tweets.

There are other solutions and tools to do this process, e.g Twarc (GitHub, 2018). Hydrator was chosen based on the recommendation of the dataset owner, Nick Ruest.

4.3 Storage

MongoDB, described in section 2.5.2, was used to store the dataset.

As the dataset contained around 80 attributes for each tweet, see appendix A.1, it was necessary to structure the datasets in a database so the correct information could be retrieved using queries.

The reason for choosing MongoDB for the database among many other alternatives is first and foremost previous experience of using the tool and furthermore, the MongoDB database is free and simple to use.

The database was structured using queries to make a new database, containing only the field: text, which contains the published post with the text from the tweet.

```
mongod --dbpath=[path to where you want to store the database]
```

Start the database and `chose path` for storing the database.

```
mongoimport --db parisattacks --collection tweets --file parisattacks.json
```

Importing the `parisattacks.json` to MongoDB:

```
  1. mongo

  2. use paris-attacks

  3. db.tweets.find({})

  4. db.tweets.find({}, {"text": 1, "_id": 0}).pretty()
```

1: Starting the database

2: Chose which database to use

3: Find `all fields` for all tweets in the `collection: tweets`

4: Find the `field: text` only in the `collection: tweets`. Excluding the `field: _id`, which is by default added by MongoDB. Using `pretty()` to display the results in a formatted way.

```
mongoexport --db paris-attacks --collection tweets --out text-paris-attacks.json -f "text"
```

Saving the queried database (containing only the `field: text`) to a new file: `text-paris-attacks.JSON`

```
db.tweets.find({}, {"text": 1, "_id": 0}).limit(10000).pretty();
```

As a result of the capacity limitations in the services, described in 4.1, a new file with 10,000 tweets, `10000tweets.json`, were made from the original `text-paris-attacks.json`.

4.4 Preprocessing

Prior to the analysis, it was necessary to make several changes to the data set, in form of filtering languages and handling file formats, in addition to preparing the analysis services for receiving the data.

A script was made prior to the analysis to handle both the issue with several languages in the dataset and the different input requirements for the services.

Language filtering

The Microsoft Azure Text Analytics service *language detection*, further described in 2.4.2, was used to detect which languages occurred the most in the dataset. The analysis was done by reading the file `10000tweets.json`.

```
{
"documents": [
  {
    "id": "1",
    "detectedLanguages": [
      {
        "name": "French",
        "iso6391Name": "fr",
        "score": 0.2161289930343628
      }
    ]
  },
  {
       "id": "1",
       "detectedLanguages": [
          {
             "name": "English",
             "iso6391Name": "en",
             "score": 0.783871007
          }
       ]
    },
],
  "errors": []
}
```

This language detection analysis shows that the main languages in the data collection are English and French. Due to the Microsoft Azure Text Analytics limitations, see section 4.1, only 5000 characters with text from the tweets are analyzed, therefore, there is a high probability that other languages occur in the dataset as well.

Input Requirements

IBM Watson Discovery
IBM Watson Discovery provides an intuitive interface and requires retrieving the data from files, one by one.

Microsoft Azure Text Analytics
Microsoft Azure Text Analytics provides only an API for the text analytics services, which means there is no interface or GUI to interact with. The API only offer input reading by manually writing content into the API. Because of this, a file reading method had to be implemented.

Microsoft Azure supports several different programming languages and in this analysis the programming technologies JavaScript and Node.js, see section 2.5.4, were used to do necessary changes to the API and for handling the process of reading data and printing the results.

Microsoft Azure requires to retrieve the data by one file of maximum 5000 characters or 1000 different document files.

Keyword Frequency Script
The Keyword Frequency script reads the data as one file.

Script

The script is written in JavaScript, see section 2.5.4. For the complete script, see appendix, section A.2.

The language filtering is based on a stop-word list, constructed to filter the top 20 most used words in the French language: être, avior, je, de, ne, pas, le, la, tu, vous, il, et, à, un, l,́ qui, aller, les, en, ça (Daily, 2018). A stop-word list filter has proven to be effective in several other cases (Li et al., 2001, Sakaki et al., 2010, Yao and Ze-wen, 2011).

Description of the script
1. Reads the file
2. Use a filtering method based on a stop-word list, to filter the top words in French.
3. *IBM Watson Discovery:* Separate each tweet, to one tweet per file, from `0-N.json`
3: *Microsoft Azure Text Analytics:* Separate each tweet, to one tweet per file, from `0-N.json`. Save the filtered tweets to one file, `filteredTweets.json`
3: *Keyword Frequency:* Save the filtered tweets to one file, `filteredTweets.json`

Other services, such as both IBM Watson and Microsoft Azure, provides services for language filtering or translation, but because of the limitations of services in the accounts, the priority for these services was to analyze the dataset. In addition, there was no need for a language translation, since the amount of English tweets already exceeded the limitations concerning the number of tweets possible to analyze.

4.5 IBM Watson Discovery

IBM Watson Discovery is an application in the IBM Watson technology, described in section 2.4.1.

The IBM Watson Discovery service have some requirements for handling the dataset.

- **One language only**
  Before creating a new data collection for analysis, the language of the documents must be specified. If the document doesn't fit the specified language, an error occurs.
- **Retrieving files one by one**
  IBM Watson Discovery's functionality requires retrieving files with one by one element in the application.

4.5.1 Analysis

The dataset was analyzed using the IBM Watson Discovery. 1,836 documents were analyzed. The current analysis was created on 15.04.18.

The analysis is divided into four categories, which are described in section 2.4.1.

Top Enitities

`enriched_text.entities.text` #Paris (1,355), paris (946), #paris (176), #bataclan (110), #fusillade (98), CNN, #Breaking.

`enriched_text.entities.type` Hashtag (1,795), twitterhandle (1,335), Location (1,072), Company (215), Facility (155), Person (151), Organization (40)

`enriched_text.entities.disambiguation.name` Paris (104), CNN (88), Reuters (56), Stade de France (34), Bataclan (theatre) (17)

General Sentiments

8% positive, 42% neutral and 50% negative documents.
`enriched_text.sentiment.label` negative (911), neutral (769), positive (156)

Related Concepts

`enriched_text.concept.text` YouTube (61), Charlie Hebdo (55), France (47), Report (46), Terrorism(44), Hostage (43), Harshad number (27), 1 Night in Paris (26).

`Related Concepts` can be linked to `DBpedia` resources.

`enriched_text.concepts.text and enriched_text.concepts.dbpedia_resource`
Report (46): *http://dbpedia.org/resource/Report* (46) and *http://dbpedia.org/resource/Hostage* (1).

`enriched_text.concepts.text and enriched_text.concepts.dbpedia_resource`
Terrorism (44): *http://dbpedia.org/resource/Terrorism* (44) *http://dbpedia.org/resource/Hostage* (3) *http://dbpedia.org/resource/Acts_ of_ the_ Apostles* (2) *http://dbpedia.org/resource/Court* (1) *http://dbpedia.org/resource/Nuremberg_ Trials* (1) *http://dbpedia.org/resource/Religion_ of_ Peace*

(1)

`enriched_text.concepts.text and enriched_text.concepts.dbpedia_resource`
Hostage (43): *http://dbpedia.org/resource/Hostage* (43) *http://dbpedia.org/resource/Suicide_methods* (5) *http://dbpedia.org/resource/Terrorism* (3) *http://dbpedia.org/resource/Death* (1) *http://dbpedia.org/resource/Report* (1)

Content Hierarchy

`enriched_text.categories.label` /travel/tourist destinations/france (1,155), /news (331), /law, govt and politics/law enforcement/police (279), /society/unrest and war (133), /news/international news (127), /art and entertainment/shows and events/concert (122), /health and fitness/disorders/mental disorder/panic and anxiety (115).

4.6 Microsoft Azure Text Analytics

Microsoft Azure Text Analytics is an application within Microsoft Azure, described in section 2.4.2.

As a result of the limitations of the Microsoft Azure Student account and the API requirements, a few changes had to be made to the Microsoft Azure Text Analytics API.

- Editing the format of IDs in the dataset to fit the API.
- Read maximum 5000 characters from the `field: text` only, because of the limitations.

An attempt to do the analysis by splitting each tweet into one file per tweet was made, due to Microsoft Azure Text Analytics' limitations. However, by analyzing 1000 different files, the documents are not analyzed as one unit, but as 1000 different units, which did not give any value for this analysis.

4.6.1 Analysis

The dataset was analyzed using the Microsoft Azure Text Analytics API. One document with 5000 characters of text from the tweets were analyzed. The current analysis was created at 12.05.18. The analysis is divided into four categories, which are described in section 2.4.2.

Language Detection

The `Language Detection` returns the language code with a score indicating the strength of the score.

```
"documents": [
    {
      "id": "1",
      "detectedLanguages": [
        {
          "name": "English",
          "iso6391Name": "en",
          "score": 0.9420731067657471
        }
      ]
    }
  ],
  "errors": []
```

Sentiment Analysis

The `Sentiment Analysis` returns a sentiment score between 0 and 1 for each document, where 1 is the most positive.

An error occurred in the analysis, which truncated the input to the first 100 tokens.

```
{
"documents": [
  {
    "score": 0.044938504695892334,
    "id": "1"
  }
],
"errors": [
  {
    "id": "1",
    "message": "Truncated input to first 100 tokens during analysis."
  }
]
}
```

Key Phrase Extraction

The Key Phrase Extraction extract key phrases to identify the main points. The key phrases which occurred most are presented here:

‘‘Paris explosions’’, ‘‘Paris attacks’’, ‘‘Bataclan concert hall’’, ‘‘innocent people’’, ‘‘hostage situations’’, ‘‘SPECIAL REPORT’’, ‘‘French police report gunfire’’, ‘‘Automatic gunfire’’, ‘‘terror attacks’’, ‘‘hostages’’, ‘‘shootings’’, ‘‘biggest terrorist attack’’, ‘‘thoughts’’, ‘‘shooting incident’’, ‘‘prayers’’, ‘‘fucking sick’’, ‘‘StadedeFrance’’, ‘‘ParisAttacks’’, ‘‘horrifying images’’, ‘‘soccer match’’, ‘‘L’explosion’’, ‘‘war zone’’, ‘‘PrayForParis’’, ‘‘terrorism’’

See appendix, section A.3.2, for the full result.

Entity Linking

The Entity Linking identifies well-known entities in the text and links the entities to more information on the web. A relevant selection is presented here:

name: Place de la République
matches: “place de la République”
wikipediaId: “Place de la République”
wikipediaURL: https://en.wikipedia.org/wiki/Place_de_la_République

name: Bataclan (theatre)
matches: “Bataclan concert hall”, “Bataclan Concert Hall”, “Bataclan Concert Hall”, “Bataclan Concert Hall”, “Bataclan Concert Hall”, “le Bataclan”
wikipediaId: “Bataclan (theatre)”
wikipediaURL: https://en.wikipedia.org/wiki/Bataclan_(theatre)

name: November 2015 Paris attacks"
matches: “attacks in Paris”, “Paris #Shootings”, “Paris #shooting”
wikipediaId: “November 2015 Paris attacks”

`wikipediaURL:` https://en.wikipedia.org/wiki/November_2015_Paris_attacks

See appendix, section A.3.1, for the full result.

4.7 Keyword Frequency Script

Keyword Frequency Script is a own developed script for keyword frequency, described in 2.4.3.

The script is developed using JavaScript and Node.js, described in section 2.5.4). The predicate for occurring words are < *three letters long* and *occurring over 300 times*. See appendix, section A.4, for the complete script.

4.7.1 Analysis

The result of the Keyword Frequency Script listed with the keyword and the number of times the keyword has occured.

The current analysis was created at 13.05.18.

```
#Paris - 5716,
#paris - 872,
#Bataclan - 789,
dead - 776,
#fusillade - 734,
attacks - 669,
people - 621,
concert - 483,
several - 466,
#Paris.  - 451,
police - 376,
explosions - 362,
hostages - 341,
reports - 341,
hostage - 323,
French - 317,
horrible - 301
```

Chapter 5

Discussion

This chapter will present a discussion of the research question, exemplified by the data analysis of the Paris attacks, in addition to the analysis services, techniques and results. Further, the different working methods and methodologies that were utilized in the research will be discussed.

5.1 Analysis Result

The research questions for this thesis aims to explore how data from social media could be used for emergency management.

The data in this analysis is collected from Twitter, containing some of the most used hashtags during the Paris attacks: #Paris, #Bataclan, #Parisattacks and #Porteuverte. The data analysis in this research is therefore directly linked to the event of the Paris attacks. In case of an unexpected ongoing event, the data connected to the event will just be a small part of a large amounts of unstructured data. The data would not have been filtered on hashtags or any other preparations made in this analysis, such as languages.

For the question regarding detection of the event, the direct linking between the data and the event will have an effect. All the tweets analyzed is associated with the incident and the event is therefore, in this case, already detected. In the analysis I will look at detecting in terms of what has been discovered. The linking between the data and the event is not as relevant for the research questions regarding what kind of information the data provides, as a detected and known event can continue to provide new information.

RQ1: How can Big Data from social media be used in emergency management?

The first question is, **How early can an ongoing emergency situation be detected?**

Related work, presented in section 2.6, shows that there is often a sharp and sudden increase on social media immediately after emergency events. In the article "An Analysis of Twitter Messages in the 2011 Tohoku Earthquake" it was revealed that the first tweet about the disaster were 1 minute and 25 seconds after the earthquake happened at the epicenter (Doan et al., 2012). In previous research, it is clearly proven that people use social media to communicate and publish information when an emergency situations occurs (Denis et al., 2014, Doan et al., 2012, Earle et al., 2010, Kongthon et al., 2014, Muralidharan et al., 2011, Sakaki et al., 2010).

The first report from the terrorist attack in Paris, the explosions near Stade de France where the international football game between France and Germany took place, was around 9.20PM. At 9.19PM, a German Twitter user, who was watching the football match, wrote about the explosions. He was asking the German national football teams twitter account if there had been an explosion or if it was harmless "Explosion in the Stade de France? Was it a bomb or was it harmless? Explosion today here in France, in the stadium". Several other Twitter users published continuous live updates, including location information and pictures, during the entire attack (BBC, 2015b). During the Paris attacks, the hashtag #fusillade went from zero to 36,000 tweets in few minutes, by 23.00, the hashtag had reached 3,9 million tweets. The hashtag #parisattacks, which seem to be used to follow information, went from zero to 1,9 million tweets in five hours (Trajkovic, 2015).

As 500 million tweets are sent every day (Stats, 2018), it can be difficult to detect an event just by watching the Twitter flow. The article "Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors" presented in section 2.6, shows that via a classifier of tweets, based on features such as keywords in the tweets, the number of words and their context, it is possible to both target events and get useful information (Sakaki et al., 2010).

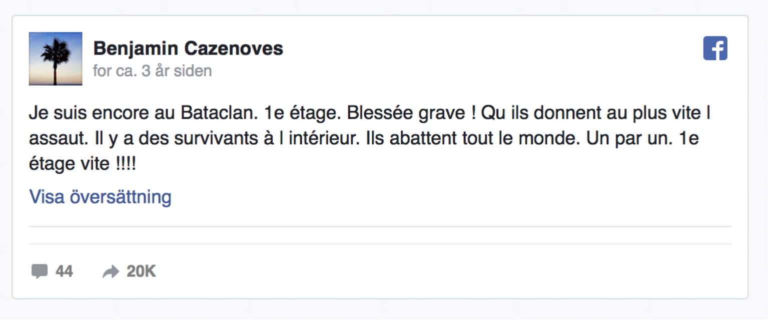
I will give an answer to these questions based on the analysis results later in this section.

Second question, **Can the analysis result provide important information for emergency management and rescuers?**

As the mentioned article "Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors" proved, useful information from social media can be detected by using criteria for the data (Sakaki et al., 2010). As described in the theory, section 2.3.1, emergency managements mission is to protects communities by coordinating and integrating all activities necessary to build, sustain, and improve the capability to mitigate against, prepare for, respond to, and recover from threatened or actual natural disasters, acts of terrorism, or other man-made disasters. The useful information in this thesis is in context with this mission.

During the Paris attacks, Twitter quickly turned into a message board on Friday night with information to help people in Paris get to safety. The hashtag #PorteOuverte — "open door" — became a vehicle for offering shelter to those in Paris who needed it. The hashtag reached one million tweets in 10 hours. The hashtag #RechercheParis, which was accompanied with descriptions of loved ones and requests for information, and used to share news when someone who had been sought was found alive, reached a million tweets within 24 hours (Goel and Ember, 2015). Social media turned into the world's primary source for responding to the terror, pleading for rescue from gunmen murdering people in the Bataclan theater, documenting pandemonium on the capital's streets and hunting for loved ones lost in the chaos of the shootings. A man trapped in the Bataclan theater described the carnage on Facebook, calling for help and urging police to storm the venue (Kayali and O'Rourke-Potocki, 2015).

Figure 5.1 – Message published on social media during the Paris attacks



Benjamin Cazenoves
for ca. 3 år siden

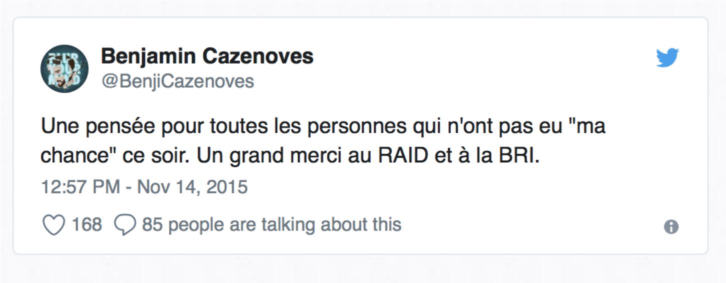
Je suis encore au Bataclan. 1e étage. Blessée grave ! Qu ils donnent au plus vite l assaut. Il y a des survivants à l intérieur. Ils abattent tout le monde. Un par un. 1e étage vite !!!!

Visa översättning

44 20K

"I am still at the Bataclan. First floor. Heavily wounded. May they launch the attack as soon as possible. There are survivors inside. They are killing everyone. One by one. Quickly, to the first floor!" (Kayali and O'Rourke-Potocki, 2015).

Figure 5.2 – Message published on social media during the Paris attacks



Benjamin Cazenoves
@BenjiCazenoves

Une pensée pour toutes les personnes qui n'ont pas eu "ma chance" ce soir. Un grand merci au RAID et à la BRI.

12:57 PM - Nov 14, 2015

168 85 people are talking about this

After the rescue, Cazenoves tweeted his thanks to the police and his prayers for those who "didn't have my luck." (Kayali and O'Rourke-Potocki, 2015).

Even the police turned to social media in the forlorn hope of getting people to stop tweeting and posting during the attacks, since it could interfere with and have a negative impact on the operation to free the hostages in the Bataclan music hall (Kayali and O'Rourke-Potocki, 2015).

In an analysis including raw data from social media, it is important to consider the possibility of misinformation. For example, users might make tweets such as "Terrorism!" or "People have been shot around me, please come help", for which "terrorism" or "shot" could be the keywords, but users might also make tweets such as "I am attending an conference about terrorism", or "I just tasted the worst Russian vodka shot". Furthermore, even if a tweet is referring to the target event, it might not be appropriate as an event report. Therefore, it is important to look at the entire analysis.

I will now give a presentation of the analysis, set in context with the event and emergency management. Further I will answer the questions around ability of detecting the situation and the information quality for emergency management and responders, exemplified by the data analysis of the Paris attacks. Lastly, I will give a summary of the entire analysis.

5.1.1 IBM Watson Discovery

Top Entities

The results mainly showed locations `(Paris and Bataclan)` and news information `(CNN)`, except from the entity `#fusillade`. The definition of fusillade is a number of shots fired simultaneously or in rapid succession (Webster, 2018a).

The entity `fusillade` can provide indications of an emergency situation, in the form of the unique significance that is only linked to special kinds of events. The locations `Paris` and `Bataclan` does not provide any significant information for emergency rescuers alone, but in the context of `fusillade`, it could provide important location information in addition to what is detected. Several shots fired simultaneously in/near Paris or Bataclan.

General Sentiments

With a result of 8% positive, 42% neutral and 50% negative documents, the overall sentiment is neutral/negative.

The result of an sentiment analysis alone cannot provide directly indications of an emergency situations or what is detected. A negative sentiment in a flow of tweets may have several reasons and is difficult to put in context. In addition, the result of an overweight of negative sentiment does not provide any main information for emergency management and rescuers, apart from a preparatory factor regarding what emotional level can meet them in the field.

Related Concepts

The results showed the concepts *Youtube, Charlie Hebdo, France, Report, Terrorism, Hostage, Harshad number and 1 Night in Paris. Charlie Hebdo* is the name of a French satirical magazine, which was the target of another terrorist attacks in 2015 (BBC, 2015a). A *Report* can be defined as a common talk or an explosive noise (Webster, 2018b). The other relevant concepts, such as *Terrorism* and *Hostage*, speaks for themselves.

These concepts, both alone or all together, provides an indication of what has been detected and takes place in this situation. `Report` can give an indication of explosions. `terrorism` gives strong indications of what kind of an emergency situation we are dealing with and `Hostage` tells something about the situation and that there may be people in need of great help/be rescued. These results give important value for the emergency management and rescuers.

Content Hierarchy

The categories */law, govt and politics/law enforcement/police, /society/unrest and war and /health and fitness/disorders/mental disorder/panic and anxiety* was the relevant top categories.

These results provide information about what type of information is displayed. The category `/society/unrest and war and /health` sets the information in context with concepts which is natural for an emergency. `/health and fitness/disorders/mental disorder/panic and anxiety` tells about the state among the population. Disorder, mental disorder, panic and anxiety are categories which can be associated with an emergency situation. Information connected to society/unrest/war and disorder/mental disorder/panic/anxiety can indicate turmoil in the society. In addition, it can provide information about the state among the population.

Result

The total analysis from IBM Watson Discovery shows the most occurring entities, sentiments, concepts and categories. Looking at the entire analysis, several indications of an emergency situation are present. `Fusillade`, `Paris`, `Bataclan`, `France`, `Report`, `Terrorism`, `Hostage`, `/society/unrest and war and /health` and `/health and fitness/disorders/mental` `disorder/panic and anxiety`.

In view of the real event, which took place in France, Paris, mainly at the Bataclan Theater, where terrorists had explosives, took hostages and killed many people, the analysis manages to pick up major keywords. Based on the high occurrences of words and concepts strongly linked to an emergency situation, we can conclude that the analysis could have provided useful information in terms of what is going on, where is the event happening, there may be injured people and people who are in need of help.

According to the research question, IBM Watson Discovery provides results that can be used to detect, coordinate and provide assistance. Regarding the time perspective for the detection of the event, it is difficult to determine, because of the already filtered data. However, we can conclude that IBM Watson Discovery manages to pick up key concepts and associate relevant information from a small amount of data. A large amount of information that is considered important also appears in the analysis.

5.1.2 Microsoft Azure Text Analytics

Language Detection

In this analysis, only one language was detected, English - with a score on 0.94, which indicates that 94% of the analyzed text is in English. This is a result of the filtering that was made prior to the analysis.

The language detection analysis turned out to be not particularly relevant to this analysis and will therefore be excluded from further results.

Sentiment Analysis

The result was 0.044, which indicates that the sentiment in the document is very negative

The result of a negative sentiment cannot provide indications of an emergency situation. A negative sentiment in a flow of tweets may have several reasons and is difficult to put in context. In addition, the result of a very negative sentiment does not give any concrete value for emergency management or rescuers in this case.

Key Phrase Extraction

The result from the analysis showed a selection of almost 100 different key phrases. Despite the large number, there are several key phrases that hit very well within concepts linked to emergency management

‘‘Paris explosions’’, ‘‘Paris attacks’’, ‘‘Bataclan concert hall’’, ‘‘innocent people’’, ‘‘hostage situations’’, ‘‘SPECIAL REPORT’’, ‘‘French police report gunfire’’, ‘‘Automatic gunfire’’, ‘‘terror attacks’’, ‘‘hostages’’, ‘‘shootings’’, ‘‘biggest terrorist attack’’, ‘‘thoughts’’, ‘‘shooting incident’’, ‘‘prayers’’, ‘‘fucking sick’’, ‘‘StadedeFrance’’, ‘‘ParisAttacks’’, ‘‘horrifying images’’, ‘‘soccer match’’, ‘‘L’explosion’’, ‘‘war zone’’, ‘‘PrayForParis’’, ‘‘terrorism’’

Several of the key phrases are strongly linked to concepts associated with an emergency situation. With a high incidence of key phrases in this category it is clear that an unusual situation is detected. The key phrases can provide useful information for emergency management and rescuers, in form of what is happening; Paris Explosion, hostage situations, French police report gunfire, biggest terrorist attack, shooting incident, L’explosion, terrorism, automatic gunfire, terror attacks, where the is event happening; Paris attacks, Bataclan concert hall, StadedeFrance, what kind of event this is; Biggest terrorist attack, terrorism, terror attacks, if someone is injured; French police report gunfire, horrifying images, hostage situations, innocent people and indications that there may be people in need of help; hostage situations, innocent people

Linked Entities

The linked entities identify entities and links them to more information on the web. The main result from the analysis is location information, linked to the actual place. Though the linked entities analysis mainly consists of informative results, an interesting result arise as well.

name: November 2015 Paris attacks"
matches: “attacks in Paris”, “Paris #Shootings”, “Paris #shooting”
wikipediaId: “November 2015 Paris attacks”
wikipediaURL: https://en.wikipedia.org/wiki/November_2015_Paris_attacks

This is by now a known event and that is why it is possible to associate the entities with the specific event. Had this, on the other hand, been a new unknown event, it may have been the link in the same way - to an earlier similar emergency situation.

The direct linking to the Paris attacks cannot provide any indications of an emergency situation. However, if the analysis links several entities to an already known event, it may be an indication that this is a similar event. If it's already clear that an ongoing emergency situation is happening, the location information can provide important information about where.

Result

Microsoft Azure Text Analytics provides varying results without any clear indications. The Key Phrase Extraction is probably the part of the analysis that provides the most value, by showing the top key phrases, which actually gives some signs of what is detected and providing information about the event. Though, in an unstructured data flow, the key phrases would not be filtered or sorted by top values and it's therefore unclear if the analysis would have given any value at all.

According to the research question, Microsoft Azure Text Analytics provides few results that can be used to detect, coordinate and provide assistance. The analysis does not make any connections between the different results. Regarding the time perspective for the detection of the event, it is difficult to determine, because of the already filtered data.

5.1.3 Keyword Frequency Script

Keywords

The result shows the main keywords in the data. Several of the main words that can be strongly linked to an event of an emergency situation.

`#Paris`, `#paris`, `#Bataclan`, `dead`, `#fusillade`, `attacks`, `people`, `concert`, `several`, `#Paris`, `police`, `explosions`, `hostages`, `reports`, `hostage`, `French`, `horrible`

Several indications of what has been discovered appear in this analysis `dead`, `#fusillade`, `attacks`, `explosions`, `hostages` and `hostage`. In addition, the following keywords can provide information for the emergency management and rescuers. What is happening; `dead`, `attacks`, `hostage`, `hostages`, where the event is happening; `#Paris`, `#paris`, `#Bataclan`, `#Paris`, `French`, what kind of event this is; `attacks`, `hostage`, `explosions`, if someone is injured; `dead`, and indications that there may be people in need of help; `dead`, `hostages`, `hostage`. However, a keyword frequency scripts ability to detect and provide important information depends on the criteria's set in the tool, which will further be described in section 5.3.

Result

Based on the high occurrences of words strongly linked to an emergency situation, we can conclude that the analysis could have provided useful information in terms of what is going on, where is the event happening and that there may be injured people and people who are in need of help.

According to the research question, the Keyword Frequency Script, provides results that can be used to detect, coordinate and provide assistance. The data is not matched against each other or any other information, it is therefore difficult to determine the capability of how early a situation can be detected or the information value.

5.2 Analysis Services

In this research, the analysis services IBM Watson Discovery and Microsoft Azure Text Analytics, in addition to an own developed script, Keyword Frequency Script, have been used for the data analysis. I will now discuss if the analysis services have proven to be useful in the case of emergency management.

RQ2: How can existing analysis services be useful for emergency management?

As a major technological innovation of recent years, social media have reshaped the nature of digital information sharing and networking. Traditionally, members of the public have relied on emergency rescuers and news media to provide information about emergency and disaster events. However, with the growth of social media, important and informative disaster-related information have been shared online from eyewitness and others experiencing the event (Denis et al., 2014, Doan et al., 2012, Earle et al., 2010, Kongthon et al., 2014, Muralidharan et al., 2011, Sakaki et al., 2010). The possibility of obtaining information directly from the public through informal sources is particularly valuable when the information is not yet covered by traditional emergency systems.

A better understanding of disasters and their effects on life, property, society and the environment is emerging. Researches are developing and accessing new ways of responding to emergencies using computers and communication technology with analysis and modelling techniques, including risk analysis, simulation, decision support system, artificial intelligence/expert systems and geographical systems to increase our knowledge about emergency situations and disasters, in order to develop new approaches for managing emergencies (Tufekci and Wallace, 1998). Today we have a large number of technology companies offering software or services for this purpose. With the extreme amount of data generated through an emergency situation, it is interesting to see if already-known analysis services can provide important support, in terms of discovering, coordinating, communicating and helping.

I will now present the different analysis services used in this research and discuss their ability to be useful managing an emergency situation.

IBM Watson Discovery

IBM Watson Discovery has an intuitive interface for both reading input data and previewing the result of the simple analysis. Furthermore, for doing a more specific analysis, an use of queries to match different categories were necessary.

IBM Watson Discovery has proven to be useful for emergency management. The analysis services make it feasible to do a comprehensive data analysis, including the possibility of matching the different analysis categories.

Microsoft Azure Text Analytics

Microsoft Azure Text Analytics provides an API for the text analytics services, which means there is no interface or GUI to interact with. The API only offers input reading by manually writing content into the API and the results are previewed in the computers terminal.

As a result of the APIs lacking features, a file reading method had to be implemented. Microsoft Azure supports several different programming languages and in this analysis the programming technologies JavaScript and Node.js, see section 2.5.4, were used to do necessary changes to the API.

Due to the limitations, described in 4.1, the analysis could contain of either one file of a maximum of 5,000 characters or a total of 1000 files. In order to analyze the greatest amount of data, the analysis was done by analyzing 1000 files with one tweet in each file. Unfortunately, the analysis was not as expected as the files were analyzed one by one and not as an entire analysis. This led to an analysis with only 5,000 characters being analyzed.

Microsoft Azure Text Analytics has not proven to be useful in this research. The analysis service requires a lot of preparation and is not suitable for unstructured big data analysis as it looks at each document as a unit and not as a part of the entire data analysis. To get the full potential of the Microsoft Azure Text Analytics, a preprocessing phase where you set up the necessary intermediary services is required.

Keyword Frequency Script

The Keyword Frequency Script is accessible through the computers terminal. The script is written in JavaScript with Node.js, see section 2.5.4. The Keyword Frequency Script is a script showing the top words occurring in the data and has no limitations as it is itself developed and can be modified according to the purpose. Though, the script could have been developed in other ways, e.g. applying a word of stop-words and “alert” if defined words in connection with emergency situations occur. In this research, this was not the angle of analysis, as the analysis should take place in such a way that you did not know which words might occur.

The Keyword Frequency script has been proven to be useful for emergency management. By monitoring on features such as keywords in the tweets and the amount of them, it is possible to both target events and get useful information. However, by further work on developing a standard of crucial keywords, a higher capacity of both detection and information can be provided.

5.2.1 Limitations of the Services

This section will discuss the limitations in the services and tools used in the analysis.

The services are not designed to make an ongoing event analysis. To be able to transfer the data directly, you would have to set up an intermediate service, in the form of a script or a similar technology, to handle the transfer between Twitter and the service.

The services also have different requirement regarding receiving the data. When obtaining data directly from the Twitter API, the data contains large amount of information, which is not necessarily relevant. As earlier mentioned, one raw tweet obtained from the Twitter API included over 80 attributes. None of the services used in the analysis have the functionality of handling a filtering process.

5.3 Implication for Practice

To be able to use data directly from social media under any circumstance, an intermediate service, in form of a script or similar technology, must be in place. An intermediate service must handle the transfer from the social media to the service, the filtering of the data to only get the necessary information from the dataset, in addition to other preparations, for example input requirements and language detection and translation.

In addition, my recommendations for preparatory work for data from social media during an emergency are as follows.

- A list of stop-words often related to or emergency situations, such as bombs, terrorism, hostage, fusillade, explosion.
- A list of hashtags often related to or used under such circumstances, for example #terror, #attack, #fusillade, #bombings.

5.4 Methods and Methodologies

The case study framework was used throughout the whole project to integrate methods that were applied in the research. The framework has a research approach with six different phases which resulted in a structured way of carrying out the research. Chapter 3 explains in detail how all the steps that were executed resulting in the analysis.

5.4.1 Criticism of the Research Method

During the planning phase of this research, it was assumed that UiB would provide the analysis services. Unfortunately, negotiation of agreements took longer than planned, in addition to one of the agreements were terminated for a period. This led to limited time to learn and understand the services, something that I should have taken into consideration during the planning phase. Another problem regarding the analysis services was the limitations in the accounts provided by UiB. This led to the fact that it was not possible to analyze large amounts of data. With larger amounts of data, the data basis in the analysis would be larger and the results would therefore be more accurate.

The reason for choosing the Paris attacks as the emergency situation in this thesis is the huge use of Twitter during the event. Twitter has been used in several other disasters, but this is one of the first times that it has been possible to follow the course of events on social media to such an extent, with detailed descriptions, images, audio and videos (Twitter, 2015). Twitter has also been highly controversial for this event because of all the detailed and sensitive information that was shared. As far as I am aware, no previous studies of the data from Twitter during the Paris attacks has been made at all. This has led to the absence of any "gold standard" or a blueprint for the data, which limited the possibility of knowing if my results were correct. Though, this is a now known event, and therefore it was possible to associate the results from this analysis with what is already known about the event.

It may be problematic to analyze an event in a language other than where it originally originated, in this case an analysis in English from an event in France. This has not been shown to be a problem for this analysis, as the amount of information in English has been greater than the capacity for the analysis. Twitter is a worldwide network and English is the most widely used language, which also was confirmed by the language detection analysis in preprocessing - language detection phase of the analysis. However, by excluding the language from where the incident occurred, there is a great possibility that important information is omitted.

In the process within the data and analysis, more work could have been put into the preprocessing phase of the data. The filtering process of English language only never reached a 100% accuracy. However, after doing several test-analysis as most of the language filtering was done, it became clear that further filtering would not affect the results. The analysis results provided varying degrees of quality, the further work section will describe how the analysis services and results could be taken to a further level.

Research Design Quality

Construct Validity

The study shows a high degree of construct validity. Several of the results in this study, correlate with previous finding in similar studies, which is presented in the theory, chapter 2.

Internal Validity

The research was designed to explore as many rival explanations as possible by initially being exploratory, as explained in section 3.2.2. This is done by using several different analysis services, with the possibility to do various testing. The research is exploratory organized for the initial phase, while transfering over to an explanatory fashion when sufficient data is gathered about a phenomenon.

Due to time limitations of the study, all causal factors related to the research were not explored. This weakens the internal validity of the study. However, the conclusions and answers found in this thesis is supported by the result of the data analysis.

External Validity

The case presented in this thesis aim to generalize towards other events within similar contexts. According to Yin (Yin, 2008), a theory must be tested by replicating the findings in a second or even a third case, where the theory has specified that the same results should occur.

This study can generally not be transmitted or replicated in other cases, but the theory and findings can be used as an underlying basis.

Realibiltiy

The steps of collection, storing, preprocessing and analyzing the data is carefully described in this thesis. To minimize errors and bias in the study, project tools were used to document the procedures and results along the way. The process within the analysis, including the tools and technologies, are well described and documented. By doing this, the study should yield at the same result each time.

Chapter 6

Conclusion and Further Work

This chapter presents the results found in the research, in addition to the further work.

In this research, a data analysis of Twitter during the Paris attacks in 2015 are presented. The goal of the analysis was to explore how data from social media can be used in emergency management. The analysis is done by using the data analysis services IBM Watson Discovery, Microsoft Azure Text Analytics and a Keyword Frequency Script. The research is backed up by literature, which all agree about the importance of establishing a method for handling big data from social media in emergency management.

6.1 Research Question

As discussed in section 1.3, this thesis seeks to answer two research questions. The following section will provide an answer to these questions.

RQ1: How can big data from social media be used in emergency management?

RQ2: How can existing analysis services be useful for emergency management?

The increasing use of social media is changing the way people communicate. During emergency situations, information available from the public can be utilized to inform situational awareness of emergencies and to help crisis coordinators respond appropriately. This information will not replace existing procedures and information sources, but it can provide a new source of data that has big potential within emergency management. Examples where social media can play a role is detection, real-time notification of an incident occurring, first-hand reports of the incident, and the state among the public.

Answering research question one, the result of this research has proven that data from social media can be used for detecting, identifying and responding to emergency situation, using the potential to track crucial information from messages published on social media. Though, further implication for practice is recommended to achieve more accurate results. Answering research question two, existing analysis services proves potential for handling the information flow, providing features for analyzing data from social media. However, raw collected data, in addition to well established analysis tools requires preparations, in forms of preprocessing and filtering, which limits the usage.

Social media technologies shows potential for leveraging public participation in disaster response. When properly employed, the benefits of social media are faster decision-making and more complete knowledge resources. The strong correlation between social media and emergency events leads to an assumption that data from social media can be an useful resource in an early warning surveillance systems as well as a tool for analyzing event information during times of disaster.

6.2 Further Work

The thesis case study gives insight in some of the challenges concerning big data from social media and emergency management. Further work on the research could lead to a deeper and more exhaustive investigation of the phenomena in the case.

The study has the potential to be further developed and investigated in several ways. A change of the technical direction, by using several other analysis services and analysis technologies, is recommended. Several analysis services can provide better and more accurate results, in addition to an indication of what is the appropriate tool for this type of analysis. It is also possible to develop own tools for this work, as it has been proven that well established analysis tools require a lot of preparation for handling the data and providing necessary useful information, which could be solved by handling all of development itself.

The case studied in this thesis could be further investigated, by analyzing a larger amount of data from the dataset. A larger amount of data will provide a better basis for the results. It would also be interesting to analyze a completely different emergency event. In addition, an analysis of unstructured data from a period of time when an emergency situation occurred can be analyzed, to quality assure both the detection capability and if the reports from the emergency situation provides any value in an unstructured data.

The evaluation of the data should be set against a gold standard, for quality assurance of the data. Involvment of experts at the field, for arguing and discussing the actual usefulness of the data, would have been preferred. To evaluate the data, interviews with people handling and assisting during an emergency situation would be interesting, to conclude if the results would have proven any valuable information for the managing and rescuing. In addition, interviews with people who actually used social media during the specific event, to provide and reach information, would have been useful, to determine if the shared information provided any value for them.

Technology is evolving all the time, and we cannot say for sure that Twitter will be one of the most widely used social media in a couple of years. It is therefore important to look at data from other social medias, to see how collection, storage and the preprocessing procedures would work out.

Unfortunately, such events, both natural and man-made, have occurred at higher rates the recent years. It is therefore important in the further work on how social media can be used for emergency management, there are several different events with different point of views, so that we can learn and understand what is important and what to look for in such an analysis.

Bibliography

J. Allen. *Recognizing intentions from natural language utterances*. 1983.

Nick Anstead and Ben O'Loughlin. Social media analysis and public opinion: The 2010 uk general election. *Journal of Computer-Mediated Communication*, 20(2):204–220. https://onlinelibrary.wiley.com/doi/abs/10.1111/jcc4.12102.

BBC. *Charlie Hebdo attack: Three days of terror - BBC News*. 2015a. http://www.bbc.com/news/world-europe-30708237.

BBC. *How the Paris attacks unfolded on social media - BBC News*. 2015b. http://www.bbc.com/news/blogs-trending-34836214.

BBC. *Paris attacks: What happened on the night - BBC News*. 2015c. http://www.bbc.com/news/world-europe-34818994.

Creative Commons. *Creative Commons Legal Code*. 2018. https://creativecommons.org/licenses/by/2.0/ca/legalcode.en.

French Language Daily. *1000 Most Common French Words - Top French vocabulary*. 2018. http://french.languagedaily.com/wordsandphrases/most-common-words.

DBpedia. *About | DBpedia*. 2018. https://wiki.dbpedia.org/about.

Lise Ann St. Denis, Leysia Palen, and Kenneth M. Anderson. Mastering social media: An analysis of jefferson county's communications during the 2013 colorado floods. In *ISCRAM*, 2014.

Son Doan, Bao-Khanh Ho Vo, and Nigel Collier. *An Analysis of Twitter Messages in the 2011 Tohoku Earthquake*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

Paul Earle, Michelle Guy, Richard Buckmaster, Chris Ostrum, Scott Horvath, and Amy Vaughan. *OMG Earthquake! Can Twitter Improve Earthquake Response?*, volume 81. 2010. http://dx.doi.org/10.1785/gssrl.81.2.246.

Simeon Edosomwan, Sitalaskshmi K. Prakasan, Doriane Kouame, Jonelle Watson, and Tom Seymour. The history of social media and its impact on business. *Journal of Applied Management and Entrepreneurship*, 16(3):79–91, 07 2011. Copyright - Copyright Nova Southeastern University, H. Wayne Huizenga School of Business and Entrepreneurship Jul 2011; Document feature - ; Last updated - 2016-04-23.

Fema. *Principles of Emergency Management | FEMA.gov*. 2017. https://www.fema.gov/media-library/assets/documents/25063.

Amir Gandomi and Murtaza Haider. Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, 35(2):137 – 144, 2015. http://www.sciencedirect.com/science/article/pii/S0268401214001066.

GitHub. *GitHub - DocNow/twarc: A command line tool (and Python library) for archiving Twitter JSON*. 2018. https://github.com/DocNow/twarc.

Vindu Goel and Sydney Ember. *As Paris Terror Attacks Unfolded, Social Media Tools Offered Help in Crisis - The New York Times*. BBC, 2015. https://www.nytimes.com/2015/11/15/technology/as-paris-terror-attacks-unfolded-social-media-tools-offered-help-in-crisis.html.

Wu He, Shenghua Zha, and Ling Li. Social media competitive analysis and text mining: A case study in the pizza industry. *International Journal of Information Management*, 33 (3):464 – 472, 2013. http://www.sciencedirect.com/science/article/pii/S0268401213000030.

Rob High. *The era of cognitive systems: An inside look at IBM Watson and how it works*. 2012. http://johncreid.com/wp-content/uploads/2014/12/The-Era-of-Cognitive-Systems-An-Inside-Look-at-IBM-Watson-and-How-it-Works_.pdf.

Martin et al. Hirsch. *The medical response to multisite terrorist attacks in Paris*. 2015.

C. Hochmuth. *How Twitter helps in emergencies, disasters*. 2015. https://www.fedscoop.com/social-medias-role-in-emergencies-disasters/.

Hydrator. *GitHub - DocNow/hydrator: Turn Tweet IDs into Twitter JSON from your desktop! Edit.* 2018. https://github.com/DocNow/hydrator.

IBM. *Watson Discovery Service: understand your data at scale with less effort - Watson.* 2018. https://www.ibm.com/blogs/watson/2016/12/watson-discovery-service-understand-data-scale-less-effort/.

Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng. Why we twitter: Understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis*, WebKDD/SNA-KDD '07, pages 56–65, New York, NY, USA, 2007. ACM.

Warren Murray Jon Henley, Ian Traynor. *Paris attacks: EU in emergency talks on border crackdown | World news | The Guardian.* 2015. https://www.theguardian.com/world/2015/nov/20/paris-attacks-france-launches-un-push-for-unified-declaration-of-war-on-isis.

Laura Kayali and Helena O'Rourke-Potocki. *The Paris attacks, as told through social media – POLITICO.* 2015. https://www.politico.eu/article/the-paris-attacks-as-told-through-social-media/.

Rob Kitchin. *The data revolution: Big data, open data, data infrastructures and their consequences.* Sage, 2014.

Alisa Kongthon, Choochart Haruechaiyasak, Jaruwat Pailai, and Sarawoot Kongyoung. *The Role of Social Media During a Natural Disaster: A Case Study of the 2011 Thai Flood*, volume 11. 05 2014.

Vasileios Lampos and Nello Cristianini. Tracking the flu pandemic by monitoring the social web. In *In Proceedings of the 2nd IAPR Workshop on Cognitive Information Processing*, pages 411–416. IEEE Press.

D. Laney. *3D Data Management: Controlling Data Volume, Velocity, and Variety.* 2001. https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf.

Weizhong Li, Lukasz Jaroszewski, and Adam Godzik. *Clustering of highly homologous sequences to reduce the size of large protein databases*, volume 17. 2001. http://dx.doi.org/10.1093/bioinformatics/17.3.282.

Dr. Merete Lunde and Prof. Dr. Rajendra Akerkar. *Project Description: Transnational Partnership for Excellent Research and Education in Big Data and Emergency Management (BDEM)*. 2017.

Microsoft. *Cognitive Services APIs Reference*. 2018a. https://westus.dev.cognitive.microsoft.com/docs/services/TextAnalytics.V2.0/operations/56f30ceeeda5

Microsoft. *Hva er Azure – Skytjeneste fra Microsoft | Microsoft Azure*. 2018b. https://azure.microsoft.com/nb-no/overview/what-is-azure/#most-popular-questions.

Microsoft. *Node.js Quickstart for Azure Cognitive Services, Text Analytics API | Microsoft Docs*. 2018c. https://docs.microsoft.com/nb-no/azure/cognitive-services/text-analytics/quickstarts/nodejs.

Microsoft. *Supported languages in Text Analytics API (Microsoft Cognitive Services on Azure) | Microsoft Docs*. 2018d. https://docs.microsoft.com/en-us/azure/cognitive-services/text-analytics/text-analytics-supported-languages.

Microsoft. *Text Analytics API overview (Microsoft Cognitive Services on Azure) | Microsoft Docs*. 2018e. https://docs.microsoft.com/en-us/azure/cognitive-services/text-analytics/overview.

A. Monappa. *How Facebook is Using Big Data: Good, Bad & the Ugly*. 2015. https://www.simplilearn.com/how-facebook-is-using-big-data-article.

MongoDB. *Introduction to MongoDB — MongoDB Manual 3.4*. 2018. https://docs.mongodb.com/manual/introduction/.

Mozilla. *About JavaScript - JavaScript | MDN*. 2018. https://developer.mozilla.org/en-US/docs/Web/JavaScript/About_JavaScript.

Sidharth Muralidharan, Leslie Rasmussen, Daniel Patterson, and Jae-Hwa Shin. Hope for haiti: An analysis of facebook and twitter usage during the earthquake relief efforts. *Public Relations Review*, 37(2):175 – 177, 2011. http://www.sciencedirect.com/science/article/pii/S0363811111000294.

Node.js. *Node.js*. 2018. https://nodejs.org/en/.

Briony J Oates. *Researching information systems and computing*. SAGE Publications Ltd, 2006.

Routledge OConnor, Balasubramanyan and Smith. *From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series.* 2010. https://pdfs.semanticscholar.org/3b95/f5e0159078bf88627deb645d3997d9d9437d.pdf.

Nick Ruest. *Paris Attack tweets.* Scholars Portal Dataverse, 2017. doi: 10864/11312. http://hdl.handle.net/10864/11312.

Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: Real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 851–860, New York, NY, USA, 2010. ACM. http://doi.acm.org/10.1145/1772690.1772777.

Rachel Schutt and Cathy O'Neil. *Doing Data Science: Straight Talk from the Frontline.* O'Reilly Media, Inc., 2013. ISBN 1449358659, 9781449358655.

A.T.M Shahjahan and Kutub Uddin Chisty. *Social Media Research and Its Effect on Our Society.* World Academy of Science, Engineering and Technology International Journal of Information and Communication Engineering, 2014. https://waset.org/publications/9998891/social-media-research-and-its-effect-on-our-society.

Homeland Security Statement of Subcommitee Chariman Susan W. Brooks. *Emergency MGMT 2.0: How SocialMedia and New Tech are Transforming Prepardness, Response and Recovery Disasters Part1 Privatsector.* 2013. https://homeland.house.gov/files/documents/06-04-13-Brooks-Open.pdf.

Statista. *Leading global social networks 2018 | Statistic.* 2017. https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/.

Internet Live Stats. *Number of Internet Users (2016) - Internet Live Stats.* 2017. http://www.internetlivestats.com/internet-users/.

Internet Live Stats. *Twitter Usage Statistics - Internet Live Stats.* 2018. http://www.internetlivestats.com/twitter-statistics/.

Christoph K. Streb. Encyclopedia of case study research. pages pages 373–374, 2010. http://sk.sagepub.com/reference/casestudy.

Techopedia. *What is Big Data? - Definition from Techopedia*. 2018. https://www.techopedia.com/definition/27745/big-data.

Jonathan Trajkovic. *ParisAttacks - How Twitter tells the story*. 2015. http://tipsandviz.blogspot.com/2015/11/parisattacks-how-twitter-tells-story.html.

Trello. *Kom i gang med Trello*. 2018. https://trello.com/guide.

S Tufekci and W Wallace. *The Emerging Area Of Emergency Management And Engineering - IEEE Journals and Magazine*. 1998. https://ieeexplore.ieee.org/abstract/document/669742/.

John W. Tukey. *The Future of Data Analysis*. Springer New York, New York, NY, 1992. https://doi.org/10.1007/978-1-4612-4380-9_31.

Twitter. *Paris attacked*. 2015. https://twitter.com/i/moments/665275299068813312.

Twitter. *Developer Policy — Twitter Developers*. 2017. https://developer.twitter.com/en/developer-terms/policy#6._Be_a_Good_Partner_to_Twitter.

Twitter. *Developer terms — Twitter Developers*. 2018a. https://developer.twitter.com/en/developer-terms.

Twitter. *GET statuses/lookup — Twitter Developers*. 2018b. https://developer.twitter.com/en/docs/tweets/post-and-engage/api-reference/get-statuses-lookup.

Twitter. *Privacy Policy*. 2018c. https://twitter.com/en/privacy.

Twitter. *About*. 2018d. https://about.twitter.com/.

Ushahidi. *Ushahidi*. 2018. https://www.ushahidi.com/.

W3Schools. *JSON Introduction*. 2018. https://www.w3schools.com/js/js_json_intro.asp.

IBM Watson. *Discovery - API reference | IBM Watson Developer Cloud*. 2018. https://www.ibm.com/watson/developercloud/discovery/api/v1/curl.html?curl#introduction.

Merriam Webster. *Fusillade | Definition of Fusillade by Merriam-Webster*. 2018a. https://www.merriam-webster.com/dictionary/fusillade.

Merriam Webster. *Report | Definition of Report by Merriam-Webster.* 2018b. https://www.merriam-webster.com/dictionary/report.

Merriam Webster. *Definition of Big Data.* 2018c. https://www.merriam-webster.com/dictionary/bigdata.

Zhou Yao and Cao Ze-wen. *Research on the Construction and Filter Method of Stop-word List in Text Preprocessing*, volume 1. 2011.

Dave Yates and Scott Paquette. Emergency knowledge management and social media technologies: A case study of the 2010 haitian earthquake. *International Journal of Information Management*, 31(1):6 – 13, 2011. http://www.sciencedirect.com/science/article/pii/S0268401210001453.

Robert K. Yin. *Case Study Research: Design and Methods (Applied Social Research Methods).* Sage Publications, fourth edition. edition, 2008. ISBN 1412960991.

Appendix A

Appendix

A.1 Data examples

This section contains example of data used in this thesis.

A.1.1 Example Tweet

This is an example of the full Tweet object. The Tweet object has a long list of root-level attributes, including fundamental attributes such as id, created_at, and text. Tweet objects are also the parent object to several child objects. Tweet child objects include user.

```
{
"contributors":
"coordinates":
"created_at":
"entities": {
  "hashtags": [
    {
      "indices": [

      ],
      "text":
    },
    {
      "indices": [


      ],
      "text":
    }
  ],
  "symbols":
  "urls":
  "user_mentions":
},
"favorite_count":
"favorited":
"geo":
"id": ,
"id_str":
"in_reply_to_screen_name":
"in\n_reply_to_status_id":
"in_reply_to_status_id_str":
"in_reply_to_user_id":
"in_reply_to_user_id_str":
"is\n_quote_status":
"lang":
"place":
"retweet_count":
"retweeted":
"source":
"text":
"truncated":
"user": {
  "contributors_enabled":
  "created_at":
  "default_profile":
  "default_profile_image":
  "description":
  "entities": {
    "description": {
      "urls": []
    }
  },
  "favourites_count\n":
  "follow_request_sent":
```

```
    "followers_count":
    "following":
    "friends_count":
    "geo_enabled":
    "has_\nextended_profile":
    "id":
    "id_str":
    "is_translation_enabled":
    "is_translator":
    "lang": "e\nn"
    "listed_count":
    "location":
    "name": "Danny Friel"
    "notifications":
    "profile_back\nground_color":
    "profile_background_image_url":
    "profile_backgr\nound_image_url_https":
    "profile_background_tile":
    "profile_banne\nr_url":
    "profile_image_url":
    "profile_image_url_https":
    "profile_link_color": "3C3940",
    "profile_sidebar_border_color":
    "profile_sid\nebar_fill_color":
    "profile_text_color": "0099B9"
    "profile_use_background_image":
    "protected":
    "screen_na\nme":
    "statuses_count":
    "time_zone":
    "translator_type":
    "url":
    "utc_offset":
    "v\nerified":
  }
}
```

A.2 Preproccessing

This section contains script used in the preprocessing phase for the data analysis.

A.2.1 Preproccessing: Script

This section presents the preproccessing script, for language filtering and input requirements. The script is developed using Javascript and Node.js, described in section 2.5.4.

```
const fs = require("fs");
const util = require("util");

const readFile = util.promisify(fs.readFile);
const writeFile = util.promisify(fs.writeFile);

async function getData() {
  return await readFile("10000tweets.json");
}


getData().then(async (data) => {
  let filedata = JSON.parse(data);
  let answer = filedata.tweets.filter(
    x =>
      !x.text.includes(" être ") &&
      !x.text.includes(" avior ") &&
      !x.text.includes(" je ") &&
      !x.text.includes(" de ") &&
      !x.text.includes(" ne") &&
      !x.text.includes(" pas ") &&
      !x.text.includes(" le ") &&
      !x.text.includes(" la ") &&
      !x.text.includes(" tu ") &&
      !x.text.includes(" vous ") &&
      !x.text.includes(" il ") &&
      !x.text.includes(" et ") &&
      !x.text.includes(" à ") &&
      !x.text.includes(" un ") &&
      !x.text.includes(" l’ ") &&
      !x.text.includes(" qui ") &&
      !x.text.includes(" aller ") &&
      !x.text.includes(" les ") &&
      !x.text.includes(" en ") &&
      !x.text.includes(" ça ")
  );

  answer = answer.map(x => ({tweetid: x.tweetid, text: x.text.replace(/[\n\r]/g, ’’)}));

  for(var i = 0; i<answer.length; i++){
    await writeFile(‘./tmp/${i}.json‘, JSON.stringify(answer[i]))
  }
});
```

A.2.2 Preprocessing: Microsoft Azure Text Analytics

This section present the final Microsoft Azure Text Analytics API. The API provided by the Microsoft Azure portal (Microsoft, 2018c). The API did not offer functionality for retrieving, read and viewing data, therefore, extensive modifications had to be made to the API.

```
using System;
using System.Collections.Generic;
using System.IO;
using System.Linq;
using System.Net.Http;
using System.Reflection.Metadata;
using System.Threading;
using System.Threading.Tasks;
using Microsoft.Azure.CognitiveServices.Language.TextAnalytics;
using Microsoft.Azure.CognitiveServices.Language.TextAnalytics.Models;
using Microsoft.Rest;
using Newtonsoft.Json;
using Newtonsoft.Json.Linq;

namespace Tweetanalyze
{

  class Program
  {
      static async Task Main()
      {
            ITextAnalyticsAPI client = new TextAnalyticsAPI(new ApiKeyServiceClientCredentials());
            client.AzureRegion = AzureRegions.Westus;

            var tweets = await ReadFilesAsync();
            var result = await client.DetectLanguageWithHttpMessagesAsync(new BatchInput(tweets));


            var tweetsWithLanguage = tweets.Select(x =>
                new MultiLanguageInput(
                    result.Body.Documents.FirstOrDefault(y => y.Id == x.Id).DetectedLanguages.First().Iso6391Name, x.Id,
                    x.Text)).ToList();

            var result2 = await client.KeyPhrasesWithHttpMessagesAsync(new MultiLanguageBatchInput(tweetsWithLanguage));
            var result3 = await client.SentimentWithHttpMessagesAsync(new MultiLanguageBatchInput(tweetsWithLanguage));

            var lines = new List<string>();
            foreach (var document in tweetsWithLanguage)
            {
                lines.Add(String.Format("Document ID: {0} , Language: {1}, Tweet: {2}", document.Id, document.Language, document.Text));
            }

            await File.WriteAllLinesAsync("../../../tweets.txt", lines);

            lines = new List<string>();
            foreach (var document in result2.Body.Documents)
            {
                lines.Add(String.Format("Document ID: {0}, Text: {1}", document.Id, tweetsWithLanguage.FirstOrDefault(y => document.Id == y.Id)?.Text));

                lines.Add("\t Key phrases:");

                foreach (string keyphrase in document.KeyPhrases)
                {
                    lines.Add(String.Format("\t\t" + keyphrase));
                }
            }
            await File.WriteAllLinesAsync("../../../pharases.txt", lines);

            lines = new List<string>();
            foreach (var document in result3.Body.Documents)
            {
                lines.Add(String.Format("Document ID: {0} , Sentiment
                   ↪ Score: {1:0.00} , Tweet: {2}", document.Id,
                   ↪ document.Score, tweetsWithLanguage.FirstOrDefault(y
                   ↪ => document.Id == y.Id)?.Text));
            }
            await File.WriteAllLinesAsync("../../../sentiment.txt", lines);


            Console.ReadKey();
        }

        static async Task<List<Input>> ReadFilesAsync()
        {
            var tweets = new List<Input>();

            for (int i = 0; i < 1000; i++)
            {
                var tweet = await
                   ↪ File.ReadAllTextAsync($@"../../../Tweets/{i}.json");
                try
                {
                    dynamic data = JObject.Parse(tweet);
                    tweets.Add(new Input(i.ToString(), data.tekst.Value));
                }
                catch (Exception e)
                {
                    //Console.WriteLine(e.Message);
                }
            }

            return tweets;
        }
```

```
  }

  class ApiKeyServiceClientCredentials : ServiceClientCredentials
  {
      public override Task ProcessHttpRequestAsync(HttpRequestMessage
         ↪ request, CancellationToken cancellationToken)
      {
          request.Headers.Add("Ocp-Apim-Subscription-Key",
             ↪ "2c2fea59a12b46e2917f349450ecb750");
          return base.ProcessHttpRequestAsync(request, cancellationToken);
      }
  }
}
```

A.3 Results of the Analysis

This section shows the full result from the Linked Entities and Key Phrase analysis from Microsoft Azure Text Analytics.

A.3.1 Microsoft Azure Text Analytics: Linked Entities

The result from the Microsoft Azure Text Analytics linked entities analysis.

```
  {
  "documents": [
    {
      "id": "1",
      "entities": [
        {
          "name": "Place de la République",
          "matches": [
            {
              "text": "place de la République",
              "offset": 3854,
              "length": 22
            }
          ],
      "wikipediaLanguage": "en",
      "wikipediaId": "Place de la République",
      "wikipediaUrl": "https://en.wikipedia.org/wiki/Place_de_la_République",
      "bingId": "fc9a0c10-b5c5-3e2c-6dfc-b542877ab954"
    },
    {
      "name": "Bataclan (theatre)",
      "matches": [
        {
          "text": "Bataclan concert hall",
          "offset": 291,
          "length": 21
        },
        {
          "text": "Bataclan Concert Hall",
          "offset": 1496,
          "length": 21
        },
        {
          "text": "Bataclan Concert Hall",
          "offset": 555,
          "length": 21
        },
        {
          "text": "Bataclan Concert Hall",
          "offset": 1350,
          "length": 21
        },
        {
          "text": "Bataclan Concert Hall",
          "offset": 3521,
          "length": 21
        },
        {
          "text": "le Bataclan",
          "offset": 712,
          "length": 11
        }
      ],
      "wikipediaLanguage": "en",
          "wikipediaId": "Bataclan (theatre)",
          "wikipediaUrl":
             ↪ "https://en.wikipedia.org/wiki/Bataclan_(theatre)",
          "bingId": "53f5b775-997d-2ddc-18e1-275791e5175c"
        },
        {
          "name": "November 2015 Paris attacks",
          "matches": [
            {
              "text": "attacks in Paris",
              "offset": 4147,
              "length": 16
            },
            {
              "text": "Paris #Shootings",
              "offset": 3195,
              "length": 16
            },
            {
              "text": "Paris #shooting",
              "offset": 4969,
              "length": 15
            }
          ],
          "wikipediaLanguage": "en",
          "wikipediaId": "November 2015 Paris attacks",
          "wikipediaUrl":
             ↪ "https://en.wikipedia.org/wiki/November_2015_Paris_attacks",
          "bingId": "33323c91-5ad9-4401-850f-76bf651e80c1"
        },
        {
          "name": "Stade de France",
          "matches": [
            {
              "text": "Stade de France",
              "offset": 4862,
              "length": 15
            },
            {
              "text": "Stade de France",
              "offset": 981,
              "length": 15
            }
          ],
          "wikipediaLanguage": "en",
          "wikipediaId": "Stade de France",
          "wikipediaUrl": "https://en.wikipedia.org/wiki/Stade_de_France",
          "bingId": "85b86bf6-3481-0eb6-4b75-b42d46d3b957"
        },
        {
          "name": "Niggas in Paris",
          "matches": [
            {
              "text": "in #Paris",
              "offset": 3039,
              "length": 9
            },
            {
              "text": "in #Paris",
              "offset": 1518,
              "length": 9
            },
            {
              "text": "in #Paris",
              "offset": 1372,
              "length": 9
            },
            {
              "text": "in #Paris",
              "offset": 1234,
              "length": 9
            },
            {
              "text": "in #Paris",
              "offset": 4730,
              "length": 9
            },
            {
              "text": "in #Paris",
              "offset": 4044,
              "length": 9
            },
            {
              "text": "in #Paris",
              "offset": 2284,
              "length": 9
            },
            {
              "text": "in #Paris",
              "offset": 4566,
              "length": 9
            },
            {
              "text": "in #Paris",
              "offset": 3543,
              "length": 9
            },
            {
              "text": "in #Paris",
              "offset": 577,
              "length": 9
            },
            {
              "text": "in #Paris",
              "offset": 109,
              "length": 9
            },
            {
              "text": "in #Paris",
              "offset": 280,
              "length": 9
            },
            {
              "text": "in Paris",
              "offset": 46,
              "length": 8
            },
            {
              "text": "in Paris",
              "offset": 2478,
              "length": 8
            }
          ],
          "wikipediaLanguage": "en",
          "wikipediaId": "Niggas in Paris",
          "wikipediaUrl": "https://en.wikipedia.org/wiki/Niggas_in_Paris",
          "bingId": "ae541289-8f4a-aa45-a601-637fb41500e3"
        },
        {
          "name": "National Police (France)",
          "matches": [
            {
              "text": "Police",
              "offset": 783,
              "length": 6
            }
          ],
          "wikipediaLanguage": "en",
          "wikipediaId": "National Police (France)",
          "wikipediaUrl":
             ↪ "https://en.wikipedia.org/wiki/National_Police_(France)",
          "bingId": "0f752781-bb14-12f8-c26a-b07b1e0bdcf7"
        },
```

A.3.2 Microsoft Azure Text Analytics: Key Phrases

The result from the Microsoft Azure Text Analytics Key Phrases Analysis.

```
  "keyPhrases": [
        "Paris explosions",
        "Paris attacks",
        "Paris shockingRT",
        "Paris ht",
        "rising RT",
        "Paris Wypierdalaj razem z nimi",
        "Bataclan concert hall",
        "innocent people",
        "ParisRT",
        "dead",
        "BBC reports",
        "hostage situations",
        "SPECIAL REPORT",
        "French police report gunfire",
        "New info",
        "FoxNews",
        "CBSNews",
        "VIDEO",
        "rising fatality count",
        "Automatic gunfire",
        "terror attacks",
        "fusillade",
        "aftermath",
        "hostages",
        "FarhanKVirk",
        "shootings",
        "Figaro",
        "SHockridgeABC15 BREAKING",
        "https",
        "biggest terrorist attack",
        "thoughts",
        "shooting incident",
        "prayers",
        "CNN affiliate reporting",
        "death toll",
        "favorite cities",
        "fucking sick",
        "favourite cities",
        "file picture",
        "DespiertaEuropa",
        "RodrigoDdeV13",
        "Sportifsurcanap",
        "BostonGlobe",
        "villamatt1874",
        "colleagues",
        "head",
        "StadedeFrance",
        "DLittleSecret",
        "une date pourri",
        "dpatrikarakos",
        "maybachmsc",
        "BREAKINGNEWS Terrorists",
        "Vamos dar",
        "PorteOuverte",
        "stadium",
        "WajahatAli",
        "streets of central",
        "Bodies",
        "ParisFriday",
        "market",
        "place",
        "ParisAttacks",
        "shootout",
        "MsJulieLenarz",
        "frenchwords",
        "noise",
        "habituallychic",
        "russian",
        "Friends",
        "vtchakarova",
        "Anschlag",
        "France24",
        "SoSad",
        "tweet",
        "hashtag",
        "crowd",
        "horrifying images",
        "violence",
        "soccer match",
        "Conflic",
        "BataclanRT",
        "juliamacfarlane",
        "lives",
        "NFLFrance",
        "L'explosion",
        "maxcarver",
        "war zone",
        "world",
        "location",
        "PrayForParisThis",
        "terrorism",
        "ParisAtleast",
        "shit"
      ]
```

A.4 Keyword Frequency Script

This section contains the self-developed script, the Keyword Frequency Script. The script is developed using JavaScript and Node.js, described in 2.5.4.

```
  const fs = require("fs");
const util = require("util");

const readFile = util.promisify(fs.readFile);
const writeFile = util.promisify(fs.writeFile);

async function getData() {
  return await readFile("10000tweets.json");
}

getData().then(async data => {
  let filedata = JSON.parse(data);

  const words = filedata.tweets
    .map(x => x.text.split(" "))
    .reduce((x, y) => {
      return x.concat(y);
    }, []);

const wordsWithoutNewLine = words.map(x => (x.replace(/[\n\r]/g, ’’)));

    const answer = words.reduce( (countWords, word) => {
        countWords[word] = ++countWords[word] || 1;
        return countWords;
    }, {});

  for(let b in answer){
      if(answer[b] > 300 && b.length > 3)
        console.log(b, answer[b]);
  }
});
```