UTRECHT UNIVERSITY

MASTER THESIS

Early warning of incidents during protest demonstrations using Twitter

Author:
Joren WOUTERS

Thesis supervisor:
Prof. dr. Remco VELTKAMP

Daily thesis supervisor:
Laurens MÜTER MSc

Second thesis supervisor:
Dr. Matthieu BRINKHUIS

A thesis submitted in fulfillment of the requirements
for the degree of Master of Science

in the

Master program Business Informatics
Department of Information and Computing Sciences

Utrecht University

Thursday 15th July, 2021

UTRECHT UNIVERSITY

Abstract

Master program Business Informatics
Department of Information and Computing Sciences

Master of Science

Early warning of incidents during protest demonstrations using Twitter

by Joren WOUTERS

While most protest demonstrations remain peaceful, some of them become violent, resulting in possibly dangerous situations. Because there is a positive correlation between social media use and protest participation, we propose to analyze incidents during protest demonstrations using social media data. In this study, a Twitter dataset is collected related to a Dutch demonstration where protesters did not comply with COVID-19 rules of the government. Following, an exploratory data analysis is performed to identify the phases of Twitter coverage after an incident during a protest demonstration. Additionally, machine learning models are trained to distinguish incident-related from non incident-related tweets. Furthermore, analysts at the Dutch national police force are interviewed to identify the information need when automatically detecting incidents during protest demonstrations. Lastly, an early warning system is created that automatically extracts tweets and detects incidents during protest demonstrations. Findings show four phases of Twitter coverage can be identified after an incident during a protest demonstration, Support Vector Machines (SVM) perform best in distinguishing incident-related from non incident-related tweets and analysts at the Dutch national police force want to obtain incident information as soon as possible. The developed system was able to detect incidents during a protest demonstration by using Twitter data, but could be improved.

Acknowledgements

I would like to thank various people for their contribution to this master thesis:

- Prof. dr. Remco Veltkamp for his pleasant supervision of this research project. His insights and especially the structure of the project helped me bring my work to a higher level.
- Dr. Matthieu Brinkhuis for his insightful feedback during this research project and his oversight as my second supervisor.
- Laurens Müter for his daily supervision of this research project. His to-the-point feedback and criticism provided me with useful insights.
- Marius Kok for structuring the collaboration with the Dutch national police force.
- Bianca de Vree and Joost Boerboom for helping with labeling the tweets dataset.
- Jacco van der Plaat for his help when dealing with technical issues.

Joren Wouters - 15th July, 2021

Contents

Preface

This report is the result of an MBI thesis project under the supervision of Prof. dr. Remco Veltkamp, Dr. Matthieu Brinkhuis and Laurens Müter MSc. The project is performed at Utrecht University and is part of the National Police Lab AI, a collaboration between Utrecht University, University of Amsterdam and the Dutch Police.

Chapter 1

Introduction

In this age, social movements are becoming increasingly dependent on the use of social media (Isa & Himelboim, 2018), as they try to achieve social change (Harlow, 2012). This social change can be accomplished by collective action, such as organizing a protest demonstration. Whereas social movements use social media (such as Facebook and Twitter) to organize and disseminate information regarding protests (Isa & Himelboim, 2018), current research also indicates that there is a positive correlation between social media use and participation in protest demonstrations.

While most protest demonstrations remain peaceful, some of them become violent, resulting in possibly dangerous situations, such as riots (ACLED, 2020) and clashes with the police (Enikolopov, Makarin, & Petrova, 2020). Armed Conflict Location & Event Data Project (ACLED) and the Bridging Divides Initiative (BDI) at Princeton University even declared that the United States is in a state of crisis, because of rising political violence (ACLED, 2020). Therefore, it becomes important to understand when incidents during protest demonstrations are happening and how law enforcement could be timely warned in order to prevent dangerous situations. Because of the reliance of social movements on social media and the correlation between social media use and protest demonstration participation, we propose to use Twitter data to perform this analysis.

Current research covers how social movements use social media and even the prediction of protests by using social media data. However, it remains unclear how incidents during protest demonstrations are covered on Twitter over time. Moreover, current research lacks the early warning of incidents during protest demonstrations by using Twitter data. Therefore, the objective of this research is to **automatically detect incidents during protest demonstrations by using Twitter data**.

In order to achieve this goal, a Twitter dataset is collected related to a Dutch protest demonstration where participants were not compliant with the COVID-19 rules. Following, an exploratory data analysis is performed to identify the phases of Twitter coverage after an incident during a protest demonstration. Additionally, several machine learning models are trained on the data to distinguish incident-related from non incident-related tweets. Furthermore, interviews with analysts of the Dutch national police force are conducted to understand the information need when automatically detecting incidents using Twitter data. Finally, a system is created that automatically extracts tweets from Twitter, detects incidents and sends a warning when an incident is detected.

Chapter 2

Related work

In this section, previous research on the early warning of protest demonstrations is described. First, we will describe how protest demonstrations use social media and the need for early warning of incidents during protest demonstrations. Then, we will discuss the broad social media landscape. Additionally, a short description of Twitter will be provided and opportunities for using Twitter as a rich source of information for data analysis will be identified. Moreover, we will discuss how Twitter can be utilized in order to detect events, classify them as incidents and automatically extract information from Twitter. Lastly, the gaps in current literature will be identified.

2.1 The need for early warning of incidents during protest demonstrations

Information technology and, more precisely, social media is becoming more important for social activism (Sandoval-Almazan & Gil-Garcia, 2014). Social movements are even becoming increasingly dependent on the use of social media (Isa & Himelboim, 2018). Therefore, it becomes interesting to analyze social movements by analyzing data from social media, and more precisely for this study, from Twitter. First, we will provide a description of social movements, their goals and how they act in combination with the Internet. Then, we will focus on protests that are using social media and how they develop over time. Ultimately, we will center around protest demonstrations and the need for research of early warning of incidents during protest demonstrations.

2.1.1 Social movements

A social movement is a network of informal interactions between a plurality of individuals, groups and/or organizations, engaged in a political or cultural conflict, on the basis of a shared collective identity (Diani, 1992). The end goal of a social movement is achieving some kind of social change (Harlow, 2012), which can be accomplished by collective action, such as a protest demonstration or a petition campaign.

In order to achieve social change, it is necessary that a social movement obtains enough highly motivated individuals to initiate a mobilization, and attract more participants and resources (Harlow, 2012). To drive mobilization and enlarge social movements, social movements can try to create a spillover effect. A spillover effect is the result or the effect of something that has spread to other situations or places (Oxford Dictionary, n.d.-b), and is created when one social movement is joined by the other, positively influencing the social movement (Meyer & Whittier, 1994). For

example, social movements often create spillover effects by using two or more hashtags together (on Twitter) that represent different issues or movements (Isa & Himelboim, 2018). Moreover, Tremayne (2014) find that the #FuckYouWashington movement helped to spread the concept of Occupy Wall Street through Twitter, whereas a spillover effect was created.

Because of the reliance of social movements on social media, it becomes important to understand how social media and collective action intertwine with one another. Van Laer and Van Aelst (2010) distinguished two main types of collective action in combination with the Internet: Internet-based action and internet-supported action. Internet-based action refers to activities that exist only because of the internet, such as email bombing and hacktivism. An example of an internet-based action is the hacking of various payment processors by hacking group Anonymous (Singel, 2010). Various payment processors, such as Paypal, Mastercard and Visa had cut off Wikileaks in 2010 because it violated their "terms of service" agreements. In response, Anonymous started 'Operation Payback' (The Guardian, 2010) flooding the sites' servers with traffic, leading to inaccessibility of the Mastercard website (also called a DDOS attack). The other type of action is Internet-supported action, which refers to the traditional tools of social movements that have become easier to organize and coordinate because of the internet, such as demonstrations and occupations. An example of an internet-supported action is the Occupy Wall Street demonstration. On September 17, 2011, thousands of people started a protest in New York demanding a need for a systemic change in the financial world (Ranney, 2014). They claimed to represent the 99% of the American population that was being taken advantage of by the wealthiest 1%, represented by large banks and corporations (with headquarters on Wall Street). Because of extensive traditional media coverage and mobilization through social media, the movement grew quickly, eventually spreading from New York to major cities all over the world.

2.1.2 Social protest

Although social media was not the only reason why the Occupy Wall Street movement got so much traction, scholars have clearly stated that social networking sites, such as Facebook and Twitter, have played an important role in the diffusion of the movement (Tremayne, 2014; Gaby & Caren, 2012; Suh, Vasi, & Chang, 2017). And Occupy Wall Street is not alone, current research has covered many examples of protest demonstrations where social media played an important role, including examples as the Egyptian revolution in 2011 (Attia, Aziz, Friedman, & Elhusseiny, 2011), the Black Lives Matter movement (Edrington & Lee, 2018) and Iran's Green Movement (Ansari, 2012). Regarding such social protests, Sandoval-Almazan and Gil-Garcia (2014) proposed a four-stage model to identify the levels of maturity and development cycle of protests using social media technologies. The individual stages are: Triggering event (1), Media response (2), Viral organization (3) and Physical response (4). Each of the stages is complementary and they follow each other in an imperfect and not totally predictable cycle. In the following paragraphs, we will briefly describe each of the stages in detail.

The first stage of a protest is the *triggering event* (Sandoval-Almazan & Gil-Garcia, 2014). A triggering event is an extraordinary event that promotes a social reaction to it, with the following characteristics: it breaks the status quo of the society, it is autonomous and citizens organize around it. The precise cause of the event is irrelevant, as long as the result is a social reaction. Moreover, this event creates synergy between the new media (i.e. social media) and traditional media (i.e. newspapers).

An example of a triggering event, is the self-immolation of Mohamed Bouazizi (a college-educated street vendor), because he was in despair over corruption and joblessness in Tunisia. An even more recent example is the death of George Floyd on 25 May 2020, when a police offer knelt on his neck for around 8 minutes until he could not breathe anymore, resulting in thousands of Black Live Matters protests in the United States (ACLED, 2020).

The triggering event creates an immediate response, resulting in the second stage: *media response* (Sandoval-Almazan & Gil-Garcia, 2014). In response to the triggering event, citizens are going to share, collaborate and cooperate using social media technologies, which fosters information aggregation for the activists and promotes a second or third information cascade, allowing late activists to join the movement. Therefore, social media can be helpful in three ways: rapidly mobilizing protesters, undermining a regime's legitimacy and increasing national and international exposure to a regime's atrocities. In addition to the activity on social media, electronic media journalism publishes information on their normal channels, such as TV, radio stations and newspapers.

Because of this mass reaction, the group starts building an online community, with efficient communication, an encrypted language (with common words and concepts) and shared ideas of co-production and collaboration, leading to the third stage: *Viral organization* (Sandoval-Almazan & Gil-Garcia, 2014). The movement creates a collective identity, gives new names to problems, pressures the government over formal channels and builds a discourse and consistent message. This viral organization influences two forms of mobilization: online mobilization (also called cyberactivism or above referred to as Internet-based action) and offline mobilization (above referred to as Internet-supported action), which requires management, consistency and strategy for the movement's discourse. For example, two months before the first Occupy Wall Street protest in September 2011, activists were using Twitter to organize and spread the movement (Tremayne, 2014).

The purpose of the last stage, *physical response*, is to place the protest in the physical world, which shows the power and strength of the social protest (Sandoval-Almazan & Gil-Garcia, 2014). By using technology and street demonstrations simultaneously, the movement can create a physical response and organize resistance, which shows the power of the organization to new activists and encourages them to promote and duplicate the movement.

The proposed model of Sandoval-Almazan and Gil-Garcia (2014) gives a good overview of the stages of a social protest, but also contains some limitations. First of all, the cycle is dependent on the influence of traditional media (Sandoval-Almazan & Gil-Garcia, 2014), meaning that when the triggering event does not have sufficient importance to escalate to mass media, it is difficult for the movement to gain attention. Secondly, the cycle implies the development of a critical mass, so that there is an online response strong enough to share the message, create a threat or maintain a protest. Without this critical mass, the cycle would not start, but it is not clear if that is the case in all situations. Moreover, the triggering event is ambiguous and difficult to assess or predict. Despite these limitations, the framework provides sufficient insights into how social protests using social media technologies develop over time and end in a physical response, such as a protest demonstration.

2.1.3 Protest demonstrations

A protest demonstration is a public meeting or a march at which people show that they are protesting against or supporting somebody/something (Oxford Dictionary,

n.d.-a). Famous examples of protest demonstrations that have actively used social media include the Occupy Wall Street demonstration in 2011 (described above), Iran's Green Movement in 2009 (where people demonstrated against a disputed election) (Ansari, 2012) and the Gezi Park protests in 2013 (where Turkish people demonstrated to protect trees) (Demirhan, 2014).

These protest demonstrations provide a strong indication that social media correlates with increased participation in protest demonstrations. However, there was no hard evidence for this correlation and recent research often theorized about whether social media promoted protest demonstration participation (Edmond, 2013; Little, 2016). Other literature provided examples where protesters joined demonstrations after it was shared on social media, such as the demonstration in Madrid in 2011 (Gerbaudo, 2012), but no sources were provided for these kinds of examples.

In order to fill this gap in existing literature, Boulianne (2015) performed a meta-analysis of social media use and participation in civic and political life. In this meta-analysis, Boulianne also specifically focused on the correlation between social media and protest activities. Findings show that there is a positive effect between social media use and participation in protest activities, but it also showed there is not always a significant correlation. However, as Boulianne already pointed out, the bulk of research uses composite indexes that combine very different activities, such as including core protest activities (e.g. participating in a protest demonstration) and other types of activities (e.g. talking to public officials) in one index. As a result, it is almost impossible to determine the true effect of social media use on protest demonstration participation.

In response to this research, Lee (2018) examined the role of social media in South Korea during the 2016 "Choi Park" scandal protests. In this case, the protesters protested against a scandal of the shadow president in South Korea, which resulted in massive protests in the form of candlelight vigils. The first candlelight vigil was relatively small (with 20.000 participants) but grew quickly to larger protests. This eventually led to a mega-protest on December 3 lead with 2.3 million protesters. A few days after the eight candlelight vigil, Lee surveyed 922 protest participants, asking about their overall Facebook use, content consumption on Facebook and political expression on Facebook. Results showed that the frequency of using Facebook was strongly positively associated with protest activity.

Even more recently, Enikolopov et al. (2020) studied the correlation between social media and protest participation in Russia. In 2011-2012, a wave of protest demonstrations started in Russia because of electoral fraud during the parliamentary elections of 2011. Because traditional media was largely controlled by the state, online social networks (such as VK - the Russian variant of Facebook) became an important source of political information. Using data from 625 Russian cities with populations over 20,000 people, Enikolopov et al. hand-collected data on protests that occurred between December 2011 and May 2012. Results show that the number of VK users in a city had a positive and statistically significant effect on the probability that a protest occurs: A 10% increase in the number of VK users in a city leads to a 4.5–4.8 percentage points higher probability of a protest being organized. Moreover, the results indicated that a 10% increase in the number of VK users leads to a 19% increase in the number of protesters.

These results indicate that there is a positive correlation between social media use and participation in protest demonstrations, which implies that social media could act as an important information disseminator for protest demonstrations. Concurrently, social movements use Facebook and Twitter to organize and disseminate information regarding protests (Isa & Himelboim, 2018). According to Howard et al.

(2011), one protester stated "We use Facebook to schedule the protest, Twitter to coordinate, and YouTube to tell the word".

While most protest demonstrations remain peaceful, some of them become violent, resulting in possibly dangerous situations, such as riots (ACLED, 2020) and clashes with the police (Enikolopov et al., 2020). Therefore, it becomes interesting to understand when these protest demonstrations are escalating and how law enforcement could be timely warned in order to prevent dangerous situations. Current research covers how social movements use social media and even the prediction of protests by using social media data (Bahrami, Findik, Bozkaya, & Balcisoy, 2018), but lacks the early warning of incidents during protest demonstrations. Hence, we emphasize the need for early warning systems that automatically detect incidents during protest demonstrations. Because social media plays an important role in information dissemination of protest demonstrations, we intend to use social media data (and more specifically, Twitter data) to perform this analysis.

Moreover, one could argue that this subject could not be more accurate than right now in today's society. Armed Conflict Location & Event Data Project (ACLED) and the Bridging Divides Initiative (BDI) at Princeton University even declared the United States in a state of crisis (ACLED, 2020). They initiated a joint project called the US Crisis monitor, collecting real-time data on political violence in the United States. After the death of George Floyd on 25 May 2020, the Black Lives Matter (BLM) movement quickly spread from Minneapolis throughout the country, resulting in over 7,750 demonstrations across 2,440 locations. Of all these demonstrations, around 220 locations became violent. Additionally, dozens of car-ramming attacks by individual perpetrators (not associated with the BLM movement) have been reported at demonstrations around the country.

Furthermore, government services are still not utilizing Twitter (and Twitter data) to the full extent, despite all the ongoing research to event and incident detection on Twitter. Already in 2012, Terpstra, Stronkman, de Vries, and Paradies noted that during the Pukkelpop storm (in Belgium), no official authorities used Twitter to interact with possible victims. And even more recently, the Dutch government released an official report on public demonstrations, stating that the possibilities of social media analysis are limited, because of the privacy of users and the speed of information on social networks (Rijksoverheid, 2020).

2.2 Social media

Over the last two decades, dozens of social media platforms have arisen, influencing the way how people communicate with each other (Mihailidis, 2014). In this period, these social media platforms have grown constantly (Ortiz-Ospina, 2019) and have been used all over the world (Hootsuite, 2020a). As of 2020, this has resulted in 3.9 billion active social media users (Hootsuite, 2020b) who are responsible for thousands of interactions worldwide every minute (Forbes, 2020).

Social media builds upon two concepts: Web 2.0 and User-Generated content. The term *Web 2.0* first arose in 2004 in a brain-storming session (O'reilly, 2009). A precise definition of Web 2.0 has been found elusive (Cormode & Krishnamurthy, 2008), but it describes a new way in which software developers and end-users started to utilize the World-Wide-Web (Kaplan & Haenlein, 2010), which is in contrast with Web 1.0. The primary difference between Web 1.0 and Web 2.0 is the role of content creators and content consumers. In Web 1.0 there were only a few content creators and a vast majority of content consumers. This is in contrast with Web 2.0, where

any user can be a content creator and numerous technological aids have been created to maximize the potential for content creation.

Additionally, social media builds on the concept of User-Generated Content. User-Generated Content refers to all the ways in which people make use of social media (Kaplan & Haenlein, 2010). The term is usually applied to describe the various forms of media content that are publicly available and created by end-users, such as images, videos and blogs. According to the OECD (2007), User-Generated Content must fit three requirements:

1. *Publication requirement*. The content should be published on a publicly accessible website or a page on a social networking site only accessible to a select group of people.
2. *Creative effort*. This means that a certain amount of creative effort was put into creating the work or adapting existing works to construct a new one.
3. *Creation outside of professional routines and practices*. This refers to the fact that the content should be created outside of professional routines and practices.

Building upon these concepts, Kaplan and Haenlein (2010) defined social media as a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of User Generated Content. Within this broad definition, various types of social media platforms can be identified. Kaplan and Haenlein classified these platforms alongside two dimensions: The level of social presence/media richness and the level of self-presentation/self-disclosure.

Social presence is defined as the acoustic, visual, and physical contact that can be achieved between two communication partners (Kaplan & Haenlein, 2010). The higher the social presence of a platform, the larger the social influence the communication partners have on each other's behavior. Closely related to social presence is the notion of media richness. According to media richness theory, the goal of communication is the resolution of ambiguity and reduction of uncertainty (Daft & Lengel, 1986). Because media differ in the degree of richness they provide (the amount of information they allow to be transmitted in a given time interval), some media are more effective than others in resolving ambiguity and uncertainty.

Additionally, self-presentation relates to the fact that in any social interaction people have the desire to control the impressions other people form of them (Goffman, 1949). Often, such self-presentation is performed through self-disclosure, defined as the conscious or unconscious revelation of personal information that is consistent with the image one would like to give (Kaplan & Haenlein, 2010).

Based on these two dimensions, Kaplan and Haenlein (2010) classified social media platforms in six categories:

1. *Collaborative projects*. Collaborative projects enable the joint and simultaneous creation of content by many end-users.
2. *Blogs*. Special types of websites that usually display date-stamped entries in reverse chronological order (OECD, 2007).
3. *Content communities*. Applications where the main objective is the sharing of media content between users.
4. *Social networking sites*. Applications that enable users to connect by creating personal information profiles, inviting friends and colleagues to have access to those profiles, and sending e-mails and instant messages between each other.

		Social presence/ Media richness		
		Low	Medium	High
Self-presentation/ Self-disclosure	High	Blogs	Social networking sites (e.g., Facebook)	Virtual social worlds (e.g., Second Life)
	Low	Collaborative projects (e.g., Wikipedia)	Content communities (e.g., YouTube)	Virtual game worlds (e.g., World of Warcraft)

TABLE 2.1: Classification of Social Media by social presence/media richness and self-presentation/self-disclosure according to Kaplan and Haenlein (2010).

5. *Virtual game worlds*. A three-dimensional environment where users appear as avatars and interact with each other as they would in real life, requiring to behave to strict rules in the context of a massively multiplayer online role-playing game.

6. *Virtual social worlds*. Three-dimensional environments where users appear as avatars and interact with each other as they would in real life, allowing users to choose their behavior more freely and essentially live a virtual life similar to their real life.

These different social media platform types, alongside with their level of social presence/media richness and the level of self-presentation/self-disclosure, are presented in Table 2.1.

With half of the world using social media (Hootsuite, 2020b), it becomes important to understand why people are using it. Research by Whiting and Williams (2013) revealed that people primarily use social media because of seven reasons:

- *Social interaction*, meaning that people communicate and interact with others.
- *Information seeking*, referring to use social media for seeking out information or self-education.
- *Pass time*, defined as using social media to occupy time and relieve boredom.
- *Entertainment*, where people use social media to provide entertainment and enjoyment.
- *Relaxation*, defined as using social media to relieve day-to-day stress.
- *Communicatory utility*, the use of social media for communication facilitation and providing information to share with others.
- *Convenience utility*, defined as providing convenience or usefulness to individuals.

Because social media is used by many people and is used for a variety of purposes, it becomes an important source of online interactions and contents sharing (Adedoyin-Olowe, Gaber, & Stahl, 2013). These online interactions include text, reviews, blogs, discussions, remarks and reactions that contain subjectivity, assessments, approaches, evaluations, observations, feelings and sentiment expressions. As a result, many organizations, governments and individuals follow the activity

on social media, because it allows them to obtain knowledge on how their audience reacts to postings. Moreover, current research has proven that social media can be used as a rich source of information for data analysis (He et al., 2016; Wang & Gan, 2017; Bovet & Makse, 2019; Terpstra et al., 2012).

Social networking sites, such as Twitter, enable the collection of large-scale data, but also give rise to major computational challenges (Adedoyin-Olowe et al., 2013), often referred to as the 3 Vs of Big Data: Volume, Velocity and Variety (Russom et al., 2011). Volume refers to the large amounts of data, Velocity considers the speed at which the data is generated and Variety refers to the variety of data formats. Despite these computational challenges, data mining techniques have made it possible to quantitatively analyze social network data to discover valuable, accurate and useful knowledge.

2.3 Twitter

Twitter is a service that allows users to send and receive short messages, called tweets. As of 2020, Twitter has an estimated user base of 340 million users (Hootsuite, 2020a) and has grown constantly over the last decade (Business of Apps, 2020). In current literature, there is no consensus regarding the classification of Twitter. Some scholars describe Twitter as a micro-blogging platform (Lasorsa, Lewis, & Holton, 2012; Gleason, 2013), while others describe Twitter as a social networking site (Hwang & Kim, 2015). Following the classification of Kaplan and Haenlein (2010), we argue that Twitter has a low to medium social presence/media richness and a high self-presentation/self-disclosure level. Therefore, Twitter is a combination of a blogging platform and a social networking site, providing users with both the ability to follow date-stamped entries in reverse chronological order and to connect with other users, by creating personal information profiles, inviting friends and colleagues, and sending instant messages between each other.

On Twitter, anyone can create a user profile and provide information such as their name, a description, profile picture and a link to a website. Additionally, Twitter users can follow other users resulting in updates of the latest activity on the users that they follow. An example of a Twitter profile is presented in Figure 2.1.

Furthermore, Twitter uses can 'tweet' about any topic with a maximum of 280 characters and each tweet can consist of a text (including links to other websites), hashtags, mentioned users and a shared location. Moreover, other users can like the tweet, comment on it or re-share it (called a *retweet*). An example of a tweet is presented in Figure 2.2.

Moreover, Twitter has a flat and flexible communicative structure: users interested in specific topics can easily find them through hashtags (Bruns & Liang, 2012). Hashtags are keywords prefixed with the hash symbol '#', which users can include in their tweets to make their tweets visible to others following the hashtag. Also, Twitter is open, meaning that non-registered users can follow these hashtags streams using the Twitter website (Bruns & Burgess, 2011). In addition, the simple network structure of Twitter enables the wide sharing of topically relevant tweets from public accounts. This is in contrast with other social media platforms, such as Facebook, where there are more complex visibility permissions and messages usually will not travel far beyond a user's immediate circle of friends, or friends of friends.

Furthermore, Twitter is used for various purposes. Research by Java, Song, Finin, and Tseng (2007) indicated that users primarily have four intentions when using Twitter. First of all, the largest and most common use of Twitter is to talk about daily

FIGURE 2.1: Example of a Twitter profile (Trump, n.d.-b)

FIGURE 2.2: Example of a tweet by Donald Trump (Trump, n.d.-a)

routine or what people are currently doing. Secondly, Twitter is used for having conversations with other users. Thirdly, people use Twitter to share information or URLs with other users. Lastly, users utilize Twitter to report news or comment about currents events on Twitter. Moreover, I. L. Liu, Cheung, and Lee (2010) find that Twitter fulfills users' needs for self-documentation, information sharing, medium

appeal (ubiquitous accessibility of Twitter) and convenience. Additionally, Twitter can be used for professional purposes. News organizations use Twitter to disseminate information (Armstrong & Gao, 2010), businesses can use Twitter to gain business value (Culnan, McHugh, & Zubillaga, 2010) and researchers use Twitter to share and acquire educational resources (Carpenter & Krutka, 2014).

Moreover, Twitter has proven itself to be a rich source of information for data analysis in various domains. First and foremost, businesses can explore Twitter data and mine it for business intelligence (Lu, Wang, & Maciejewski, 2014). Business intelligence is the process of transforming raw data into useful information for more effective strategic, operational insights and decision-making purposes so that it yields real business benefits (Duan & Da Xu, 2012). One field of interest for business intelligence is revenue prediction, which can be performed by analyzing social media to understand product adoption and sentiment. For example, He et al. (2016) mined competitive intelligence by comparing consumer opinions and sales performance of two competitors (Apple and Samsung) based on publicly available Twitter data. They collected 229,948 tweets mentioning the iPhone 6 or Galaxy S5 for a period of four months after the release of the iPhone 6. By using opinion mining and sentiment analysis, they analyzed the differences in volume, (purchase) intention and sentiment between the market leader (Apple) and one of its competitors (Samsung). Based on the Twitter data, they estimated the amount of sold phones by multiplying the mean volume of daily tweets by the purchasing intention score. They find that this indicator (3.96) was very similar to the shipment gap of sold phones (4.04), thereby showing that publicly available Twitter data can be used as a source for competitive intelligence.

Elections are another domain that have effectively used Twitter data. A main field of research regarding elections is the use of social media data to predict election results (Anstead & O'Loughlin, 2015). For example, Wang and Gan (2017) tried to predict the election results of the French 2017 election. By using a popularity estimator on election-related tweets, the popularity of the two candidates (Macron and Le Pen) was estimated. The estimator calculates the popularity of a candidate, based on the number of positive, negative and neutral tweets about the candidate. Using this estimator, the final result of the French election was predicted and resulted in only a 2% difference from the real voting results. This indicates that Twitter data might be a reliable predictor of election results. Furthermore, in recent years, there has been a rise of bots and disinformation (also referred to as "fake news") on social media in the context of political propaganda (Ferrara, 2017). As a result, fake news detection on social media has recently become an emerging research topic that is attracting tremendous attention (Shu, Sliva, Wang, Tang, & Liu, 2017). An example of such research, is the research by Bovet and Makse (2019), focused on the US 2016 election. They collected over 171 million tweets over approximately 5 months mentioning the two top candidates (Donald Trump and Hillary Clinton). Using domain-level classification, they analyzed 30 million tweets from 2.2 million users linking to news outlets. Based on their analysis, they find that 25% of these spread tweets were either containing fake or extremely biased news.

Another domain that benefits from the analysis of Twitter data are emergencies. Social media platforms provide active communication channels during emergencies, such as disasters by natural hazards (Imran, Castillo, Diaz, & Vieweg, 2015). These kinds of crises generate a situation that is full of questions, uncertainties and the need to make quick decisions, often with minimal information. Therefore, the automatic extraction of useful information based on publicly available social media data is especially interesting for first responders and decision-makers to gain

insights into the situation as it unfolds. For example, Terpstra et al. (2012) analyzed 97,0000 tweets that were published shortly before, during and after a storm hit the Pukkelpop 2011 festival in Belgium. When the storm hit the festival, the tweet activity increased exponentially, peaking at 576 tweets per minute. Results showed that festival-goers were surprised by the storm and eventually shared many tweets containing uploaded pictures and videos of damages to the festival. As a consequence, this study showed clear opportunities for two-way crisis communication between authorities, media and citizens. For example, Twitter crisis managers could interact with citizens and media by confirming or refuting rumors and by taking emotional responses into account in their crisis communication.

The aforementioned examples show that Twitter can be used as a rich source of information for data analysis in various domains. Therefore, in this research project, we attempt to use Twitter data for the early warning of incidents during protest demonstrations.

2.4 Early warning techniques using Twitter

An early warning system for incidents using Twitter generally uses three methods: the detection of an event, the classification of whether this event is related to an incident and automatic information extraction from Twitter data so that law enforcement can act on it (Imran et al., 2015). In this section, we will first discuss the timeline of incident reporting on Twitter and then discuss each of the different methods used in early warning.

2.4.1 Timeline of incident reporting on Twitter

According to Klein, Laiseca, Casado-Mansilla, López-de Ipiña, and Nespral (2012), the Twitter reporting process of an incident can be divided into three phases, as presented in Figure 2.3. First, several witnesses individually report an incident on Twitter. Then, in the second phase, the followers of these first-time reporters will spread this information on Twitter. Finally, mass media will cover the incident, usually several hours after the incident has occurred.

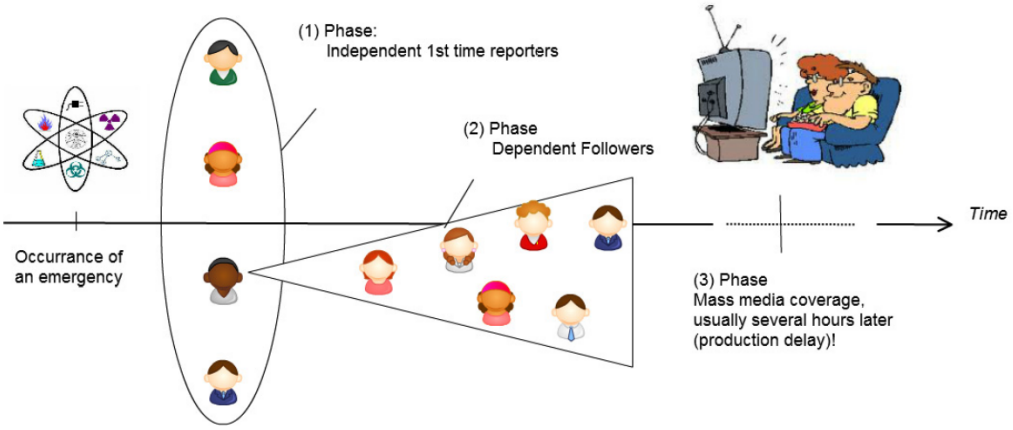

FIGURE 2.3: Incident reporting timeline by Klein et al. (2012)

Although the model by Klein et al. (2012) is minimal, it seems to have an overlap with other literature. Hu et al. (2012) focused on Twitter coverage after the news of

Osama Bin Laden's death leaked through Twitter. They find that the news was first spread by people from the media. Subsequently, mass media cover the news with reports, resulting in more mentions and the second phase of coverage. Lastly, celebrities use their social influence to help spread the news and stimulate discussions. These three phases of Twitter coverage are presented in Figure 2.4. However, Hu et al. did not provide definitions for "media people", "mass media" and "celebrities", and no explanation was provided for how these labels were determined. Moreover, Hu et al. only took the 100 most mentioned users into account.

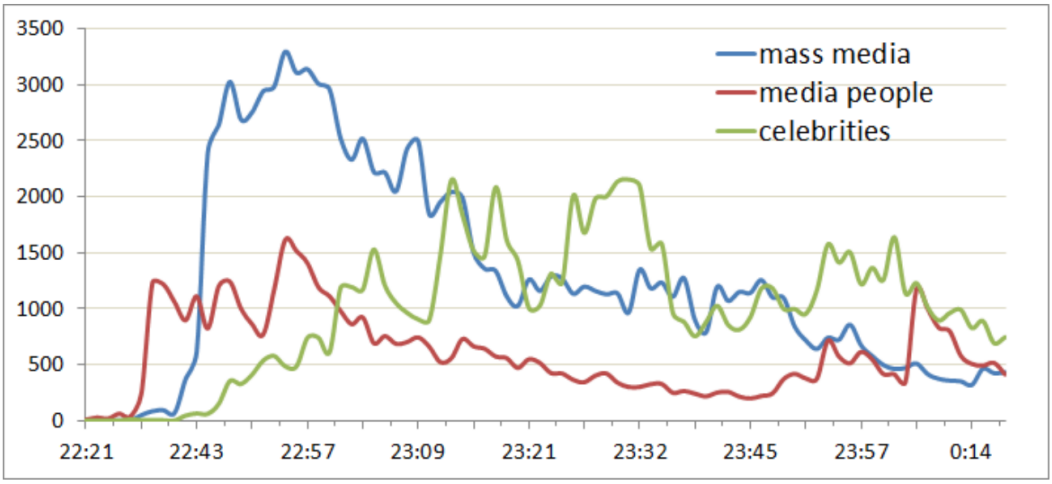

FIGURE 2.4: Three phases of Twitter coverage after Osama Bin Laden's death leaked to Twitter, in number of tweets per minute (Hu et al., 2012)

Although Klein et al. (2012) did not provide any empirical evidence for their model and Hu et al. (2012) only provided evidence for one event, these studies provide some insight into how incidents and events are covered on Twitter over time.

2.4.2 Event detection

An event on social media is a way of referring to an observable activity at a certain time and place that involves or affects a group of people in a social network (Saeed et al., 2019). By systematically analyzing the content published on Twitter, such events can be detected, which is referred to as event detection.

Following, the different forms of event detection will be discussed, namely unspecified versus specified event detection and new versus retrospective event detection. Thereafter, several event detection methods will be described.

Unspecified and specified event detection

The first distinction in current literature regarding event detection is the presence of prior information about the event (Saeed et al., 2019; Atefeh & Khreich, 2015). When an event is already known or planned, processing data concerning known information (such as location, time, keywords and users) to detect events is called specified event detection (SED). An example of specified event detection is research by Robinson, Power, and Cameron (2013), focused on detecting earthquakes using Twitter. The earthquake detector checks for the keywords "earthquak" and "#eqnz" in the Twitter stream and applies burst detection by observing features in fixed time

windows against historical word frequencies. New earthquakes are detected when the observed frequencies are much higher than usual word frequencies in the past.

In contrast, when one wants to detect an event without prior information, this is referred to as unspecified event detection (UED) (Saeed et al., 2019). Unknown events are typically driven by emerging events, breaking news and general topics that attract the attention of Twitter users (Atefeh & Khreich, 2015). They are typically detected by exploiting the temporal patterns of the Twitter stream by monitoring bursts in keywords and concepts that highlight events. For example, Mathioudakis and Koudas (2010) introduced TwitterMonitor, a system that detects trends on Twitter. First, the Twitter stream is monitored to identify 'bursty keywords', keywords that suddenly appear in tweets at an unusually high rate. Subsequently, related keywords are grouped and related to a specific trend. Once a trend is identified, TwitterMonitor attempts to compose a more accurate description of it by incorporating context extraction algorithms over the history of the trend and keywords that are correlated with it.

New and retrospective event detection

Event detection can also be classified depending on the task at hand and the type of the event (Atefeh & Khreich, 2015). New event detection (NED) refers to the continuous monitoring of the Twitter stream for discovering new events in real-time. NED is typically suitable for detecting unknown real-world events or breaking news. Additionally, retrospective event detection (RED) involves the task of detecting events from historical data (Saeed et al., 2019). Historical data can be clustered or classified to detect significant events that happened in the past.

Event detection methods

In current literature, there are two common approaches to detecting events on Twitter: keyword-burst approaches and location-burst approaches. Keyword-burst approaches assume that word frequencies related to the event increase over time (Imran et al., 2015). The observed keyword frequencies are compared with historical keyword frequencies and if there is a significant increase in the frequency of the keyword, an event is detected. For example, Marcus et al. (2011) introduced *Twitinfo*, a system for detecting, summarizing and visualizing events on Twitter. Twitinfo collects tweets based on a search query and bins the number of tweets per time window (e.g. 5 minutes). Automatically, an exponentially weighted moving average of several time windows is calculated and if the number of tweets in the next time window is significantly higher, an event is detected.

A similar approach to the keyword-burst approach is the location-burst approach, which is often used in research on detecting traffic incidents. Instead of looking at the frequency of specific keywords in tweets, it monitors the number of tweets in a certain geographical region. When the number of tweets in a specific region significantly increases, an event is detected. For instance, research by Xu, Li, Wen, and Huang (2019) used a location-burst approach and was focused on detecting traffic incidents in Toronto using Twitter. In Figure 2.5, it is demonstrated that most of the tweets related to traffic incidents were sent in downtown areas of Toronto.

Additionally, some scholars go beyond burst approaches and also take social features, topical features and Twitter-centric features into account, as used in research by Becker, Naaman, and Gravano (2011) focused on distinguishing tweets about real-world events and non-events. First of all, they captured the interaction of users

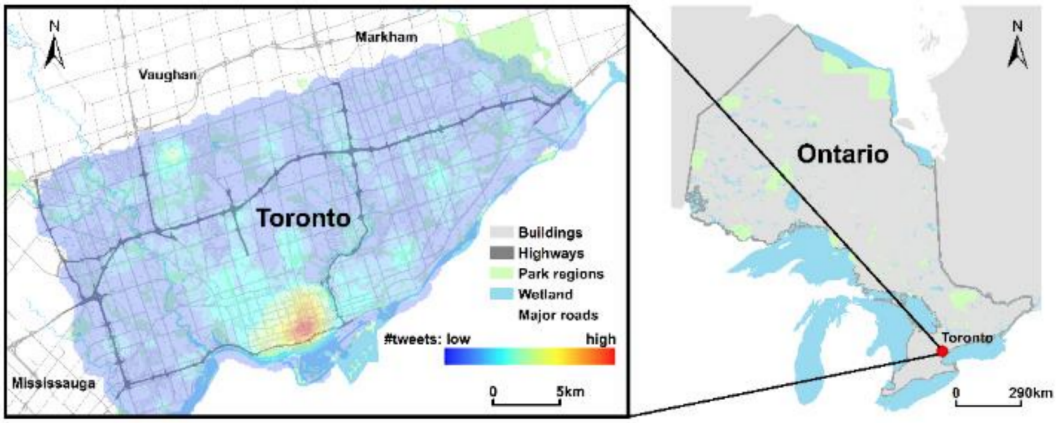

FIGURE 2.5: Geographical map of Toronto showing the number of tweets as reported in Xu et al. (2019).

(such as retweets, replies and mentions) to possibly identify differences between events and non-events messages. Secondly, they used topical features (which describe the coherence of a topic), based on the hypothesis that event clusters tend to revolve around a central topic. Additionally, they captured the use of Twitter-centric features, such as the percentage of tweets that use hashtags and the percentage of tweets that contained the most frequently used tag.

Moreover, some studies do not only use clustering but also utilize classification to detect events. For example, Alsaedi, Burnap, and Rana (2017) use a combination of classification, online clustering and summarization to detect disruptive events based on a Twitter data stream. First of all, event-related tweets are distinguished from non event-related tweets by classifying them according to two labels: [Event] and [Non-event]. Secondly, clustering is performed to identify the topic of an event based on temporal, spatial and textual features. Lastly, the tweets within the clusters are automatically summarized, so that the output can be interpreted by policy and decision-makers.

2.4.3 Classification of incidents

Once an event is detected, it can be determined whether the event is an incident or not. Whether an event can be classified as an incident, depends on the content of tweets and whether these tweets are related to an incident. In order to automatically distinguish incident-related tweets from non incident-related, supervised and unsupervised classification techniques can be utilized (Imran et al., 2015). Supervised classification "learns" a machine learning model from features of labeled cases in order to label new, unseen data items. On the other hand, unsupervised classification refers to a family of methods that seek to identify and explain important hidden patterns in unlabeled data. This research project is limited to supervised classification.

Most studies focused on supervised incident classification of Twitter data do not follow one uni-formal process, but it seems they follow common steps. A supervised classification approach of Twitter data typically involves the labeling of tweets according to their relevance to the incident, preprocessing of the data, selecting which features must be taken into account and using an algorithm to create a machine learning model (Elsafoury, 2020; Qian et al., 2016; Salas, Georgakis, & Petalas, 2017;

Nguyen, Liu, Rivera, & Chen, 2016). The first three steps are often performed in arbitrary order. Following, we will first discuss the act of data labeling and accompanying problems. Additionally, we will cover pre-processing of the data. Moreover, we will shortly discuss selecting features of tweets. Lastly, we will examine the different algorithms that are most commonly used to train a machine learning model on the data.

Tweet labeling

Before a supervised classification algorithm can "learn" a machine learning model of labeled cases, tweets should be labeled according to their relevance to an incident. In current research, there is no strict approach regarding the labeling of incidents. However, it seems that there is a trend regarding two approaches.

First of all, some studies use a binary approach, using labels for tweets that are related or not related to a specific event. For example, Salas et al. (2017) studied the use of Twitter for supporting real-time incident detection in the United Kingdom. For this study, they collected 3,956,871 that were labeled into two classes: Traffic (traffic-related) or Non-traffic (not related to traffic). In addition to this first approach, some studies use multiple labels that describe specific events. In research by Nguyen et al. (2016), focused on detecting traffic incidents using Twitter, they used over ten labels to describe different types of incident-related tweets. A possible drawback of this approach is that the prediction accuracy of under-representative labels decreases.

However, data labeling is often a time-consuming and costly task to do (Dabiri & Heaslip, 2019). Therefore, current literature proposes several solutions to solve this problem. One of these solutions is crowdsourcing. Crowdsourcing is used in many studies and is proven to be an easy, cheap and fast way of labeling data (Snow, O'connor, Jurafsky, & Ng, 2008). With crowdsourcing, the data labeling task can be divided over workers that label tweets individually, sometimes in return for money (Estellés-Arolas & González-Ladrón-De-Guevara, 2012). For example, Elsafoury (2020) used crowdsourcing to label 6693 tweets related to the Gezi Park protest in 2013. Every worker needed to answer two questions for every tweet: (1) Is the tweet related to the Gezi Park protest in 2013? and (2) Does this tweet report/discuss a violent incident? For each question, two options were given: "Yes" and "No". Each tweet was labeled by three workers and only the tweets that received the same label from all three workers were considered in the final dataset.

Another solution to the data labeling problem is clustering. Clustering is an unsupervised learning technique that divides data into groups of similar objects (Aggarwal, 2018). The goal of clustering is to increase the dissimilarity between groups and the similarity within groups. Thus, instead of labeling the complete dataset, one could label only part of the dataset and use a clustering technique to determine the labels of the unlabeled cases.

Furthermore, we will cover a relatively new and hybrid solution to the data labeling problem, called active learning, which combines manual and automatic labeling (Imran et al., 2015). With active learning, the algorithm is allowed to choose the data from which it learns from (Settles, 2009). Typically, the task starts with a small number of labeled cases and a large number of unlabeled cases. Based on the small number of labeled cases, the algorithm determines which cases the human annotator must label next. In Figure 2.6, a typical active learning procedure is demonstrated. The idea behind active learning is that the combination of manual and automatic labeling leads to better results with less training, and it is proven that it actually can work.

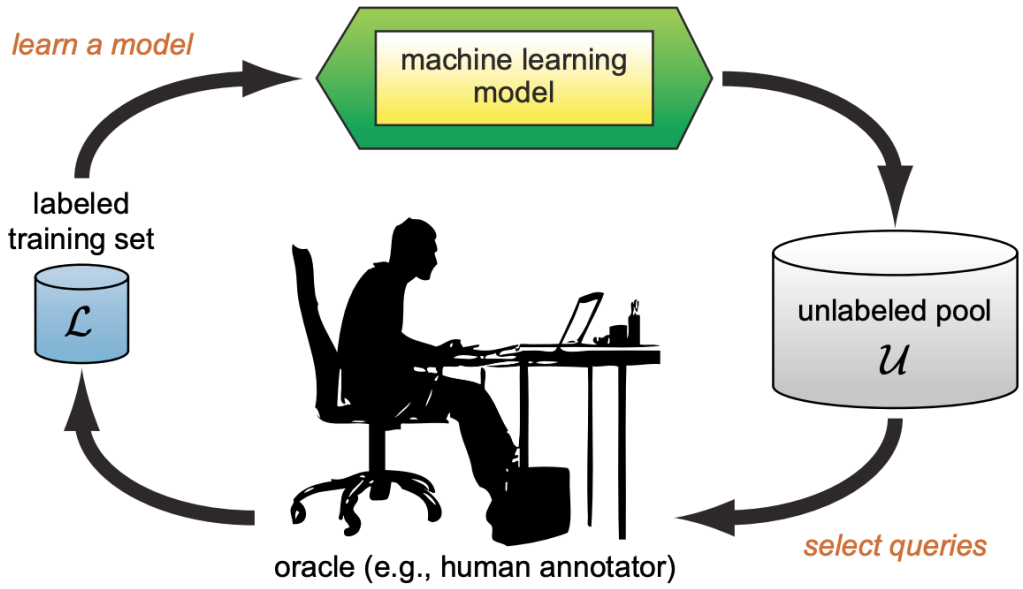

FIGURE 2.6: Active learning cycle, described in Settles (2009)

Data pre-processing

Before the data can be analyzed by a classification algorithm, the data should be pre-processed. While data mining tools are designed to handle structured data (Vijayarani, Ilamathi, & Nithya, 2015), the text of tweets is generally considered as unstructured data. To impose structure on text, several operations can be performed. First of all, all words can be transformed to lowercase characters. Since uppercase and lowercase forms of words are assumed to have no difference, all uppercase words are converted to their lower case forms (Uysal & Gunal, 2014). Additionally, each tweet can be tokenized. This means that the text in the tweet is broken into words or other meaningful elements called *tokens* (Kannan & Gurusamy, 2014). Consequently, all the stop words in a tweet can be removed. Stop words are common words that occur very frequently, such as 'and', 'are' and 'this'. Stop words account for 20-30% of all word counts and are not useful for classification. Thus, to improve the efficiency and effectiveness of classification, stop words should be removed. Furthermore, words in the dataset can be minimized to their stem (called *stemming*). This is the process of conflating the variant forms of a word into a common representation, the stem. For example, the words "presentation", "presented", "presenting" could all be reduced to the stem "present". Moreover, it is worth noting that there is no unique combination of preprocessing tasks that provides successful classification results for every domain and language studied (HaCohen-Kerner, Miller, & Yigal, 2020). Therefore, it is important to test out several combinations of preprocessing tasks.

In addition to the general operations for text mining, additional operations specifically for tweets can be identified. These operations include removing user mentions, HTTP addresses, hashtags, digits and words that are less than two characters long (Elsafoury, 2020).

Feature selection

Once structure on the data is imposed, it must be determined which features are selected to train the machine learning model. Because of the richness of Twitter data, it provides a great opportunity to use a large variety of features (Saeed et al., 2019). Saeed et al. (2019) identified several categories of features regarding Twitter analysis.

First of all, keyword-based features can be used, referring to the occurrence of words in tweets. A standard approach to keyword-based features is the bag-of-words model. Using the bag of words model, each tweet (also referred to as a document) is represented by a bag of words, which is the set of words it contains along with a count of how often it appears (Witten, 2004). Based on how many times specific words appear in a tweet, the algorithm classifies a certain category. By default, the bag-of-words model measures the frequency of unigrams (one word), but this could also be extended to bigrams (two-word occurrences) or trigrams (three-word occurrences), or a combination of those. For example, in research by Salas et al. (2017), they tried to classify incident-related tweets using various n-gram features. Results showed that unigrams provided the best result in terms of accuracy (90,71%) followed by a combination of unigrams and bigrams (89,7%) and a combination of unigrams, bigrams and trigrams (88,35%). A drawback of only relying on term frequency is that highly frequent words can dominate the classification. Therefore, an alternative measure to term frequency was introduced, called Term Frequency/Inverse Document Frequency (TF-IDF). TF-IDF does not only take the word frequency into account, but also normalizes each word by Inverse Document Frequency, thereby reducing the weight of terms that occur more frequently in the collection (Aggarwal & Zhai, 2012). However, both of these approaches only take the frequency of words into account, but not the word's meaning or the order of words. Moreover, as the number of words in a tweet corpus is very large and only a small subset of words is used in each tweet, the resulting matrix suffers from sparsity and curse of dimensionality (Dabiri & Heaslip, 2019). To overcome these problems, word embeddings can be utilized. Word embeddings map words to vectors of numbers (also called word vectors) in such a way that words with similar meaning tend to be closer to each other in vector space (Dabiri & Heaslip, 2019). Important examples of word embeddings tools are Word2vec (created by Google) and FastText (an extension of Word2vec).

Secondly, Saeed et al. (2019) identified Twitter-based features, such as the use of hashtags, time-stamps and number of retweets. For example, hashtags are considered explicit content descriptors and frequently appear in event contents.

Additionally, location-based features can be utilized, referring to the geotagging of tweets (Saeed et al., 2019). If a tweet is geotagged, the geographic information (coordinates) of the user at the moment of tweeting are shared with Twitter. Geotagging is one of the important features which is widely used by research studies and plays an important role in spatial event detection. A problem related to geotagging is that approximately only 2% of the total tweets are geotagged (T. Hua, Chen, Zhao, Lu, & Ramakrishnan, 2013). Proposed solutions for this problem are inferring the location of tweets based on their locality or by mentioned locations in tweets (Saeed et al., 2019).

Lastly, language-based features, such as nouns, verbs and part-of-speech (POS) tags can be important (Saeed et al., 2019). These features are most authoritative in terms of describing and expressing event-related information. Part-of-speech (POS)

tagging assigns tags to each of the words, such as nouns, verbs, adjectives and determiners (Yang, Tan, Selvaretnam, Howg, & Kar, 2019). These tags are then used to capture desired entities, such as names and locations, and are often used in Named Entity Recognition (described in Section 2.4.4).

Classification model using algorithms

After the data is labeled, pre-processed and the features have been selected, a machine learning algorithm can be applied. This step aims to distinguish real-time incidents from irrelevant events and is input for the following step (information extraction).

Event detection techniques that focus on detecting a specific type of event based on tweets, such as an incident during protest demonstrations, mainly rely on supervised learning approaches (Atefeh & Khreich, 2015). Several supervised classification algorithms have been proposed for the early detection of specific events, including Naive Bayes, Support Vector Machines (SVM), logistic regression (Xu et al., 2019) and gradient boosted decision trees, which are often used in combination with frequencies of words. In addition, several classifiers can also be combined into one classifier, which is called an *ensemble* (Imran et al., 2015). Currently, there is no consensus on which is the best performing algorithm and the choice for the algorithm is largely dependent on the specific problem setting (Imran et al., 2015).

But recently, in 2019, Dabiri and Heaslip performed a study on event detection of traffic incidents using supervised deep-learning algorithms. In this study, recurrent neural networks (RNN) and convolutional neural networks (CNN) were used to learn long-term dependencies between tweet words and to capture local correlations between consecutive words. This deep-learning approach improved over state-of-the-art methods, such as SVM and Naive Bayes.

2.4.4 Information Extraction

As soon as an incident is detected, useful information in tweets can be extracted. Analyzing these tweets manually could lead to several problems. First of all, it takes time to analyze these tweets manually, which could result in large delays between the detection of an incident and the information extracted about the incident. Secondly, there could be inconsistency between participants in analyzing the tweets, i.e. person A perceives tweets differently than person B. Therefore, we can rely on automatic information extraction (IE) from tweets.

Information extraction is the task of automatically extracting structured information from unstructured (e.g. plain text documents) or semi-structured data (e.g. web pages) (Imran et al., 2015). The result of the information extraction task is to transform the unstructured data into a machine-readable format, so that the data can be processed, filtered, sorted and aggregated by machines (W. Hua, Huynh, Hosseini, Lu, & Zhou, 2012). In the context of incidents during protest demonstrations, IE can be used to extract incident information from tweets. For example, the sentence "5 injured and 10 dead in Antofagasta" can be transformed to

<people-affected=5, reporttype=injury, location=Antofagasta, Chile>,
<people-affected=10, report-type= fatalcasualty, location=Antofagasta, Chile>

First, we will cover the challenges of Twitter regarding Information Extraction. Following, entity extraction (a common task in IE) will be discussed.

Challenges of Twitter regarding Information Extraction

Although Twitter provides opportunities for information extraction, it also has its challenges. First of all, tweets on Twitter are length-limited, meaning that users can post a maximum of 280-characters per tweet. This results in ungrammatical tweets, which makes traditional NLP tools inappropriate to use (W. Hua et al., 2012). Additionally, this limitation of characters can result in limited context information, possibly making it hard to determine an entity's type (X. Liu, Zhang, Wei, & Zhou, 2011). Secondly, tweets are often written in an informal manner, containing noisy texts, such as abbreviations, symbols and misspellings, bringing great difficulties to analyzing the content and meaning of tweets.

Named Entity Recognition (NER)

Named entity recognition (NER) is the task of identifying mentions of named-entity types such as persons, organizations and locations from text (Nadeau & Sekine, 2007). In general, there are two main approaches to NER called rule-based approaches and statistical approaches (W. Hua et al., 2012).

Rule-based approaches define heuristics to identify named entities within documents in a particular domain, which are determined by experts in the domain (W. Hua et al., 2012). One of the advantages of this approach is that the execution time of rule-based systems is shorter than other methods. Moreover, developers can easily control the rules to obtain optimization for specific domains. But this approach also has some limitations. First of all, it is required that experts define the rules for extraction, which can be rigid and not general enough. Secondly, rule-based approaches are often less effective than statistical approaches (Imran et al., 2015).

Statistical approaches solve the problem of entity recognition in two phases. First, they decompose unstructured texts (into tokens or word chunks) and second, they label the parts of the decomposition (W. Hua et al., 2012). The main statistical approaches use hidden Markov models, conditional Markov models, maximum-entropy Markov models or conditional random fields (Imran et al., 2015; W. Hua et al., 2012). Currently, conditional random fields (CRF) based methods are state-of-the-art and outperform all previous machine learning-based methods (W. Hua et al., 2012). For example, Imran, Elbassuoni, Castillo, Diaz, and Meier (2013) applied a CRF based method to the extraction of information from tweets. Their approach was subdivided into two steps. First, tweets were classified according to categories, such as "Caution and advice" and "Casualties and Damage". Afterward, specific entities were extracted based on the category. For example, for "Infrastructure damage" tweets the reported damage is extracted, while for "donations" tweets, the item being offered in donation is extracted.

2.5 Gaps in current literature

Following this literature review, three gaps in current research are identified. First of all, to the best of our knowledge, no study in current literature detects incidents during protest demonstrations by using Twitter data. Most studies related to incident detection by using Twitter data are focused on traffic incidents and most studies related to event detection with Twitter data are focused on trending topics or (real-time) events in general.

Secondly, most studies related to incident detection with Twitter describe the identified incidents, but do not provide a clear-cut point on when the system decides an event is classified as an incident. Some studies define a point on when the system decides a Twitter event is classified as an incident, but these points are chosen arbitrarily and no justification is provided for this point, such as in Dittrich and Lucas (2014).

Thirdly, scholars have proposed several models to describe the coverage of incidents and events on Twitter (Klein et al., 2012; Hu et al., 2012). However, it remains yet unclear how incidents during protest demonstrations are covered on Twitter over time. Therefore, this study is intended to describe the different phases of Twitter coverage after an incident has occurred during a protest demonstration.

Chapter 3

Research Questions

Based on the research objective and the gaps in current literature, three research questions are defined.

Scholars have proposed several models to describe the coverage of incidents and events on Twitter (Klein et al., 2012; Hu et al., 2012). However, it remains unclear how incidents during protest demonstrations are covered on Twitter over time. Therefore, the first research question is:

> **[RQ1]**: *"What phases of Twitter coverage after an incident during a protest demonstration can be identified?"*

In order to answer this question, an exploratory data analysis is performed on a dataset related to protest demonstrations in the Netherlands from 31 May to 7 June 2020. Moreover, there is currently no research that detects incidents during protest demonstrations by using Twitter data. Therefore, the research project is extended by the development of multiple machine learning models using standard supervised classification algorithms aimed at distinguishing incident-related tweets from non incident-related tweets. The goal is to identify the differences in performance between several standard supervised classification algorithms in the context of incident-related tweet prediction during protest demonstrations, and will provide an answer to the second research question:

> **[RQ2]**: *"What are the differences in performance between several standard supervised classification algorithms in the context of incident-related tweet prediction during protest demonstrations?"*

Performance in the context of incident-related tweet prediction is aimed at distinguishing incident-related tweets from non incident-related tweets. To answer this question, the developed machine learning models will be compared according to several performance metrics.

Furthermore, current literature related to incident detection based on tweets describes the identified incidents, but does not provide a justified clear-cut point on when a system decides an event is classified as an incident. To incorporate this gap and to obtain a broader perspective of the information need of Open-Source Intelligence Team (OSINT) analysts of the Dutch police, the third research question is defined as:

> **[RQ3]**: *"What is the information need of OSINT analysts at the Dutch national police force when automatically detecting incidents using Twitter data?"*

In order to answer this question, interviews with analysts of the Open-Source Intelligence Team (OSINT) at the Dutch national police force will be conducted.

Chapter 4

Approach

4.1 Design Cycle

The research project is framed around the Design Cycle, a research method that results in a validated treatment (Wieringa, 2014). The Design Cycle is part of the engineering cycle, which consists of the following tasks:

1. *Problem investigation*: the design of a treatment is prepared by learning more about the problem to be treated. The goal of this stage is to determine what phenomena must be improved and to identify stakeholders, goals, problems, effects and contributions to goals.

2. *Treatment design*: Requirements are specified, available treatments are identified and one or multiple artefacts are designed for the treatment.

3. *Treatment validation*: Effects, trade-offers and requirements satisfied by the artefact are determined. The goal of this stage is to determine whether one of the designed artefacts would treat the problem.

4. *Treatment implementation*: The designed artefacts are implemented in a real-life situation.

5. *Implementation evaluation*. The goal of this stage is to evaluate a treatment after it has been applied in the original problem context. The same questions are asked as in the problem investigation stage, but with a different goal.

The engineering cycle is often utilized in long-term research projects, where the designed artefact can be implemented in a real-life situation. However, with design science projects, transferring new technology to a real-life situation is not part of the research project (Wieringa, 2014). Therefore, the Design Cycle only carries out the first three tasks of the engineering cycle: Problem investigation, treatment design and treatment validation (as presented in Figure 4.1). In the following sections, the three individual tasks with their according research approaches will be described in more detail.

4.1.1 Problem investigation

The first task is to get an understanding of the problem to be treated. In order to get this understanding, current literature is examined by conducting a literature review. This activity is performed to get a better understanding of social movements, social media and protest demonstrations, and incident detection using Twitter data.

In addition to the literature review, exploratory data analysis will be performed to better understand the phases of Twitter coverage after an incident during a protest

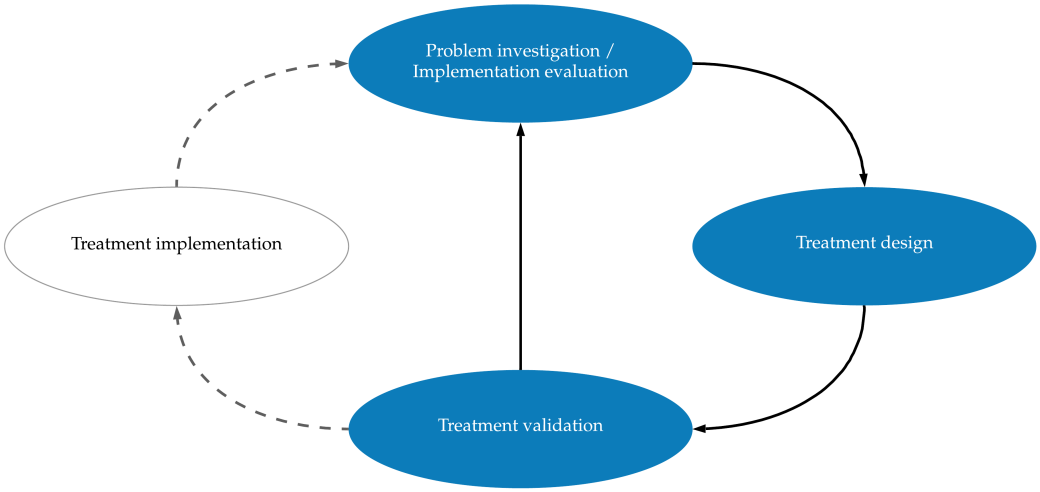

FIGURE 4.1: Design cycle as part of the engineering cycle (Wieringa, 2014)

demonstration [RQ1]. Therefore, a dataset consisting of Dutch tweets related to protest demonstrations from 31 May to 7 June 2020 was collected and will be analyzed.

4.1.2 Treatment design

Based on the results of the previous task, two types of treatments are identified and designed for the treatment.

Model treatment

The goal of this research project is to automatically detect incidents during protest demonstrations by using Twitter data. Therefore, several machine learning models will be developed that can distinguish incident-related tweets from non incident-related tweets. In order to develop these models, tweets related to a Dutch protest demonstration were collected and multiple machine learning models will be trained on the tweets using several standard supervised classification algorithms.

System treatment

Secondly, a system will be designed and developed that can detect events on Twitter, determine whether these events are related to incidents and can provide a warning if an incident is detected. To determine when this warning of detected incidents during protest demonstrations will be provided, semi-structured interviews with OSINT analysts of the Dutch national police force will be conducted [RQ3]. Moreover, the system will utilize one of the developed machine learning models, as described in Section 4.1.2.

4.1.3 Treatment validation

In order to validate the machine learning models and system described in Section 4.1.2, we will validate these treatments separately.

Model validation

To validate whether the machine learning models can distinguish incident-related tweets from non incident-related tweets, the models will be evaluated using several accuracy metrics, such as F-Measure, Precision, Recall, Area under the ROC Curve (AUC) and accuracy, as described in Section 5.6.3. Moreover, the models will be compared to determine the differences between the standard supervised classification algorithms in the context of incident-related tweet prediction during protest demonstrations [RQ2].

System validation

Additionally, it will be validated whether the designed system can detect events on Twitter, determine whether these events are related to incidents and can provide a warning if an incident is detected. In order to do this, the system will be evaluated using a similar dataset as the training dataset, containing tweets from a protest demonstration not previously used in the training of the machine learning models. Moreover, the main validation method that will be utilized is when the system detects the incident compared to when the incident occurred in real life.

4.2 CRISP-DM

In order to provide answers to the first and second research questions, data analysis on Twitter data is performed. To structure this data analysis, CRISP-DM will be utilized, which is the most used method for data mining projects (Piatetsky, 2014). CRISP-DM is a hierarchical process model that provides an overview of the life cycle of a data mining project and divides the process into six main phases (Chapman et al., 2000). This research project uses an adaptation of the CRISP-DM method because it does not share the business narrative of CRISP-DM. Therefore, the first phase is changed from Business Understanding to Domain Understanding.

Following, we will shortly reflect on each of the phases and describe the activities:

1. *Domain understanding*
   This phase consists of a literature review of related work in the context of social movements, protest demonstrations, Twitter and incident detection using Twitter.

2. *Data understanding*
   In this phase, the initial Twitter datasets related to a protest demonstration will be collected. The data will be analyzed, data quality problems will be identified and the data will be explored.

3. *Data preparation*
   During data preparation, the Twitter datasets will be preprocessed so that they can be used for the next phase. Tasks include selecting the data and cleaning the data (with a specific focus on Natural Language Processing of the tweet text).

4. *Modeling*
   In this phase, various machine learning algorithms will be selected and utilized to create machine learning models. Some techniques have specific requirements on the data, so it could occur that going back to the previous phase is required.

5. *Evaluation*
   In the Evaluation phase, the created machine learning models will be compared based on a set of performance metrics.

6. *Deployment*
   In this phase, the designed system (as described in Section 4.1.2) will be developed and tested on a Twitter dataset related to a similar protest demonstration, but not used during training of the machine learning models.

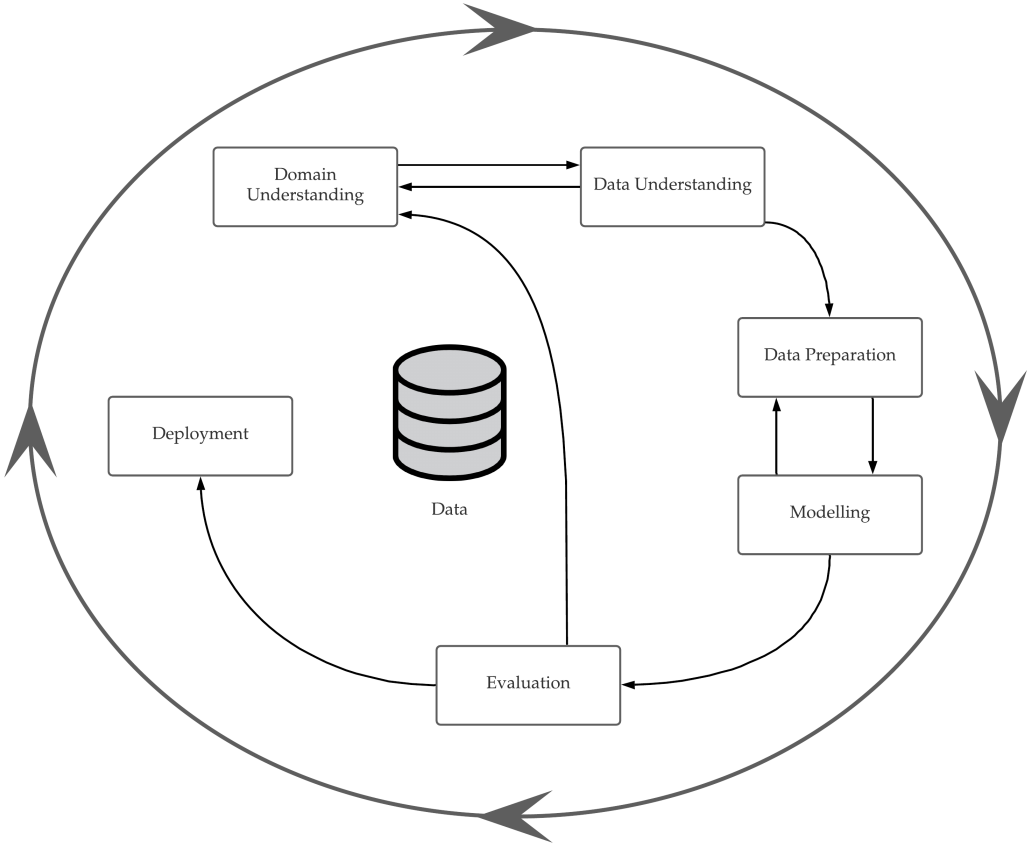

FIGURE 4.2: Adaptation of CRISP-DM used in this research project

4.3 Reliability and validation

In this section, the reliability and validation of this study are discussed.

4.3.1 Reliability

The reliability of a study refers to the extent to which results are consistent over time, an accurate representation of the total population under study and if the results can be reproduced under a similar methodology (Golafshani, 2003). To ensure the reliability of this research project, all the used methods are explained in Section 5. Moreover, both quantitative methods (data analysis) and qualitative methods (interviews) are used to test the research questions. To ensure the reproducibility of this research project, all tools, scripts and techniques are described in this document. Moreover, the code used in this project is made publicly available on a Github repository (Wouters, 2021).

4.3.2 Validity

We distinguish three types of validity: construct validity, internal validity and external validity.

Construct validity refers to whether the research truly measures the concept it was intended to measure. The main concept of this study is the detection of incidents during protest demonstrations which is measured by using Twitter data. To ensure the used methods are valid, we have based the used methods on existing literature.

Internal validity refers to the degree to which the treatment causes the outcome and confounding is avoided. This study is aimed at finding relationships between Twitter data and incidents during protest demonstrations. By including a large number of variables during the data analysis, we aim to reduce the number of confounding variables that influence this relationship.

External validity refers to the generalizability of the research findings. The findings of this research project are limited to the defined scope, which is the detection of incidents during protest demonstrations by using Twitter data.

4.4 Relevance

In this section, the relevance of this research project is discussed from an academic and practical perspective.

4.4.1 Academic relevance

In Section 2.5 the gaps in current literature are described. This study intends to contribute to current literature by:

- Identify phases of Twitter coverage after an incident during a protest demonstration
- Develop machine learning models in the context of distinguishing incident-related tweets from non incident-related tweets during protest demonstrations
- Provide insights in the information need of OSINT analysts at the Dutch national police force when automatically detecting incidents using Twitter data

Additionally, this study contributes to the limited literature focused on analyzing tweets in the Dutch language.

4.4.2 Practical relevance

With this study, we intend to develop a system that can detect incidents during protest demonstrations, by using Twitter data. Moreover, with this system, a clear-cut point is provided on when the system decides an event is classified as an incident. By providing this clear-cut point, law enforcement could be timely warned when an incident during a protest demonstration is detected. Currently, government services are not utilizing Twitter data to the full extent (see Section 2.1.3). Therefore, the system created in this study could be used in existing government systems, such as systems of the police.

Chapter 5

Methods

5.1 Background information

At 17:00, 1st June 2020, a protest demonstration in Amsterdam (the Netherlands) was organized to protest against police violence as a response to the death of George Floyd. It was estimated by the Dutch police that around 250 people would attend the protest demonstration (Rijksoverheid, 2020). But surprisingly, more than 10,000 protesters attended the protest demonstration.

At the time of the protest demonstration, rules were imposed by the Dutch government to control the spread of COVID-19. These rules included that people needed to have a distance of 1.5-meters and it was not allowed to get together in groups larger than 4 people.

However, because of the large number of participants of the protest demonstration, the protesters were not able to keep a distance and the group of approximately 10,000-14,000 people (at its peak) was larger than the allowed number of 4 people. A detailed timeline of all the events related to the protest demonstration on 1st June 2020 in Amsterdam is presented in Appendix A.

5.2 Data collection

Tweets were extracted in the week of 31 May to 7 June 2020 using the *tweepy* Python package that provides access to the Twitter Search API (Standard level API access). Tweets containing the word "demonstratie" (Dutch for "demonstration) and written in the Dutch language (according to the Twitter API) were collected. From this dataset, only the tweets published on 1st June 2020 were selected and used for this research project.

For each tweet, the following variables were extracted:

1. *created_at*, the datetime at which the tweet was created in UTC timezone.
2. *id*, a unique tweet ID, as determined by Twitter.
3. *text*, the text of the tweet with a maximum of 140 characters. If the tweet exceeded 140 characters, the text is truncated.
4. *coordinates*, the GPS coordinates of the user when the tweet was created (if the user has not disabled this feature).
5. *hashtags*, the hashtags mentioned in the tweet.
6. *place*, a geographical location of the tweet as determined by the user. The user can set this geographical location manually per tweet, which implicates that

the place does not necessarily represent the actual geographical location the tweet was sent from.

7. *lang*, the language of the text in the tweet.
8. *retweet_count*, number of times a tweet has been retweeted at the moment of API extraction.
9. *favourite_count*, number of times a tweet has been favorited at the moment of API extraction.
10. *user_id*, an unique user ID, as determined by Twitter.
11. *user_location*, the user-defined location for its account's profile.
12. *user_screen_name*, the screen name that a user identifies themselves with on Twitter.
13. *org_tweet_created_at*, if a tweet is a retweet, it contains the datetime of the original tweet in UTC timezone.
14. *org_tweet_user*, if a tweet is a retweet, it contains the user ID associated with the original tweet.

Because the initial dataset (as described above) does not contain the full text of a tweet (if the tweet was truncated) and does not have all relevant variables of a tweet (e.g. images, media and URLs are missing), it might negatively affect the performance of a machine learning model. Therefore, it was necessary to create an additional dataset containing the full text and more relevant variables of a tweet. In order to create this, the tweet ids of all unique tweets in the initial dataset were selected and used to get the full text and more variables of the tweets. Due to this research project starting approximately four months after the creation of the initial dataset, this resulted in a decrease of unique tweets. This is a direct result of tweets that were deleted by users, users that were suspended by Twitter or users that deleted their Twitter account. Just as with the initial dataset, tweets were extracted using the Twitter Search API. The most important difference with the initial dataset in terms of variables is the introduction of:

1. *full_text*, the full text of a tweet.
2. *entities*, which contains the hashtags, user mentions, URLs and media shared in a tweet.

This dataset is from now on referred to as the unique tweets dataset. The initial dataset is referred to as the complete tweets dataset.

5.3 Data preparation

Subsequently, the datasets were prepared so that they could be used for exploratory data analysis and the development of machine learning models. This process was divided into five sub-steps: variable preprocessing, derive variables from existing variables, text preprocessing, the creation of additional datasets and imbalanced data set handling. For each of the steps, Python was used as a programming language in combination with Jupyter Notebooks where the data was processed using the *pandas* package. Furthermore, all the code used in this project is publicly available on a Github repository (Wouters, 2021).

5.3.1 Variable preprocessing

In this section, we will describe the preprocessing of variables on each of the datasets. For all variables holds that they were converted to the right data types.

Complete tweets dataset

Some features of the complete tweets dataset required additional preprocessing due to a multitude of reasons. First of all, the *created_at* and *org_tweet_created_at* variables were in UTC timezone and were converted to UTC+2 timezone (local time in Amsterdam during the protest demonstration) using the *pandas* package. Secondly, the *coordinates*, *hashtags* and *place* variables were formatted in JSON and were converted to lists.

Unique tweets dataset

Just as with the complete tweets dataset, some variables in the unique tweets dataset also required some additional preprocessing. First of all, the *created_at* and *org_tweet_created_at* variables were in UTC timezone and were converted to UTC+2 timezone (local time in Amsterdam during the protest demonstration). Secondly, the *coordinates* variable was formatted in JSON and was converted to list format. Thirdly, the variable *entities* (which contains the hashtags, user mentions, URLs and shared media of a tweet in JSON format) was divided into four seperate variables: *hashtags*, *user_mentions*, *urls* and *media* (all in list format).

Moreover, if a tweet shared a website page, the thumbnail URL of the website page was scraped off the website and added to the *media* variable, by using the *requests* and *bs4* packages.

5.3.2 Derive variables from existing variables

Complete tweets dataset

Because the complete tweets dataset lacked some variables, it was necessary to derive variables from the tweet text. Therefore, several variables were created:

- *retweeted*, a boolean value that describes whether a tweet is a retweet or not. This was extracted by performing a Regex function.
- *user_mentions*, list with mentioned users in a tweet. This was extracted by performing a Regex function.
- *user_mentions_types*, list with the types of the mentioned users in a tweet.
- *type_user*, type of the user that sent the tweet.

Regex functions were performed on the text of the tweet by using the *re* Python package. The *user_mentions_types* and *type_user* were labeled according to the process described in Section 5.4.1.

Unique tweets dataset

For the unique tweets dataset, the *user_mentions* variable was already available, which did not require any additional derivation from the text. Moreover, for each user that tweeted the user type was represented in the *type_user* variable.

Additionally, the *has_media* variable was introduced, which is a boolean variable that describes if a tweet has media (contains images or video). Also, if a tweet shares a website URL with a thumbnail, the *has_media* will be set to True.

5.3.3 Text preprocessing

In this section, the preprocessing of text in the unique tweets dataset is described. Because the complete tweets dataset is not used for modeling, we will not cover that in this section.

In order to mine text using a machine learning algorithm, structure must be imposed on the data. Therefore, the following tasks were performed on the text of a tweet:

- All text was converted to lowercase.
- Words with less than two characters were removed.
- URLs were removed by performing a Regex function.
- Punctuation was removed unless it was part of a digit (such as 5.000). This exception was determined because it was expected that the occurrence of "1,5" would appear often in the dataset. By removing punctuation for such terms, it could possibly lose its meaning.
- Emojis were removed, by using the *demoji* package.
- User mentions were removed by performing a Regex function. A user mention is when a tweet contains "@screen_name". Both the "@" symbol and the screen name were removed from the text.
- Non-alphanumeric characters, line-breaks and tabs were removed by performing a Regex function.
- Dutch stop words were removed by using the *nltk* package. One stop word was added to the Dutch stopwords: "RT" (which indicates that a tweet is a retweet).

All Regex functions were performed by using the *re* package in Python.

Moreover, for the preprocessed text in a tweet (text after performing the steps above), it was considered to create two text variables: one consisting of the preprocessed tweet text containing the hashtag contents of a tweet and one only consisting of the preprocessed tweet text. Also, some algorithms require the text data in a list and others in a string. To combine these two challenges, four variables were added to the dataset:

1. *preprocessed_text*, refers to the preprocessed text with hashtag contents in string format.
2. *preprocessed_text_no_hashtag*, refers to the preprocessed text without hashtag contents in string format.
3. *preprocessed_text_tokenized*, refers to the preprocessed text with hashtag contents in tokenized format.
4. *preprocessed_text_no_hashtag_tokenized*, refers to the preprocessed text without hashtag contents in tokenized format.

Because labelers were allowed to take the full context of a tweet into account (as described in Section 5.4.2), it is decided that if a tweet shares a URL, the preprocessed web title of that URL was also added to the four variables described above. The titles of web pages were extracted using the *requests* and *bs4* packages.

5.3.4 Additional datasets

To understand how an incident was covered on Twitter over time [RQ1], it was necessary to label users according to specific user types. Therefore, all users that tweeted and all users that were mentioned at least once in the complete tweets dataset were extracted using the Twitter Search API, which resulted in two additional datasets:

- Users, which contains all users that tweeted in the complete tweet dataset.
- Mentioned users, which contains all users that were mentioned at least once in the complete tweets dataset.

For each user, the following information was extracted:

1. *screen_name*, the screen name that a user identifies themselves with on Twitter.
2. *followers_count*, the number of followers a Twitter user has at the moment of API extraction.
3. *friends_count*, the number of friends a Twitter user has at the moment of API extraction.
4. *listed_count*, the number of public lists that the user is a member of at the moment of API extraction.
5. *created_at*, the datetime of creation of the Twitter user in UTC timezone.
6. *favourites_count*, the number of tweets a user has liked in the account's lifetime up to the moment of API extraction.
7. *verified*, a boolean value that describes whether a Twitter user is verified.
8. *tweet_count*, represents the total number of tweets at the moment of API extraction.
9. *description*, refers to the description of the Twitter profile.

Moreover, these datasets do not include all the (mentioned) users represented in the dataset. Users that deleted their Twitter profiles or were suspended by Twitter are not included.

5.3.5 Imbalanced dataset handling

After labeling the unique tweets according to the labeling process described in Section 5.4.2, the unique tweets dataset appeared to have a class imbalance. To handle this problem, two additional datasets were created to balance the dataset more to the incident-related class. On the first dataset, all incident-related tweets were back-translated using the Google Translation API. This was performed in three steps:

1. The tweet was translated to English and French

2. The English/French tweet was translated to Dutch
3. If the back-translated tweet was the same as the original tweet, the tweet was discarded. Otherwise, the back-translated tweet was added to the dataset.

English and French were chosen because we were familiar with those languages. Therefore, potential problems during back translation could be more easily identified and resolved. Google Translation was chosen because it was considered a sufficient translator and because we were familiar with using Google Cloud functions.

On the second dataset, in addition to back translation, random word replacement was conducted on each incident-related tweet. This includes that for each incident-related tweet 3 random words were replaced by synonyms. If a random word was chosen which did not have any synonyms, an other random word was chosen. 3 was chosen because of two reasons. First of all, the tweet with the least number of words contained 4 words. Secondly, this would account for replacing 10% of the contents of each tweet (on average), because the average number of words in a tweet was 34 words (as presented in Section 6.1.2). 10% was considered as a reasonable amount to replace while keeping the most information of the original tweet.

Other activities were also considered for handling the imbalanced dataset, such as oversampling, undersampling, SMOTE and other data augmentation techniques (e.g. random swapping of words and random deletion of words). Oversampling the minority class was not utilized because it could lead to more overfitting of the machine learning models on the training data set. Undersampling the majority class was not conducted, because it could potentially lead to an information loss because of the lost cases of the majority class. Synthetic Minority Over-sampling Technique (SMOTE) (Chawla, Bowyer, Hall, & Kegelmeyer, 2002) was not performed, because it would involve creating two separate balanced datasets for TF-IDF vectorized matrices and for word embeddings, making it hard to compare the performance of the machine learning models of the various algorithms. Moreover, random swapping of words was not utilized, because this would not have any influence on the algorithms that use TF-IDF vectorized matrices. Lastly, random deletion of words was considered, but this could potentially lead to an information loss.

5.4 Data labeling

5.4.1 User labeling

To determine what type of users were sending tweets and what type of users were mentioned in tweets, Twitter users needed to be labeled according to user types. To label users to their user type, four groups of interest were determined:

1. Users that were mentioned at least 5 times
2. Users that tweeted at least 5 times
3. Verified users that were mentioned at least once
4. Verified users that tweeted at least once

Verified users were considered relevant because of their verified status on Twitter. Furthermore, with the other two groups, a trade-off was made between the number of users that needed to be labeled and the amount of information these labeled users would provide. Thereby, 5 seemed to be the 'sweet spot' for both mentioned

and tweeted users. When lowering the threshold to 2, the number of tweeted users quadrupled and the number of mentioned users increased by 256%. When setting the threshold to 10, only 30% of the number of tweeted users and 54% of mentioned users remained (compared to the threshold of 5).

After extracting the users using the Twitter Search API (as described in Section 5.3.4), the users were partly automatically labeled (based on keywords in their Twitter profile description or screen name) and partly internally labeled by the author of this research project, as described in Appendix B. The users were labeled according to their profile picture, screen name and description of their Twitter profile. If these elements did not provide sufficient information, a Google search was conducted with the name of the Twitter profile. Based on the search results on Google (and websites pages in these search results), the label of the user was decided upon. For approximately 80% of all labeled users this additional step was performed.

5.4.2 Tweet labeling

Following, the tweets present in the unique tweets dataset were labeled into two classes: [Incident-Related] and [Not Incident-Related]. Incident-Related refers to expressions of law-violating behavior, such as observations of violence and riots. Furthermore, non-compliance with the COVID-19 rules (not keeping a distance of 1.5-meters and people larger in groups than four) also suffices for the label Incident-Related. Additionally, if a tweet is labeled as Incident-Related, a second label needs to be fulfilled that describes the type of incident, according to 3 categories: [COVID-19] (describing non-compliance with the COVID-19 rules), [Violence] and [Riots].

When labeling the tweets according to the aforementioned labels, labelers were allowed to take the full context of a tweet into account. This includes the text of a tweet, the media of a tweet (if the tweet shared media, such as video and images) and the title and thumbnail of a webpage (if the tweet shared a URL).

The tweets were internally labeled by six labelers, consisting of the author and daily supervisor of this research project, two employees at the Dutch national police force and two students directly related to this research project. This group of labelers conducted weekly meetings to discuss questionable tweets or problems during the labeling process. Moreover, an in-house label tool (called Tweeti) was utilized to label the tweets, which kept track of the labeling process and how many times each tweet was labeled. Each labeled tweet was labeled by multiple labelers. If a tweet was updated at least ten times (meaning that two or more labelers did not agree with each other's labels), the tweet was internally discussed at the end of the labeling process to assign the right class. In Appendix C, descriptive statistics of the labeling process and individual labelers are presented.

5.5 Exploratory data analysis

The goal of the exploratory data analysis is to get familiar with the data, find correlations in the data and identify phases of Twitter coverage after an incident during a protest demonstration [RQ1]. For the exploratory data analysis, both the complete tweets dataset and the labeled unique tweets dataset will be utilized (as described in Section 5.2).

Because the complete tweets dataset did not contain the full text of a tweet, the emphasis was less on the contents of the tweet, but more on the user types of users that tweeted, used hashtags and the mentioned users in tweets. Moreover, although

the full text of a tweet was not extracted, we argue that the intent of a tweet can be captured using the truncated version of the tweet text. Furthermore, the overall pattern that was observed is that other users were mainly mentioned at the beginning of the tweet, thereby the complete tweets dataset can provide good insights in which users are mentioned over time. Following the approach of Hu et al. (2012), the five most mentioned user types were charted over time with their respective number of mentions.

For the unique tweets dataset, the emphasis is on incident-related tweets. Thereby, the focus is on the contents of the tweet, the shared media of tweets (URLs, images and videos) and what user type sent the tweet.

5.6 Modeling

To identify the differences between standard supervised classification algorithms in the context of incident-related tweet prediction during protest demonstrations [RQ2], several machine learning models are trained and tested on the second prepared balanced dataset (as described in Section 5.3.5). The dataset was divided into a training set (80%) and a test set (20%) by performing a randomized stratified split. The training set was used to train the machine learning models and the test set was used to evaluate the machine learning models. The goal of the modeling task is to predict a tweet into two classes: [Incident-Related] and [Not Incident-Related].

5.6.1 Selection of algorithms

From the literature, the most used standard supervised classification algorithms for the early detection of specific events using Twitter include Naive Bayes (NB), Support Vector Machines (SVM), Logistic Regression (LR) and Gradient Boosted Decision Trees (Atefeh & Khreich, 2015; Xu et al., 2019). Furthermore, convolutional neural networks were found to perform well on the topic of traffic incident detection with tweets and improved over state-of-the-art methods, such as SVM and Naive Bayes (Dabiri & Heaslip, 2019). Therefore, five machine learning algorithms are selected for this study: Naive Bayes (NB), Support Vector Machines (SVM), Logistic Regression (LR), Gradient Boosted Decision Trees (GBDT) and Convolutional Neural Networks (CNN).

5.6.2 Feature selection

For each of the machine learning algorithms, two models were created using the following features:

1. Model that trains on the preprocessed text of a tweet (without the hashtag contents).
2. Model that trains on the preprocessed text of a tweet (with the hashtag contents).

Intuitively, the hashtag in a tweet can be regarded as a special type of text. Therefore, a distinction has been made between the full text of a tweet and the full text of a tweet without the hashtag contents.

To train the convolutional neural network on the preprocessed text, the text was converted to a word embedding using the Dutch pre-trained word vector of FastText

(Grave, Bojanowski, Gupta, Joulin, & Mikolov, 2018). For the other algorithms, the text was converted to a Term Frequency / Inverse Document Frequency (TF-IDF) vector. TF-IDF was chosen over term frequencies, because with term frequencies highly frequent words can dominate the classification. Regarding the convolutional neural network, the *keras* package was utilized. Regarding the other algorithms, *scikit-learn* package was used.

Moreover, for the convolutional neural network, the code and setup of Dabiri and Heaslip (2019) was utilized. For each of the models of the other machine learning algorithms, the hyperparameters were optimized by conducting a 3-fold cross-validation grid search, aimed at optimizing the score of the F-measure.

Furthermore, 10-fold cross-validation was performed to detect overfitting of the machine learning models on the training data.

5.6.3 Model evaluation

To evaluate the machine learning models in the context of incident-related tweet prediction during protest demonstrations, the models predicted the labels of the test set and the following metrics were calculated for each model: Accuracy, precision, recall, Area Under the ROC Curve (AUC) and F-Measure. The definitions of these metrics are based on the paper by Hossin and Sulaiman (2015) and presented in Table 5.1.

Performance metric	Definition
Accuracy	Ratio of correct predictions over the total number of instances evaluated
Precision	Used to measure the positive patterns that are correctly predicted from the total predicted patterns in a positive class
Recall	Used to measure the fraction of positive patterns that are correctly classified
Area Under the ROC Curve	Measure that reflects the overall ranking performance of a classifier
F-Measure	Represents the harmonic mean between recall and precision values

TABLE 5.1: Supervised classification performance metrics.

When evaluating the machine learning models, there will be an emphasis on the F-Measure of the Incident-Related class. Accuracy is not a reliable indicator because we have an imbalanced dataset (27/73% ratio). Thereby, blindly predicting the majority class will lead to an accuracy of 0.73 due to the class imbalance. Moreover, the goal of the modeling task is to distinguish incident-related tweets from non incident-related tweets, and therefore, the minority class (Incident-Related) is more important. The F-Measure is a harmonic mean between recall and precision on the minority class.

5.7 Semi-structured interviews with Dutch police

To provide an answer to [RQ3], interviews with three analysts of the Open-Source Intelligence Team (OSINT) at the Dutch national police force are conducted. It was decided to focus on the OSINT team because these analysts work with open-source

data, such as Twitter data. Together, the analysts account for 36 years of experience within the Dutch police working at multiple departments. Moreover, the interviews were organized in a semi-structured approach, so specific topics were covered, but the conversation is free to vary and likely to change between participants (Miles & Gilbert, 2005). A benefit of semi-structured interviews is finding out the Why, in addition to How many or How much.

The interviews are structured into two phases. The goal of the first phase is to get a general understanding of OSINT, how social media analysis is performed and current problems that arise when performing social media analysis. Also, this phase covers what OSINT analysts consider as an incident during a protest demonstration.

The second phase of the interview is focused on detecting incidents by using social media analysis. During this phase, incident-related tweets from the unique tweets datasets are shown to the participants and questions about these tweets are asked, such as *"Do you want to receive a warning after observing this tweet?"* and *"What information do you want of a detected incident?"*. The goal of this phase is to understand the information need of an OSINT analyst after detecting a potential incident by using Twitter data [RQ3]. The questions that were used as a basis for the semi-structured interviews are presented in Appendix D. The consent form of the semi-structured interviews is presented in Appendix E.

5.8 Designed system

The goal of this research project is to automatically detect incidents during protest demonstrations by using Twitter data. Therefore, a system is designed and developed that can automatically acquire tweets from Twitter, preprocess these tweets, detect if there is currently an incident and automatically send a warning. First, the components of the system are described and elaborated on. Next, the evaluation of the system will be discussed.

5.8.1 Components of the system

The components of the system are presented in Figure 5.1 and the pseudo-code for the system is presented in Appendix F. The tweets are processed in a queue-wise manner. This indicates that every time a tweet is acquired, it is added to a queue, while a second processing thread will process the tweet. By separating the acquiring and processing tasks, blocking of the acquisition task by the processing task is prevented when processing large volumes of tweets.

Tweet acquisition and preprocessing

First, tweets with the word "demonstratie" (Dutch for "demonstration) will be automatically acquired using the Twitter Streaming API. The tweets are acquired one by one and will be individually preprocessed, similar to the preprocessing steps used in Section 5.3. After a tweet has been preprocessed, it will be forwarded to the event detection module.

Event detection

The event detection module is based on the work of Marcus et al. (2011) and consists of three parts: binning a tweet to the current time window, detecting whether there

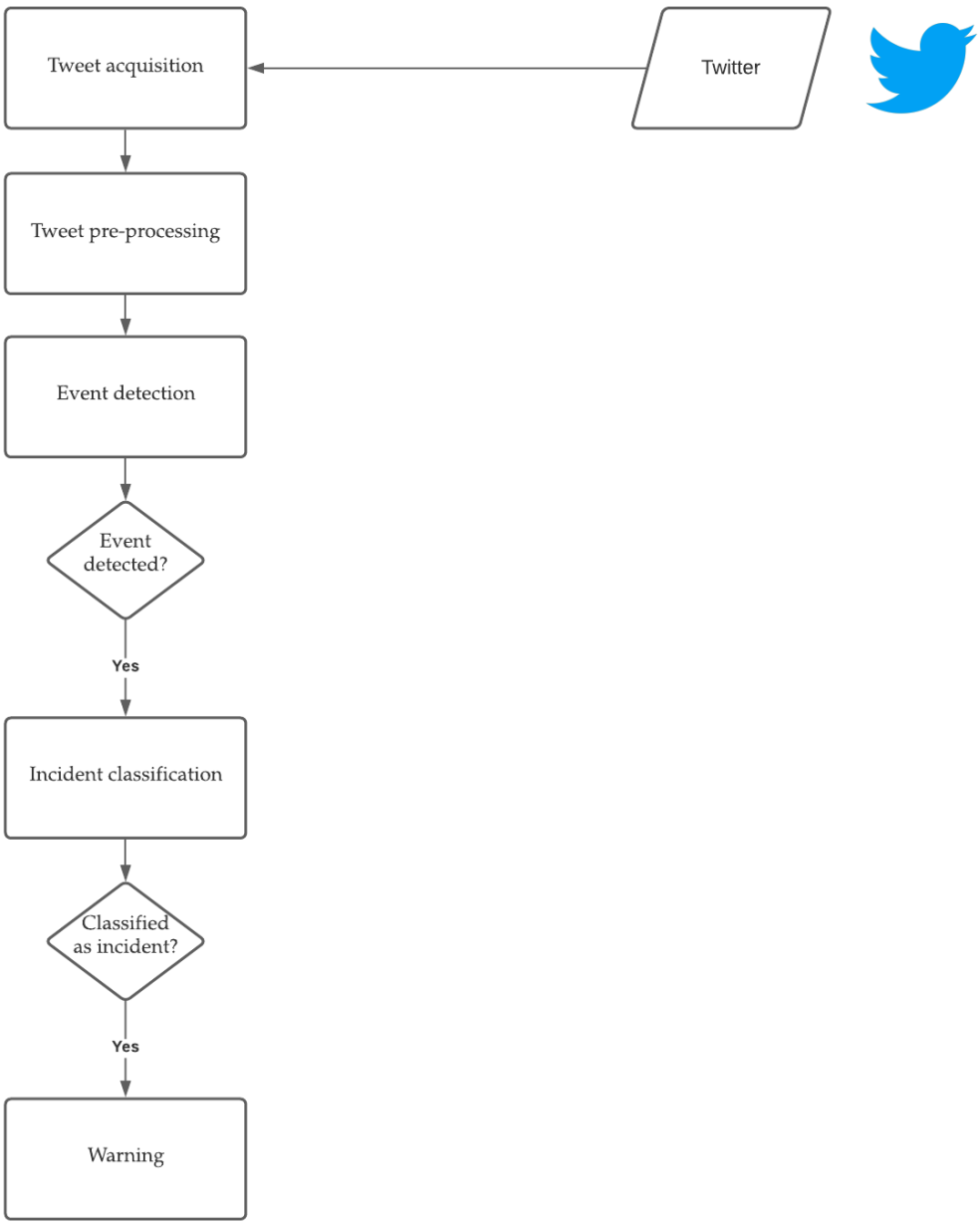

FIGURE 5.1: Designed system for automatically detecting incidents during protest demonstrations

is a significant increase in the number of tweets in the current time window compared to the historically weighted running average, and updating the historically weighted running average. By binning tweets to the current time window, the tweet arrival rate of each time window can be calculated. Then, the number of tweets in the current time window is compared with the historically weighted running average. If the number of tweets is significantly higher than the historically weighted running average, an event is detected. Moreover, at the end of each time window, the weighted running average will be updated. This process will be explained in more detail below.

First, a tweet will be binned to a specific time window, for example by 5 minutes. If no time windows exist yet, the first time window will be created and the start of the

time window will be set to the time at which the tweet was sent. If a time window does exist, it will be checked if the tweet falls inside the current time window. If the tweet falls inside the current time window, it will be added to the time window. Otherwise, the system will create a new time window and automatically add the tweet to the new time window.

Secondly, at the end of each time window, the number of tweets of that time window will be compared to the historically weighted running average. If the number of tweets in the current time window is more than two mean deviations from the current mean, and the number of tweets in the current time window is higher than in the previous time window (as employed by Marcus et al.), the increase is considered significant and an event is detected. Once an event is detected, the incident detection module will be invoked.

Thirdly, at the end of each time window, the historically weighted running average is updated. Given the number of tweets in the current time window, the mean and mean deviation values are updated, according to Equations 5.1, 5.2 and 5.3. These update steps require that $\alpha < 1$ (according to Marcus et al.). If no mean and mean deviance values exist, the mean will be initialized to the mean of the number of tweets in the last n_windows and the mean deviance will be set to the variance of the number of tweets in the last n_windows (we use n_windows = 5, as employed by Marcus et al.).

$$diff = |mean - n\_tweets\_in\_window| \tag{5.1}$$

$$newmean = \alpha * n\_tweets\_in\_window + (1 - \alpha) * mean \tag{5.2}$$

$$newmeandev = \alpha * diff + (1 - \alpha) * meandev \tag{5.3}$$

Incident classification

Once an event is detected, it needs to be determined whether that event is related to an incident. Therefore, all tweets in the time window of the event will be classified into [Incident-Related] and [Not Incident-Related] using a machine learning model. The machine learning model utilized in the system is the one with the highest F-Measure on the Incident-Related class of our modeling task (as described in Section 5.6). If one of the tweets in the time window of the event is classified as [Incident-Related], an incident is detected and a warning will be created. The choice for detecting an incident after one incident-related tweet is based on the results of the semi-structured interviews with OSINT analysts (as described in Section 6.3.5).

5.8.2 System evaluation

In order to determine whether the designed system can automatically detect incidents during protest demonstrations, the system will be evaluated using a dataset containing tweets from a similar protest demonstration as the machine learning models were trained on, but not previously used in the training of the machine learning models. The goal of the system evaluation is to detect incidents as soon as possible and automatically provide a warning once an incident is detected.

At 17:00, 3rd June 2020, a protest demonstration in Rotterdam (The Netherlands) was organized against police violence as a response to the death of George Floyd. It was estimated by the Dutch police that around 200-4,000 people would attend

the protest demonstration (Rijksoverheid, 2020). Eventually, it was estimated that around 4,000-5,000 people attended the protest demonstration. At the time of the protest demonstration, rules were imposed by the Dutch government to control the spread of COVID-19. These rules included that people needed to have a distance of 1.5-meters and it was not allowed to get together in groups larger than 4 people. However, according to several mass media websites (NOS, 2020; AD, 2020), people were not able to keep a distance of 1.5-meters and were thereby breaking the COVID-19 rules. Around 18:12, the mayor of Rotterdam and the Dutch police decided to end the protest demonstration. Because this protest demonstration was centered around the same topic as our training data and protesters were also breaking the COVID-19 rules, the tweets of the day of this protest demonstration will be used to evaluate the system. Moreover, because the Dutch police and the mayor of Rotterdam ended the protest demonstration around 18:12, our time window of interest is 14:00 - 18:15.

Tweets were extracted on 3rd June 2020, using the Twitter Search API (Standard level API access). Tweets containing the word "demonstratie" (Dutch for "demonstration") and written in the Dutch language (according to the Twitter API) were collected. For each tweet, the same variables were extracted as for the unique tweets dataset (as described in Section 5.3). Moreover, it must be noted that due to time constraints of this research project, this dataset was not labeled and therefore it is not known beforehand if this dataset contains any incident-related tweets. Also, because the complete tweets dataset (as described in Section 5.2) does not contain the full text of a tweet and does not have all relevant variables, it was necessary to extract the tweets again (just as with the unique tweets dataset). This was approximately performed 8 months after the creation of the complete tweets dataset and resulted in a decrease of tweets. This is a direct result of tweets that were deleted by users, users that were suspended by Twitter or users that deleted their Twitter account.

To evaluate the system, the dataset mentioned above was fed into the system using a web socket that provides the tweets in a streaming manner. When evaluating the system, two time windows were used, a one-minute time window and a five-minute time window. Larger time windows were not considered because of the nature of the system, which only detects incidents after the end of a time window. For example, if a 15-minute time window was utilized, and an incident-related tweet was sent at 17:02, it could be that the system would detect this incident 13 minutes later (at 17:15), because of the large time window. Because this situation is not preferred, only a one-minute time window and a five-minute time window were evaluated. Moreover, for each of the time windows, three values of α were tested: 0.125 (value proposed by Marcus et al.), 0.25 and 0.5.

Chapter 6

Results

6.1 Exploratory Data Analysis

In this section, the results of the exploratory data analysis will be described. Two datasets are explored: the complete tweets dataset and the unique tweets dataset. The complete tweets dataset consists of all tweets, but does not have all relevant variables, such as the full text of the tweet. The unique tweets dataset only consists of unique tweets (not retweets) and does contain all relevant variables. A more elaborate description of the two datasets is provided in Section 5.2. In Section 5.5, the scope of the exploratory data analysis is reflected upon.

6.1.1 Analysis complete tweets dataset

Using the Twitter Search API, 25,350 tweets were collected, sent by 11,228 users on 1st June 2020 containing the word "demonstratie" and written in the Dutch language. 17,094 tweets (67%) are retweets, meaning that Twitter users did not create the tweet themselves, but shared an existing tweet on their Twitter account. Moreover, 8,256 unique tweets (33%) were sent, meaning that the Twitter user did create the tweet themselves. Furthermore, only 8 tweets (0.03%) contained GPS coordinates and 348 tweets (1.4%) contained a user-determined place. A user-determined place does not necessarily indicate that the tweet is sent from that specific location. From those 348 tweets, only 29 tweets provided "Amsterdam" as the place (the place of the protest demonstration). Table 6.1 presents the descriptive statistics of the complete tweets dataset.

5,647 tweets (22%) contain a hashtag and in total 544 different hashtags were present in the tweets. The most used hashtags and the number of times they were

Metric	Value
Number of tweets	25350
Number of users	11228
Average number of tweets per user	2.26
Number of retweets	17094
Number of unique tweets	8256
Number of tweets with GPS coordinates	8
Number of tweets with user-determined place	348
Number of tweets containing a hashtag	5647

TABLE 6.1: Descriptive statistics of the complete dataset.

mentioned are presented in Figure 6.1. #halsema (used 1,993 times) and #femkehalsema (used 184 times) relate to the mayor of Amsterdam, Femke Halsema. #amsterdam (used 1,465 times) and #dam (used 152 times) relate to the place of the protest demonstration: Dam Square in Amsterdam. #blacklivesmatternl (used 297 times) relates to the subject of the protest demonstration, which was a protest against police violence as a response to the death of George Floyd. #anderhalvemeter (used 287 times) is Dutch for 1.5-meters and relates to the COVID-19 rules imposed by the government in The Netherlands (citizens should always have a distance of 1.5-meters to each other). #op1 (used 236 times) relates to the television program Op1 in The Netherlands, which had a broadcast on 1st June 2020 with Femke Halsema (mayor of Amsterdam) as a guest. Lastly, #coronavirusnederland (used 146 times) is Dutch for 'COVID-19 Netherlands' relating to the COVID-19 virus. These hashtags provide some general insights into the subject of the tweets, which seems to center around the mayor of Amsterdam, Black Lives Matter, the COVID-19 rules at the time of the protest demonstration and the television program Op1.

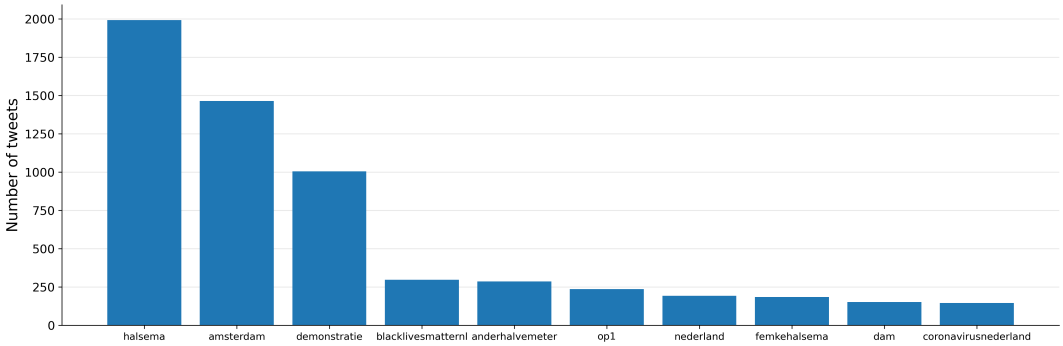

FIGURE 6.1: Most used hashtags and the number of tweets they were mentioned in (in the complete tweets dataset).

Furthermore, primarily ordinary Twitter users are sending tweets. In order to determine the type of a Twitter user, all Twitter users that tweeted or were mentioned, were extracted using the Twitter Search API, resulting in 9,833 users that tweeted and 3,063 mentioned users. Following, four groups of interest were established and internally labeled according to the labeling process described in Section 5.4.1, which resulted in a dataset of 1,475 labeled users. Results show that 98% of all tweets are sent by ordinary people, while only 2% of tweets are sent by specific user types, as presented in Table 6.2. Regarding the specific user types, media people (236 tweets), politicians (158 tweets) and mass media (91 tweets) account for the most sent tweets of user types. These are followed by writers (20 tweets), political parties (14 tweets), political organizations (6 tweets), political activists (5 tweets), employees at government organizations (3 tweets), comedians (3 tweets) and musicians (1 tweet).

User type	Number of sent tweets
Ordinary people	24813
Media people	236
Politician	158
Mass media	91
Writer	20
Political party	14
Political organization	6
Political activist	5
Part of government organization	3
Comedian	3
Musician	1

TABLE 6.2: Number of sent tweets per user type in the complete tweets dataset.

To further understand which accounts create the most reactions amongst Twitter users, the most mentioned users were identified. @andriesgknevel (Dutch TV presenter) was the most mentioned user with 1,012 mentions, followed by @telegraaf (a Dutch newspaper) with 955 mentions and @dijkhoff (politician) with 830 mentions. The most mentioned users and the number of times they were mentioned in tweets are presented in Figure 6.2. Together the 10 most mentioned users were mentioned 6,732 times, accounting for 27% of all tweets.

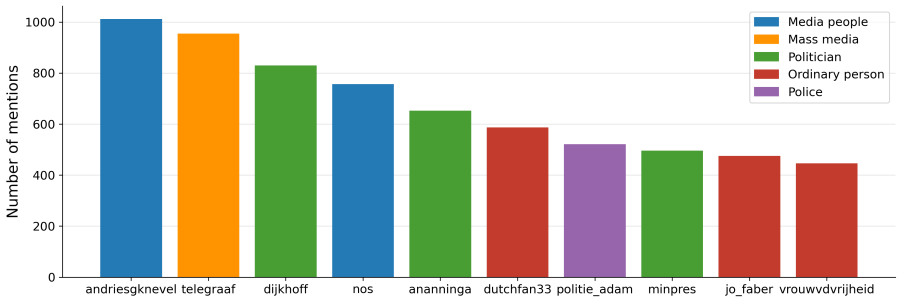

FIGURE 6.2: Most mentioned users and the number of times they were mentioned in tweets (in the complete tweets dataset). Colors depict the type of the user.

Moreover, we found that ordinary people was the most mentioned user type (10,370 mentions), followed by politicians (4,496 mentions), media people (3,745 mentions) and mass media (3,026 mentions), as shown in Figure 6.3. This shows that politicians, media people and mass media are together mentioned 11,267 times, accounting for 44% of all tweets. When comparing the user types of Twitter accounts that tweeted with the user types of Twitter accounts that were mentioned, it can be found that the mentioned user types are more equally distributed.

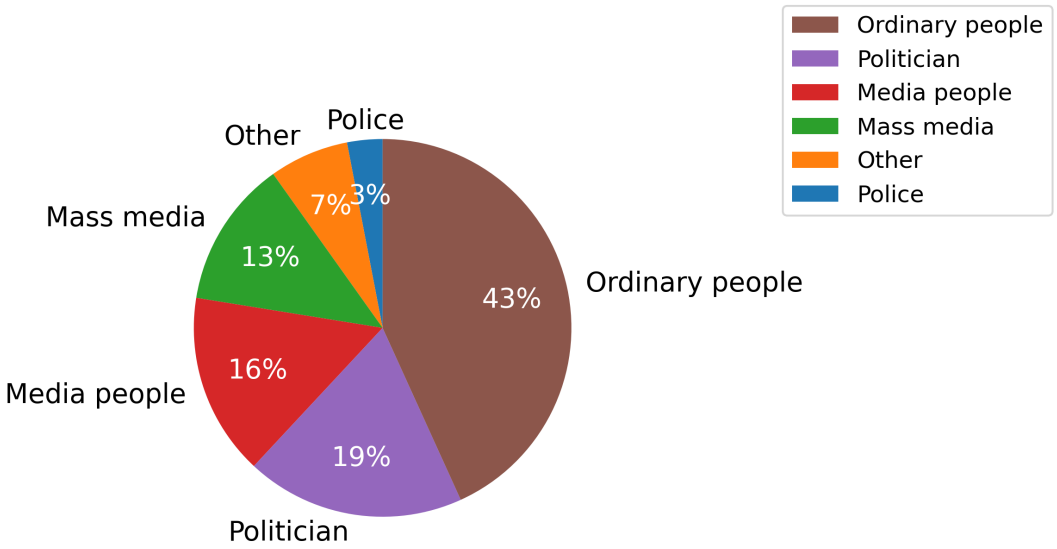

FIGURE 6.3: Percentages of mentions by user type (in the complete tweets dataset).

When analyzing the tweets over time, it can be noticed that during the day around 30-80 tweets are sent per 15 minutes. However, this suddenly increases at 16:45-17:00 to 132 tweets in 15 minutes, which is an increase of 83% compared to the previous 15 minutes (in which 73 tweets were sent), as presented in Figure 6.4. This finding shows that just before the start of the protest demonstration, more tweets are sent containing the word 'demonstratie'. Because between 15:00-16:00 the protest organizers make their preparations at Dam Square (presented in Appendix A), our time window of interest is 15:00 - 00:00. From the figure, it can be noticed that starting at 16:45 the number of tweets increases with almost every 15-minute time window, eventually peaking at 1,368 tweets in 15 minutes at 22:30. Around 15:00 - 17:15, almost 80% of all tweets are retweets, indicating that people are not sending unique tweets, but possibly agreeing with tweets that already exist. After 17:15, the number of retweets decreases to around 65% of all tweets, as presented in Figure 6.5.

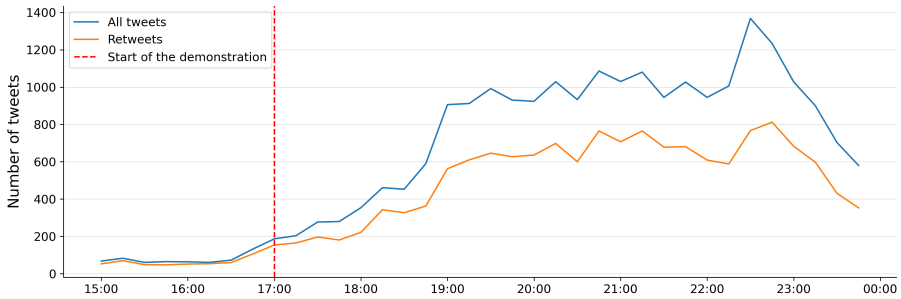

FIGURE 6.4: Number of tweets per 15 minutes on 1st June 2020 after 15:00 containing the word "demonstratie" written in the Dutch language.

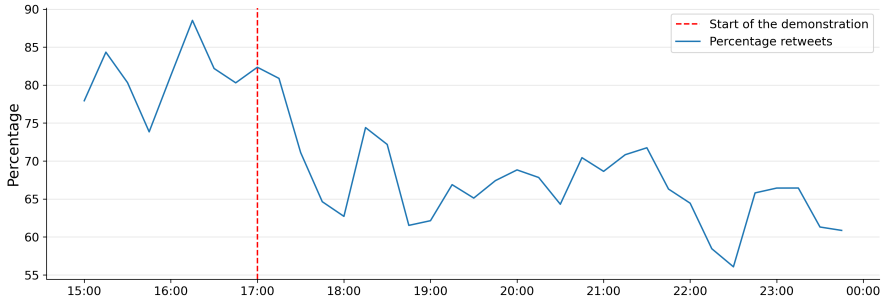

FIGURE 6.5: Percentage of retweets relative to all tweets per 15 minutes on 1st June 2020, after 15:00.

Following the approach of Hu et al. (2012), the five most mentioned user types are charted over time with their respective number of mentions, as presented in Figure 6.6. This shows that there are five "bumps" (sudden increases of specific user types), as depicted with the red lines under the x-axis of the figure, which will be discussed in further detail.

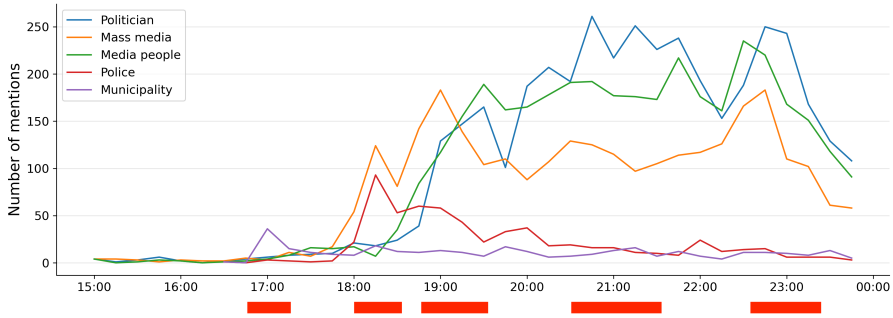

FIGURE 6.6: Number of mentions per 15 minutes mentioning a Twitter account from one of the categories. Red lines under the x-axis relate to sudden increase of mentions.

The first bump, starting just before 17:00, shows that the number of mentions regarding municipality accounts suddenly increases. At that time, the protest demonstration at Dam Square is about to start. However, upon manual examination of the tweets mentioning municipality accounts, it was found that these tweets are unrelated to the protest demonstration on 1st June 2020 in Amsterdam. One Twitter user shared a tweet that a protest demonstration was scheduled for the next week (on 7th June 2020) in Amsterdam, thereby mentioning the Amsterdam municipality Twitter account. This tweet was retweeted 33 times by other Twitter users, explaining the peak in mentions of municipality accounts.

Around 18:00 - 18:30, the number of mentions of mass media and police accounts increases. At that time, the protest demonstration at Dam Square is already active for an hour. Especially @NOS (mass media account, 138 mentions) and @politie_adam (police of Amsterdam Twitter account, 85 mentions) are mentioned often.

The third bump, starting at 18:40 and ending around 19:30, shows an increase in the number of mentions of mass media accounts, media people accounts and politicians. Regarding mass media accounts, the most mentioned accounts are @NOS (191 mentions), @telegraaf (141 mentions) and @at5 (97 mentions). When analyzing accounts from media people, three users are mentioned often: @Rhoogland (Dutch

journalist, 119 mentions), @SanderSas (Dutch journalist, 70 mentions) and @andriesgknevel (TV presenter, 53 mentions). Regarding politicians, especially @ANanninga (Dutch politician, 164 mentions) and @MinPres (prime minister of the Netherlands, 60 mentions) are mentioned often.

The fourth bump, starting around 20:30 towards 21:45, shows an increase in the number of mentions of politicians and media people accounts, while the number of mentions of mass media accounts decreases. Moreover, politicians are mentioned more often than people from the media. When analyzing the mentions of politicians, @dijkhoff is mentioned the most with 362 mentions, followed by @keesvdstaaij (215 mentions) and @ANanninga (139 mentions). Regarding media people accounts, the most mentioned accounts are @andriesgknevel (Dutch TV presenter, 328 mentions), @marcelvink888 (Dutch journalist, 100 mentions) and @SanderSas (Dutch journalist, 95 mentions).

The last bump, starting at 22:30 and ending around 23:15, shows an increase in the number of mentions of media people, politicians and mass media accounts. Regarding mass media accounts, @op1npo (Dutch TV program, 114 mentions), @telegraaf (104 mentions) and @adnl (mass media account, 66 mentions) are mentioned the most. When analyzing media people accounts, @andriesgknevel (120 mentions) is mentioned the most, followed by @SanderSas (73 mentions) and @jackvangelder (Dutch TV presenter, 37 mentions). Lastly, regarding politicians, @fransweisglas (155 mentions), @dijkhoff (115 mentions) and @keesvdstaaij (74 mentions) are mentioned most often.

6.1.2 Analysis unique tweets dataset

Using the Twitter Search API, 5,549 unique tweets were collected, sent by 3,859 users on 1st June 2020, containing the word "demonstratie" and written in the Dutch language. These tweets were manually labeled (according to the labeling protocol described in Section 5.4.2) into two classes: [Incident-Related] and [Not Incident-Related]. Only 2 tweets contained GPS coordinates and 257 tweets (4.6%) contained a user-determined place. From those 257 tweets, only 16 tweets contained "Amsterdam" (the place of the protest demonstration) as the place. Table 6.3 shows the descriptive statistics of the unique tweets dataset. Because the protest demonstration started around 17:00 and ended around 19:00, our time window of interest for this dataset is 15:00 - 20:00.

Metric	Value
Number of tweets	5549
Number of users	3859
Average number of tweets per user	1.43794
Average number of words in a tweet	34.2737
Average number of characters in a tweet	217.867
Number of tweets with GPS coordinates	2
Number of tweets with user-determined place	257
Number of tweets containing a hashtag	2021

TABLE 6.3: Descriptive statistics of the unique tweets dataset.

From the 5,549 unique tweets, 479 tweets (8.6%) were incident-related, while 5,070 tweets (91.4%) were not incident-related. None of the incident-related tweets were related to riots or violence, indicating that every tweet that is classified as

[Incident-Related] relates to not being compliant with the COVID-19 rules imposed by the Dutch government. This is in line with the expectations, because no acts of riots or violence have been reported during or after the protest demonstration. In Figure 6.7, the number of all tweets and incident-related tweets over time are presented. To provide better insights in the incident-related tweets over time, these are presented in 6.8. Remarkably, most incident-related tweets are not sent at the beginning of the protest demonstration (when people were already breaking the COVID-19 rules), but later on the day, showing peaks around 18:00 and 19:00.

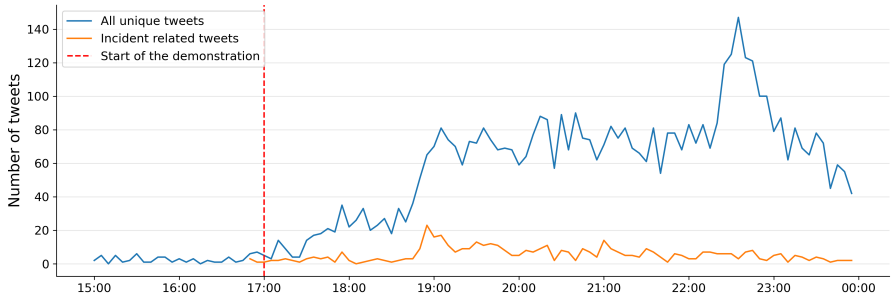

FIGURE 6.7: Number of unique tweets and unique incident-related tweets per 5 minutes on 1st June 2020, after 15:00.

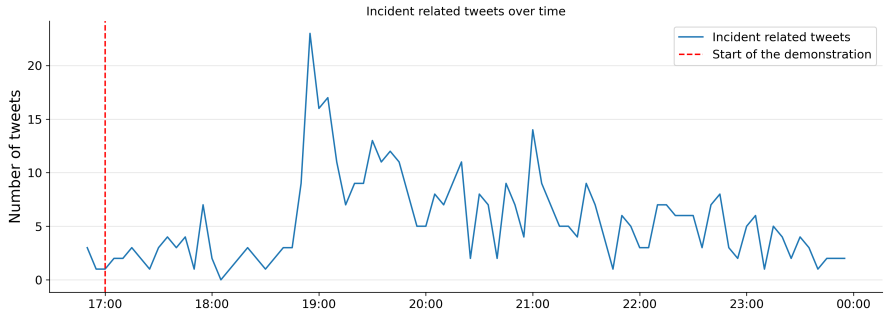

FIGURE 6.8: Number of unique incident-related tweets per 5 minutes on 1st June 2020.

While most incident-related tweets are sent by ordinary users (92%), as shown in Figure 6.9, the first incident-related tweet was sent by someone from the media (@_MikeMuller, a Dutch journalist). This tweet contained the following text: *"At the #BlackLivesMatter demonstration on Dam Square. Organization expected 250/300 attendees, but it is a lot busier on Dam Square"* with an accompanying video of Dam Square.

To further understand if the first incident-related tweets are shared by other Twitter users, the incident-related tweets are combined with the complete tweets dataset. In Table 6.4, the screen name, user type, time of the tweet and the number of user mentions at 18:00 and 19:00 of the first ten incident-related tweets are presented. From this table, it can be noticed that three users obtain the most mentions: @itisme_Patty (an ordinary Twitter user), @telegraaf (a mass media account) and @_MikeMuller (a journalist).

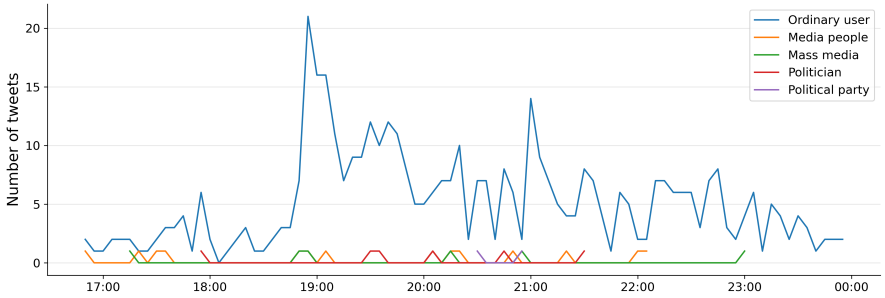

FIGURE 6.9: Number of unique incident-related tweets per 5 minutes on 1st June 2020, grouped by the type of the user.

Screen name	User type	Time	User mentions at 18:00	User mentions at 19:00
_MikeMuller	Media people	16:50	12	19
JosvanSon	Ordinary user	16:54	2	3
eelko76	Ordinary user	16:54	6	17
johnhoving	Ordinary user	16:55	0	0
macharoesink	Ordinary user	17:04	1	2
cockspan	Ordinary user	17:05	0	0
itisme_Patty	Ordinary user	17:06	39	88
BoschLaurens	Ordinary user	17:12	0	0
StaartjesL	Ordinary user	17:13	0	0
telegraaf	Mass media	17:17	12	51

TABLE 6.4: Screen name, user type, time of the tweet, and number of user mentions at 18:00 and 19:00 of the first ten incident-related tweets.

Moreover, we found that most incident-related tweets contain some form of media (an image, video or website with a thumbnail). In fact, from all 479 incident-related tweets, 382 tweets contain some form of media (79.7%). After performing a correlation analysis on the unique tweets dataset, it was found that that the variable *has_media* showed a correlation score of 0.56 with *has_Incident_Related*, indicating a strong correlation between whether a tweet contains some form of media and whether the tweet is incident-related. To further illustrate the number of tweets containing media over time, these are presented in Figure 6.10.

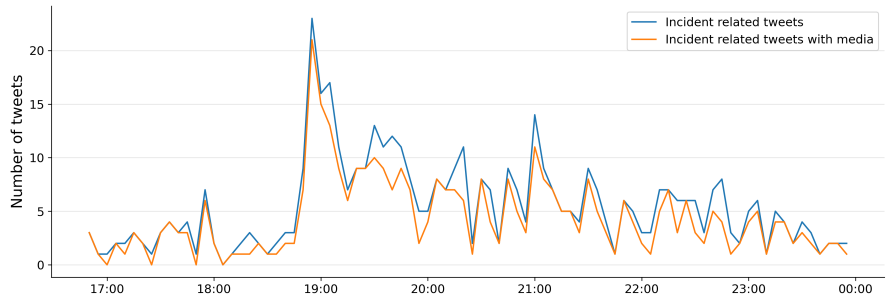

FIGURE 6.10: Number of unique incident-related tweets and unique incident-related tweets with media per 5 minutes on 1st June 2020.

When analyzing the incident-related tweets containing media, it shows that at the beginning of the protest demonstration most tweets share images or videos, as shown in Figure 6.11. At 17:17, the first article about the protest demonstration is shared by @telegraaf (10th incident-related tweet) on Twitter. From that moment, the number of tweets containing a website with a thumbnail increases, peaking at 19:00. Furthermore, when examining all incident-related tweets, we found that 54% of all incident-related tweets contained a link to an external website. Figure 6.12 shows the most mentioned websites, which account for 94,7% of all links in incident-related tweets. This shows that almost all websites in incident-related tweets are mass media websites.

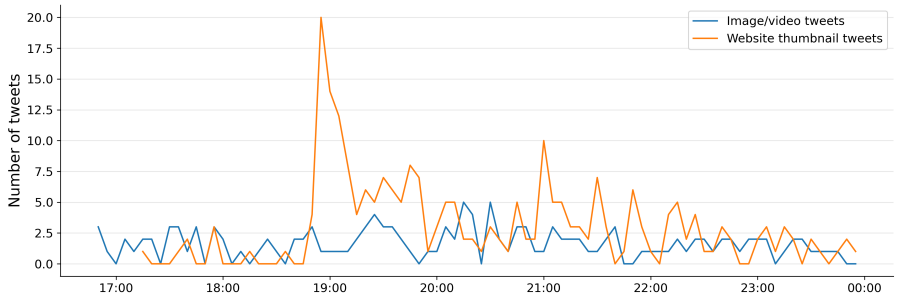

FIGURE 6.11: Number of unique incident-related tweets containing image/video and unique incident-related tweets containing website thumbnails per 5 minutes on 1st June 2020.

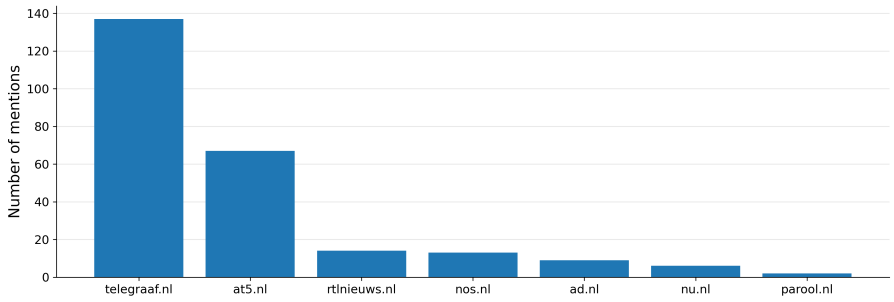

FIGURE 6.12: Most mentioned websites and the number of times they were mentioned in unique incident-related tweets.

When examining the individual tweets, the same trend can be observed. In the first 30 minutes after the first incident-related tweet (16:50 - 17:20), 12 incident-related tweets were sent. From those 12 tweets, almost all of them (10 tweets) are sharing some form of media. @telegraaf shares a URL (sharing their article about the protest demonstration) and the 9 other tweets (containing media) share images and videos of the protest demonstration.

Between 17:20 and 18:00, more Twitter users start sharing incident-related tweets. In the next 40 minutes, 25 incident-related were sent. Likewise, almost all tweets (21 tweets) in this timeframe share some form of media. 7 tweets share website URLs in this timeframe, leaving 14 tweets sharing videos or images. At this time, it can also be found that other mass media have published articles about the protest demonstration, such as AD (@ADnl on Twitter) and RTL Nieuws (@RTLnieuws on Twitter).

In the second hour of the protest demonstration (18:00-19:00), more Twitter users start sharing incident-related tweets, increasing from 37 to 51 tweets compared to the first hour. It seems that around 18:50, more users start sharing articles created by mass media in their tweets. Especially telegraaf.nl is mentioned often with 17 mentions.

During the last hour of the protest demonstration (19:00-20:00), the number of incident-related tweets peaked together with the number of incident-related tweets mentioning websites. This indicates that most incident-related tweets are sharing information about the incident, but also refer to a URL.

6.2 Modeling

In this section, the findings of the modeling task will be described. First, we will describe the more balanced dataset the machine learning algorithms were trained on. Following, each of the machine learning models will be evaluated according to the model evaluation described in Section 5.6.3.

6.2.1 More balanced dataset

The unique tweets dataset contained a class imbalance, with 9% incident-related tweets and 91% non incident-related tweets. Therefore, two additional more balanced datasets were created, following the methods described in Section 5.3.5.

The first dataset resulted in 6,499 tweets, where 5,070 tweets (78%) were not incident-related and 1,429 tweets (22%) were incident-related. This indicates that

8 back-translated tweets were the same as the original tweet and were therefore discarded. The second dataset resulted in 6,978 tweets, of which 5,070 tweets (73%) were not incident-related and 1,908 tweets (27%) were incident-related. To compare these three datasets, a simple Naive Bayes model was trained on each of them. The results of this comparison are presented in Appendix G and show that the second balanced dataset performs better in terms of F-Measure score on the Incident-Related class. Therefore, this dataset will be used to train the machine learning models as described in Section 5.6.

6.2.2 Model evaluation

As described in Section 5.6, for each algorithm two machine learning models are developed. One model is trained on the text of a tweet not containing the hashtag contents (referred to as model 1) and the other model is trained on the text of a tweet containing the hashtag contents (referred to as model 2). A complete overview of the findings can be found in Table 6.5, which presents the results of each performance metric for each machine learning model. For each metric, the highest scores have been printed in bold. F1 (Train) and F1 (Val) relate to the results of the 10-fold cross-validation task on the training data, the other metrics relate to predicting the data on the test dataset.

As can be observed from Table 6.5, all machine learning models provide an F-Measure between 0.85 and 0.92 on the Incident-Related class, which can be generally considered as a high F-Measure. The same is true for the AUC score, which is between 0.96 and 0.98 for all machine learning models. Moreover, across all algorithms, no large differences between model 1 and model 2 can be observed, indicating that using or not using the hashtag contents does not provide an effect on the predicting performance of the model.

Also, it shows that each machine learning model overfits on the training data. For almost all machine learning models, there is a difference of around 0.10 between the F-Measure on the training set and the validation set, while performing 10-fold cross-validation. Only the CNN models overfit less than the other models, but this could be the result of the number of epochs the models were trained on. Following the approach of Dabiri and Heaslip (2019), both CNN models were trained on a maximum of 20 epochs. Model 1 was trained on 19 epochs (because the 20th epoch did not provide an improvement on the validation set), while model 2 was trained on 20 epochs. For model 1 (as presented in Figure 6.13), it shows that at the 19th epoch the model starts to overfit with a difference of about 5 percentage points between the F-Measure on the training set and the validation set. The validation F-Measure provides a sudden drop, which can indicate overfitting. Model 2 (as presented in Figure 6.14 shows less overfitting, because the validation F-Measure remains stable and there is no sudden drop. While the CNN models overfit less, this could be the result of fitting the model on 20 epochs. When increasing the number of epochs, it could result in more overfitting, just as with the other machine learning models.

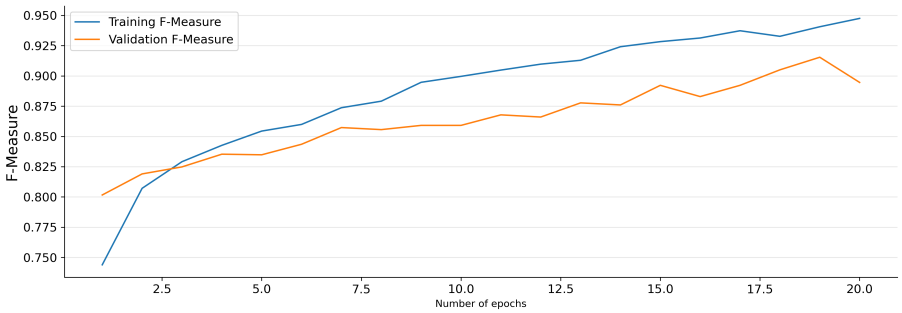

FIGURE 6.13: Performance of model 1 of the CNN algorithm in terms of F-Measure on the training set and validation set for varying number of epochs.

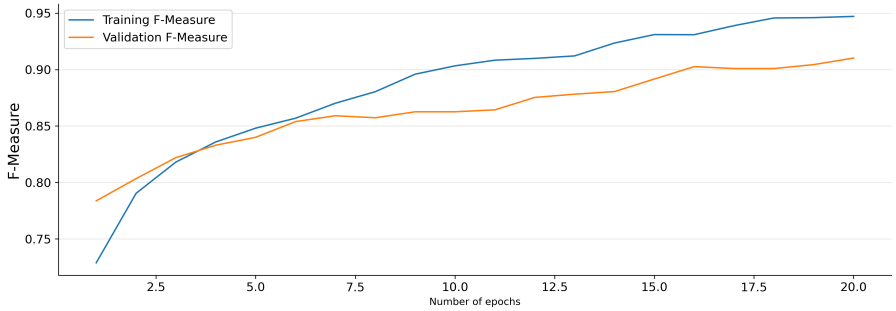

FIGURE 6.14: Performance of model 2 of the CNN algorithm in terms of F-Measure on the training set and validation set for varying number of epochs.

All in all, model 1 of the SVM algorithm provides the highest score of the F-Measure (0.915) on the Incident-Related class. Therefore, this machine learning performs better than the other models in terms of distinguishing incident-related tweets from non incident-related tweets. This model is closely followed by model 2 of the SVM algorithm (F-Measure of 0.914 on Incident-Related class) and model 1 of the Naive Bayes algorithm (F-Measure of 0.885 on Incident-Related class).

To further understand the differences between the machine learning algorithms, the most important feature words of the classification task are identified. For Naive Bayes, the most important feature words are determined by the log probabilities of the algorithm. For Logistic Regression, the most important feature words are determined by the coefficients of the algorithm. And for Gradient Boosted Decision Trees, the most important feature words are determined by the feature importance of the algorithm. To identify the most important feature words, the best performing model of each of the aforementioned algorithms was chosen in terms of the highest score of the F-Measure on the Incident-Related class. Regarding Logistic Regression, model 1 was chosen because both models show the same F-Measure on the Incident-Related class, but model 1 shows a higher F-Measure on the Not Incident-Related class. Both Support Vector Machines and Convolutional Neural Networks do not provide importance or coefficients of feature words of machine learning models and therefore the most important feature words could not be determined for those algorithms. Appendix H shows the ten most important feature words per algorithm. These results show that there is some overlap of importance of feature words between the

algorithms. For example, the words "dam" and "telegraaf" both occur in the most important feature words of Naive Bayes, Logistic Regression and Gradient Boosted Decision Trees. However, there are also some differences between the algorithms, such as "landinwaarts" (which only occurs in the most important feature words of Logistic Regression) and "verbijsterd" (which only occurs in the most important feature words of Naive Bayes). In Section 8.2, we will further elaborate on this finding.

Classification algorithm	Model	F1 (Train)	F1 (Val)	F1 (Not-Related)	F1 (Related)	Precision (Not-Related)	Precision (Related)	Recall (Not-Related)	Recall (Related)	AUC	Accuracy
NB	1	0.943	0.845	0.959	0.885	0.948	0.914	0.969	0.859	0.973	0.939
	2	0.946	0.850	0.957	0.882	0.949	0.904	0.965	0.861	0.974	0.937
LR	1	0.980	0.881	0.957	0.884	0.950	0.904	0.965	0.864	0.971	0.938
	2	0.995	0.879	0.956	0.884	0.957	0.883	0.956	0.885	0.973	0.936
SVM	1	0.997	0.911	**0.969**	**0.915**	0.958	**0.947**	**0.981**	0.885	0.975	**0.955**
	2	**0.998**	**0.914**	**0.969**	0.914	**0.959**	0.939	0.978	**0.890**	**0.977**	0.954
GBDT	1	0.997	0.878	0.956	0.884	0.957	0.883	0.956	0.885	0.966	0.936
	2	0.997	0.878	0.952	0.872	0.949	0.880	0.956	0.864	0.964	0.931
CNN	1	0.948	0.895	0.949	0.859	0.937	0.893	0.963	0.827	0.958	0.926
	2	0.947	0.910	0.945	0.845	0.929	0.890	0.963	0.804	0.958	0.926

TABLE 6.5: Performance metrics of the machine learning models. For each metric, the highest scores are printed in bold. F1 relates to the F-Measure. Model 1 is trained on the text of a tweet not containing the hashtag contents. Model 2 is trained on the text of a tweet containing the hashtag contents. Related refers to the [Incident-Related] class, Not-Related refers to the [Not Incident-Related] class.

6.3 Semi-structured interviews

In this section, the findings of the semi-structured interviews will be described. Three interviews with analysts working at the Open-Source Intelligence Team (OSINT) of the Dutch national police force were conducted. The process and questions of the interview are described in Section 5.7.

First, a description of the OSINT team will be provided and what its duties are in the scope of the Dutch national police force. Subsequently, the current process of social media data collection and analysis will be described and problems of OSINT analysts when performing social media analysis will be identified. Lastly, we will describe what an incident is during a protest demonstration (according to the OSINT analysts), when they want to receive a notification (if a system can detect an incident based on tweets), which information they want to receive by the system and how this information will be utilized. In the rest of this section, we will refer to the participants as P1, P2 and P3. Italic text between double quotation marks are quotes of the participants.

6.3.1 OSINT at the Dutch national police force

The Open-Source Intelligence Team (OSINT) at the Dutch national police force enriches police intelligence with open-source information. While this task is performed at all police units nationwide, the national team performs the complex and specialist work that is required. This always leads to an information product, which is an official police report (containing criminal offenses) or an intelligence product (which can come in many forms). An example of an intelligence product is the comparison of information available in police systems about a company and information about that company on the internet.

Furthermore, OSINT primarily performs two tasks: Monitoring and identification, as explained by P1: *"Monitoring is reasonably explainable, monitoring silent marches, actions, you name it. Identification is that, for example, we want to trace a criminal who threatens the prime minister with an online alias and we do not know who it is. But also identification of locations and photos, we want to know where it is. Then, we can set up online identification investigations."* Additionally, within the last years, there has been more emphasis on social unrest at the OSINT team. Regarding the topic of social unrest, the main goal is anticipation. For example, if the police discover signals that there is a call to loot, a police force could be deployed at a specific location so that dangerous situations can be prevented.

Moreover, P1 explained that there are five different OSINT levels, related to the conducted tasks at hand and the depth of those tasks:

1. Level 1 are desk clerks that obtain information, such as a video about someone being robbed. The desk clerk interprets the information and passes it to the other levels.

2. Levels 2 and 3 are the largest groups, consisting of officers that perform simple internet research, such as investigating where a demonstration is, if there is a threat on Facebook or when a silent march is announced on Twitter. Monitoring such cases is for levels 2 and 3, where level 3 has a more coaching role.

3. Levels 4 and 5 support specific investigations. This includes criminal investigations, but sometimes also proactive investigations (initiated by the police). For example, they analyze operational OSINT reports (created by level 2 and

3 officers) and translate them to strategic management information. Moreover, level 5 is specifically focused on scientific research, which investigates long-term phenomena, such as the influence of Trump or the influence of Huawei in the Netherlands.

To perform these analyses, OSINT uses open-source information, which mainly originates on the Internet. Anything that is 'open' via the Internet (meaning that everyone can access it without a password) could be used by OSINT analysts. Specific sources that are consulted include websites, social media (primarily Facebook and Twitter), Snapchat and Telegram.

6.3.2 Current process of social media data analysis

The initiation of social media data analysis at OSINT is primarily reactive. For example, when a police department gets information that Wilders (a Dutch politician) is coming to a city, a social media analysis is initiated. But this could also be started because of other signals, such as a trending topic on Twitter.

When collecting social media data, it is primarily performed manually. According to P1, there are some tools acquired by the Dutch police, but manual work is still required: *"With those tools, you have to enter what you want to filter on and that is quite difficult. For example, if you know that Wilders is coming to Venlo, then you type in 'venlo', 'wilders, 'extreme right, 'extreme left', which could be groups that will demonstrate there. Then if you enter it, you will get a certain result."* P2 agreed with this and added: *"It is mostly purely manual. On the internet, you have some tools that can provide some bulk information, for example, the number of tweets on a certain topic. So you do use some available tools to indicate quantities, but in the end, it is just reading and interpreting them manually."*

Moreover, the timeframe of a social media analysis by OSINT officers could vary from hours to days, which depends on the specific case at hand. Occasionally, analyses could take longer, because they are less urgent or cannot be mapped so quickly, as P2 notes: *"Sometimes I have analyzed movements, but based on the available data I was not able to provide a sufficient analysis. Then, you need to wait a couple of weeks for the right data."*

6.3.3 Problems of social media data analysis

During the semi-structured interviews, several problems related to social media data analysis were identified. First of all, P3 noted that social media data collection is bounded by the legal authorizations of a police officer. P2 elaborated on this subject and described several liberties and rights of citizens, such as the freedom of speech, freedom of demonstration and privacy of citizens. P2 noted: *"We are bounded by the legal authorizations within the use of Article 3 of the Police Act, which states what we are allowed to do. So we cannot just go and see all the tweets of a person, see what someone is saying on Facebook. And for our work, it does not really matter who says something, but more on what is said and what the consequences are. Therefore, we stay away from all forms of invasion of privacy or tracking specific people as much as possible, also because it provides no added value for us."* Furthermore, P2 added, is that there is a distinction in the amount of personal information between several social media platforms. For example, Facebook is a more personal platform than Twitter. Therefore, with Facebook, personal information is more easily obtained, which makes it more difficult.

Secondly, P1 described an additional problem. While legislation allows the police to perform online research with social media data, the use of social media data

by police violates the general terms Terms and Conditions of these social media platforms. As a consequence, companies are reserved when it comes to providing social media data to the police. In order to solve this problem, P1 states that you need to stay creative, which indicates that you need to perform the analysis manually or set up your own scraper (legitimately, according to law). But this still remains a *"cat and mouse game"*, because when a scraper is set up, it is often not working after a few months.

Thirdly, the amount of manual work related to social media data analysis is one of the largest problems at OSINT. Because of the limitations of automated online monitoring of social media, police officers are required to perform the social media analysis manually. But often, police officers cannot keep up with the amount of data, as P2 noted: *"If it is very busy, it is impossible to read. For example, there have been incidents at the beginning of the year with riots (called "avondklokrellen" in Dutch), then 8000-9000 tweets are sent per hour, that is 2/3 tweets per second, good luck! Then you are bound to your own capacity, to your capabilities to read those tweets. And the moment you look aside because you have seen something interesting, you are a few hundred tweets behind."* P1 and P2 also noted that some software packages are used at the Dutch police, but these are not sufficient because they do not filter enough messages, the sentiment of tweets is not indicated properly and it is not possible to cluster information.

6.3.4 What is an incident during a protest demonstration?

After general questions on social media data analysis, the participants were asked what they perceived as an incident during a protest demonstration. P1 primarily provided examples of incidents he/she is looking for, such as vandalism, looting, open violence, assault and threats, but no specific definition was provided. Also, P3 did not provide any definition, but said it was reliant on the knowledge and experience of the police officer: *"What we do notice now is that from the knowledge and experience of the team with doing the manual work, you actually get a 'fingerspitzengefühl' to know when something is about to happen. But we have not automated that yet. So we cannot say from an objective analysis 'if we have those elements, then we know things are going wrong'."*

Moreover, P1 and P2 stated that incident detection is not a specific subject of interest for OSINT teams, but only when these incidents cause social unrest. This is also related to adherence to the COVID-19 rules. While the Dutch police (in general) are interested when people are not adhering to the rules, OSINT is mainly interested in events causing social unrest, for example when someone starts a campaign against the COVID-19 rules or when thousands of people violate the COVID-19 rules. Therefore, the protest demonstration on 1st June 2020 in Amsterdam was of interest to OSINT. Also, P2 is interested when the expected behavior of people changes: *"There is a kind of normal behavior that you expect during events. But are there any signs that there are deviations from this? And what is that deviation? And what is the consequence of that deviation? That's what it mainly is for us."* On the other hand, as P2 mentions, incident detection is of interest for the Real-Time Intelligence Center (RTIC). Each region has its own RTIC department, which is in direct contact with police officers on the street to provide them with real-time up-to-date information.

6.3.5 When to receive a warning of an incident?

Following, the participants were required to review the incident-related tweets of the unique tweets dataset (described in Section 5.2) one by one, and were asked when they want to receive a warning of a potential incident (if an automated incident detection system was in place). As presented in Table 6.6, all participants wanted to receive a warning after one incident-related tweet. P1 and P3 stated that whether to receive a warning depends on the specific contents of the tweet at hand, while P2 stated that he/she always wants to receive a warning after one incident-related tweet. Additionally, it was asked why the participants want to receive a warning after one incident-related. P1 stated that the information can be used to potentially downsize the demonstration, which could be impossible after a few hours. P2 noted that by obtaining the warning, the tweet could be analyzed and eventual consequences could be identified.

Besides, all participants stated that they want to receive a warning in real-time, as soon as possible after the tweet(s) has been sent. Furthermore, the participants were introduced to the problem that a system could make mistakes and that the system could provide false warnings around 3-4 times per day. However, none of the participants perceived this as a problem. P2 even stated that he/she would rather receive ten false warnings than that one real warning would be missed. P1 noted that it did not matter, but it depends on the percentage of warnings that are real and false. Also, according to P1, it depends on the number of tweets that were related to the warning: *"It depends on the number of tweets. If I need to look at 50 tweets to authenticate a warning, it becomes less relevant to me."* P3 also stated that it did not matter, but it would be pleasant if he/she could indicate to the system that it was a false warning, so that the system could automatically learn from it.

Question	P1	P2	P3
After how many incident-related tweets, do you want a warning?	1	1	1
Is this dependent on the incident-related tweet?	Yes	No	Yes
When do you want to receive a warning?	Real-time	Real-time	Real-time
Are multiple false warnings on a day a problem?	No	No	No

TABLE 6.6: Answers of participants about when to receive a warning of a potential incident based on tweets.

6.3.6 What information of an incident?

When a system detects an incident based on tweets, the participants were asked what information they would want to receive from the system. As Table 6.7 shows, all participants want to receive the type of the incident, the location of the incident and the incident-related tweets the warning is based on. The reason why they also want to receive the incident-related tweets, is because they want to stay in control (stated by P1), interpret the situation better and receive more context (stated by P2 and P3), and to validate the source of the information (P2).

Moreover, P2 would like to obtain the date and time of the incident (just as P1), visual material of the incident (images and video) and to receive some information about the sentiment around the demonstration: *"Of course, it is very difficult to detect emotion in forms of text and social media, because you also have sarcasm and people who deliberately say things in a certain way. But it is very interesting to see what atmosphere it is related to. Can you tell what the atmosphere is like, based on certain tweets? Because the fact that it is busy in itself is not a bad thing in the first place. It could be very busy, but if there are signs that people are talking about the atmosphere or that it is a grim atmosphere, that*

Information	P1	P2	P3
Type of incident	✓	✓	✓
Date and time of incident	✓	✓	✗
Location of incident	✓	✓	✓
Visual material of the incident	✗	✓	✗
Sentiment around incident	✗	✓	✗
Future times and timelines	✗	✓	✗
Estimate of the crowd	✗	✗	✓
Where nearest police officers are	✗	✗	✓
How many police officers are on location	✗	✗	✓
Whether the situation is under control	✗	✗	✓
Incident-related tweets	✓	✓	✓
"Null information" on Twitter	✗	✓	✗

TABLE 6.7: Information OSINT analysts want to receive after detecting a potential incident based on tweets.

is, of course, very interesting." Additionally, P2 would also like to receive some "null information", such as when the topic is first discussed or when a certain hashtag is used for the first time on Twitter. Also, P2 stated that shared times and timelines are relevant to extract, also when future times are shared by Twitter users.

Furthermore, P3 stated that he/she also wants to receive an estimate of the crowd, where the nearest police officers are, how many police are already there and whether the situation is still under control.

Following, the participants were asked why they needed the information and what were they going to do with it. P1 and P2 both stated that the information would be analyzed and if there is indeed something going on, an information product would be created for management, which could make decisions based on that information, such as sending more police to a certain location. In addition, P2 also addressed that RTIC could use this information and would probably send the incident information as soon as possible to police officers on the streets. P3 would send the information of a potential incident to her supervisor, and when it is relevant enough, leadership of the police would be informed.

6.3.7 Automatic incident detection during protest demonstrations

Lastly, the participants were asked if a system that automatically detects incidents during protest demonstrations using Twitter data could potentially benefit them. All participants stated that it would benefit them, which is primarily because of the manual work that is currently required to perform the analysis. Moreover, P3 thought that it would also help to objectify incidents, because an incident is currently a *"gut feeling based on knowledge and experience"*. Additionally, P3 stated that the system could help with the local distribution of tasks. For example, when an incident is detected in Amsterdam, the system knows that Amsterdam belongs to the North Holland area and that the warning is automatically sent to the people who are responsible for North Holland.

6.4 System evaluation

To evaluate the system, a similar dataset (but not used for training the machine learning models) was fed into the system, as described in Section 5.8.2. Using the Twitter Search API, 1,894 tweets were collected, sent by 1,428 users between 14:00 - 18:15 on 3rd June 2020, containing the word "demonstratie" and written in the Dutch language. 1,182 tweets (62%) are retweets, meaning that the Twitter user did not create the tweet themselves, but shared an existing tweet of another user. The other 712 tweets were unique tweets, which account for 38% of all tweets. Furthermore, 0 tweets contained GPS coordinates and 35 tweets contained a user-determined place. From those 35 tweets, 6 tweets mentioned "Rotterdam" (the place of the protest demonstration) as the place. Table 6.8 presents the descriptive statistics of the dataset.

Metric	Value
Number of tweets	1894
Number of users	1428
Average number of tweets per user	1.33
Number of retweets	1182
Number of unique tweets	712
Number of tweets with GPS coordinates	0
Number of tweets with user-determined place	35

TABLE 6.8: Descriptive statistics of the dataset used for system evaluation.

The tweets mentioned above were fed into the system in a streaming manner. Two time windows were analyzed (one-minute time window and five-minute time window) and for each of the time windows, three values of α were tested (0.125, 0.25 and 0.5). Results showed that using a five-minute time window and a α value of 0.125 showed the best results in terms of detecting the most events. Increasing α results in less detected events. As shown in Figure 6.15, the system is capable of detecting events when a sudden increase in tweets is occurring. Only the event detected at 17:25 seems trivial, because the increase in tweets at that time does not seem significant. Moreover, the first incident is detected at 17:50. In that time window, an event is detected and one incident-related tweet is observed by the system. This incident-related tweet was sent at 17:45, containing the text *"LIVE | Thousands of demonstrators at #Erasmusbrug via @Telegraaf #Rotterdam #Demonstration #BlackLivesMatter"* with a link to a website article named *"Violation of COVID-19 rules during demonstration in Rotterdam"*. This indicates that the system can detect an incident 5 minutes later than observing the first incident-related tweet about it. Consequently, three other incidents were detected in the time window of 17:50 (detected at 17:55), 18:05 (detected at 18:10) and 18:10 (detected at 18:15).

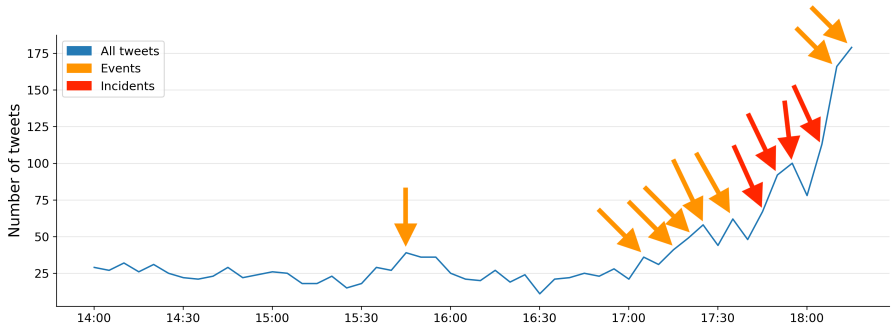

FIGURE 6.15: Events detected using the system on a five-minute time window with value α value of 0.125. Orange arrows indicate detected events, red arrows indicate detected incidents.

When lowering the time window from five minutes to one minute, more events are detected, but fewer incidents. In total, 24 events are detected when using a one-minute time window with a value of 0.125 for α, as shown in Figure 6.16. However, only one of these events is classified as an incident. That incident is detected at 17:50, based on a tweet sent at 17:50, containing the same text and link to a website as the tweet mentioned above (although it is a different tweet). These results show that lowering the time window influences the number of incidents that are detected (fewer incidents), but do not influence when these incidents are observed (both time windows perceive the incident at the same time).

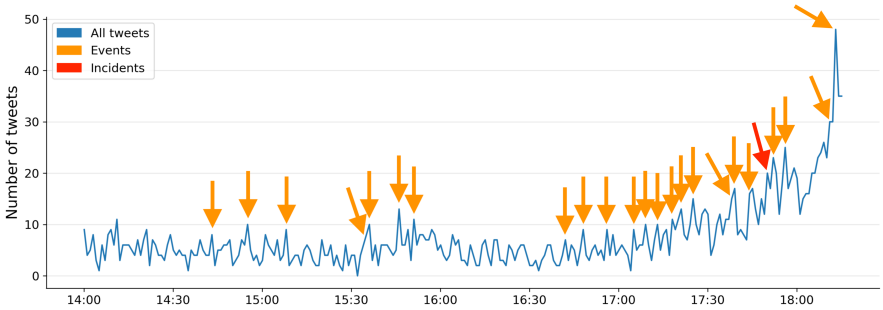

FIGURE 6.16: Events detected using the system on a one-minute time window with value α value of 0.125. Orange arrows indicate detected events, red arrows indicate detected incidents.

Knowing that the protest demonstration was ended approximately at 18:15 by the Dutch police and the mayor of Rotterdam, detecting the first incident at 17:50 does not appear as "early". Therefore, the dataset was manually examined to analyze whether the dataset did contain any incident-related tweets before 17:45. Upon manual examination of the tweets in the dataset, it seems that the first incident-related tweets are sent around 17:15-17:20, with tweets containing the following text: *"The images I see of the demonstration are not cheerful. Distance please! #demonstration #Rotterdam #Erasmusbrug"* and *"Although the organization has called for people to keep a distance of 1.5 meters, this turns out to be quite difficult in practice"*. However, the system was not able to classify these tweets as incident-related and therefore did not detect an incident. When further analyzing this finding, the labels of all tweets in the dataset were predicted using the machine learning model of the system. This resulted in detecting 10 incident-related tweets. Five of those incident-related tweets were found

by the system. The other five tweets were not found, because the system did not detect an event during that time. Therefore, it seems that the used machine learning model is not capable of detecting the incident-related tweets in this dataset. In Section 8, we will further elaborate on this finding.

Chapter 7

Conclusion

In this chapter, the conclusions to the three research questions and the research objective are provided.

7.1 Conclusion of research questions

In Section 3, three research questions were formulated. In this section, we will answer the research questions based on the results (as described in Section 6).

> **[RQ1]**: *"What phases of Twitter coverage after an incident during a protest demonstration can be identified?"*

Based on the exploratory data analysis, four phases of Twitter coverage are identified. The first phase consists of several Twitter users tweeting independently about the incident during the protest demonstration, thereby sharing some form of media (such as video or images). These tweets can come from multiple types of users, but are most often shared by ordinary users.

In the second phase, approximately starting 30 minutes after the first phase, more Twitter users start sharing incident-related tweets. These tweets often contain a form of media, such as images, videos and website URLs. Moreover, during this phase, mass media Twitter accounts and police Twitter accounts are mentioned the most by other Twitter users.

During the third phase, approximately starting one hour after the second phase, the number of incident-related tweets peaks. Also, these tweets often contain a URL to a website page and contain less often a video or image. Furthermore, mass media accounts, media people accounts and politicians are mentioned the most by other Twitter users during this phase.

In the last phase, approximately starting 90 minutes after the third phase, the number of incident-related tweets decreases over time. Just as with the previous phase, mass media accounts, media people accounts and politicians are mentioned the most by other Twitter users. Moreover, the number of mentions of these user types slightly increases during this phase compared to the last phase.

> **[RQ2]**: *"What are the differences in performance between several standard supervised classification algorithms in the context of incident-related tweet prediction during protest demonstrations?"*

Based on the evaluation of the machine learning models, we found that Support Vector Machines (SVM) provides the best performance in the context of incident-related tweet prediction during protest demonstrations, because it provides the highest F-Measure on the Incident-Related class (compared to the other algorithms).

If we analyze the algorithms that used counted vectors as input (NB, LR, SVM, GBDT), we found that there are not many differences between them. All algorithms show some overfitting on the training data, a relatively high F-Measure and provide a high Area Under the ROC Curve.

When comparing the algorithms that used counted vectors with convolutional neural networks, we observe that, in general, the counted vector algorithms provide a higher F-Measure on the Incident-Related class, although the difference is small. Moreover, we found that the convolutional neural networks overfit less on the training data than the other algorithms.

Lastly, it can be observed that including or not including the hashtag contents in the training data does not have any significant impact on the performance of the machine learning models.

> **[RQ3]**: *"What is the information need of OSINT analysts at the Dutch national police force when automatically detecting incidents using Twitter data?"*

When a system automatically detects incidents based on Twitter data, OSINT analysts want to receive a warning as soon as possible. This means that they already want to receive a warning after one incident-related tweet. Moreover, they want to receive the warnings real-time, as soon as possible after the incident-related tweets have been sent.

When receiving a warning of an incident, the analysts both wanted summary information about the incident and the specific tweets this information was based on. The requested summary information includes what type of incident happened, the date and time of the incident, and the location of the incident. Other suggestions include visual material extracted from tweets (images and videos), general sentiment around the incident, shared times and timelines, and information about the police presence. In addition to summary information, the analysts also want to obtain the individual tweets, because they want to stay in control, interpret the situation better and receive more context, and to validate the source of the information.

7.2 Conclusion of research objective

The objective of this research project was to automatically detect incidents during protest demonstrations by using Twitter data. In order to realize this goal, current literature on social movements, protest demonstrations, Twitter and incident detection was examined. Next, an exploratory data analysis was performed to understand what tweets were sent by Twitter users and to identify the phases of Twitter coverage after an incident during a protest demonstration. Furthermore, several machine learning models were trained and tested on a protest demonstration-related dataset. Additionally, semi-structured interviews with OSINT analysts of the Dutch national police force were conducted. Finally, a system was designed and developed that automatically extracts tweets from Twitter, detects incidents and sends a warning when an incident was detected.

We argue that the objective of this project has been achieved. As shown in Section 6.4, the designed system was able to automatically detect incidents on another dataset. On this dataset, the first incident was detected 50 minutes after the start of the protest demonstration. However, we do think the system and developed machine learning models could be improved, as will be discussed in Section 8.

Chapter 8

Discussion

In this chapter, the results of this research project will be reflected upon, improvements for future research projects will be identified, limitations of the project will be addressed and ideas for future work will be proposed.

8.1 Interpretation of four phases

In this section, the four phases of Twitter coverage after an incident during a protest demonstration are elaborated on. In order to interpret these phases, the tweets in the datasets were manually examined.

In the first phase, several Twitter users tweet independently about the incident. In those tweets some form of media (such as video or images) is shared, indicating that Twitter users deem it important to share the real-life situation with other Twitter users. Also, it is interesting to find that the first incident-related tweet was sent by someone from the media, showing that media people are 'on top of the news'.

During the second phase, more people share incident-related tweets. Just as with the first phase, these tweets can contain some form of media. However, in this phase, also website URLs are shared, which did not happen in the previous phase. This is probably because websites did not have any published articles about the demonstration in the previous phase. Moreover, the number of mentions of mass media and police accounts increases during this phase. Upon manual examination of the tweets, it seems that Twitter users are finding out about the incident during the protest demonstration through mass media, explaining the increase in mentioned mass media accounts. Furthermore, users express their concerns and ask why the Dutch police are not intervening in the protest demonstration (because people are not compliant with the COVID-19 rules) relating to the increase in police accounts. Also, it is important to note that users are not tweeting much about that an incident has happened (or is happening), but more about their concerns and opinions about the incident.

In the third phase, the number of incident-related tweets peaks together with the number of websites shared within the incident-related tweets. During this phase, the number of mentions of mass media accounts, media people accounts and politicians increases. Upon manual examination, it was found that more people are finding out about the news through mass media, retweeting the news and tweeting their concerns and opinions about it. Moreover, some people from the media have expressed their concerns about the incident, resulting in mentions by other Twitter users. Furthermore, politicians are mentioned often by Twitter users, which could be because of two reasons. On the one hand, some politicians have expressed their concerns about the incident on Twitter, and other Twitter users retweet those tweets or reply to them. On the other hand, individual users are mentioning politicians and express

their concerns about the incident and the decisions made during the protest demonstration.

In the last phase, the number of incident-related tweets decreases, while the number of mentions of politicians, mass media accounts and media people accounts increases. This could indicate that the content of the tweets is shifting more towards a discussion on Twitter. Upon manual examination of the tweets, it seems that the contents of the tweets remain the same, users expressing their concerns about the protest demonstration and the non-compliance with COVID-19 rules. Regarding politicians, more politicians have shared their concerns about the protest demonstration, resulting in mentions by other Twitter users. With mentions of media people accounts, the same trend can be observed. The subject and the approach of the tweets remain the same (people are concerned about the protest demonstration and non-compliance with COVID-19 rules), but just more media people are tweeting about it, resulting in more mentions. Also, users are responding to the appearance of the mayor of Amsterdam (Femke Halsema) at the Dutch television program Op1, which started at 22:15 (as presented in Appendix A). In this program, Femke Halsema explained some of the decisions that were made during the protest demonstration. As a consequence, media people are tweeting about the appearance of Femke Halsema, and getting retweeted or replied to.

8.2 Interpretation of machine learning algorithms

After evaluating the performance of the machine learning algorithms, it becomes interesting to understand why the differences in performance of the machine learning algorithms occur. One reason could be noise in the training data, which occurs because of text disfluencies, such as spelling errors, abbreviations and non-standard words (Agarwal, Godbole, Punjani, & Roy, 2007). It could be that some algorithms are better at handling noisy data than others. For example, the study of Agarwal et al. (2007) found that Support Vector Machines perform better in terms of accuracy than Naive Bayes in the context of text classification when handling noisy text data.

Moreover, as shown in Section 6.2.2, the algorithms provide different importance weights for feature words in the classification. This difference in importance weights for feature words could explain the differences in the performance of the machine learning algorithms. To further understand why differences in the performance of machine learning algorithms occur in the context of incident-related tweet prediction, future work is required.

8.3 Unexpected results

During this research project, some unexpected results were discovered. First of all, the obtained phases of Twitter coverage after an incident during a protest demonstration were inconsistent with the frameworks of Klein et al. (2012) and Hu et al. (2012). In the proposed framework of Klein et al., first, several witnesses of an incident will tweet independently about the incident, which is then spread by their followers. As the last phase, this is picked up by the mass media, several hours later. However, during our exploratory analysis, we found that someone from the media sent the first incident-related tweet and that a mass media account already covered the incident on Twitter approximately 20 minutes after the start of the protest demonstration. This indicates that mass media does not pick up the incident several hours later, but covers the incident pretty soon after it occurred. Regarding the

framework of Hu et al. (2012), first, people from the media are mentioned the most, followed by mass media accounts (in the second phase) and celebrities (in the third phase). However, we found that that mass media and police accounts are mentioned the most in the second phase, followed by mass media, media people and politicians (in the third and fourth phase). There could be multiple underlying reasons for the difference in results. While Hu et al. was focused on breaking news on Twitter, this research project was focused on the coverage of incidents on Twitter. This difference in the scope of the studies could explain why different phases are obtained. Moreover, Hu et al. uses Twitter data from the United States, while this study uses Twitter data from the Netherlands. There could be a difference in how Twitter is used between countries, which could explain a difference in the obtained phases. Also, Hu et al. only took the 100 most mentioned users into account, whereas this study had a more elaborate approach by focusing on four user groups, which could have an impact on the obtained phases.

Secondly, the findings of our machine learning models are inconsistent with the results of Dabiri and Heaslip (2019). While the convolutional neural networks of Dabiri and Heaslip improve over state-of-the-art methods, such as Naive Bayes and Support Vector Machines (SVM), our results show that algorithms trained on TF-IDF vectorized matrices (Naive Bayes, SVM, Logistic Regression and Gradient Boosted Decision Trees) provide better performance results in the context of incident-related tweet prediction during protest demonstrations. A multitude of reasons could underlie this difference in findings. Whereas Dabiri and Heaslip focuses on traffic incidents, this research project focuses on incidents during protest demonstrations. There could be differences in the type of tweets that are sent in each of the scopes, for example, tweets in the context of traffic incidents could be more precise. Furthermore, in the study of Dabiri and Heaslip more data was used. In total, 51,100 cases were used to train the machine learning models, while in this study, 6,978 cases were used to train the machine learning models. Additionally, the dataset used by Dabiri and Heaslip showed no class imbalance (50/50 ratio when using two classes), which is in contrast with our dataset, which showed a class imbalance (27/73 ratio). Because of this class imbalance, several balancing tasks were performed, which was not the case in the study of Dabiri and Heaslip. Moreover, the labeled dataset used in this study was primarily labeled by one person, possibly resulting in bias on the dataset. Lastly, the machine learning models of Dabiri and Heaslip were trained on English text using English word vectors, while the machine learning models of this research project were trained on Dutch text using Dutch word vectors. Differences between the qualities of these word vectors may exist.

Thirdly, the findings of the system evaluation on a dataset from a similar protest demonstration as the machine learning models were trained on, showed unexpected results. While the used machine learning model (model 1 of SVM) showed a high score of the F-Measure on the Incident-Related class of the test set during the modeling task, it was only able to detect 10 Incident-Related tweets on the dataset used for system evaluation, while it seems that there are more Incident-Related tweets in this dataset. An underlying reason for this finding is that the used machine learning model contains high variance, which is consistent with our findings, because the results of the modeling task show that the model showed some overfitting on the training dataset. Another underlying reason could be that there is bias in the labeled dataset. Moreover, it is important to note that the dataset the system was evaluated on, was not labeled. Therefore, it could be that this dataset does not contain many Incident-Related tweets.

8.4 Improvements of results

To improve the performance of the machine learning models in future research projects, several ideas are proposed. First of all, the machine learning models could be trained on more data, and preferably, on data from multiple protest demonstrations where incidents have occurred. Thereby, the machine learning models could learn from more data, which could results in less overfitting on the training data and better performance of the machine learning models on unseen data. Secondly, the machine learning models that use TF-IDF vectorized matrices could use bigram (two-word occurrences) and trigram (three-word occurrences) vectorized matrices to train the data on, instead of unigrams, which could have a positive impact on the predicting performance of the machine learning models. Thirdly, other data balancing tasks could be performed, such as utilizing Synthetic Minority Over-sampling Technique (SMOTE) or other data augmentation techniques, such as random swapping of words and random deletion of words. Moreover, a different word vector could be used for the convolutional neural networks, in addition to the FastText vector that was used in this research project. Furthermore, in addition to convolutional neural networks, other neural networks could be used to train the data on, such as Long-Term Short-Term Memory (LSTM) recurrent neural networks (as used in Dabiri and Heaslip (2019)). The reason why LSTM was not utilized in this research project, is because Dabiri and Heaslip (2019) found better performance with convolutional neural networks. Lastly, the convolutional neural networks were trained on a maximum of 20 epochs (following the approach of Dabiri and Heaslip). By increasing the number of epochs, the neural network could provide better performance. But likewise, it could also result in more overfitting on the training data.

In order to improve the designed system in future research projects, several improvements are suggested. First and foremost, the designed system only extracts tweets containing the word 'demonstratie'. It would be pleasant if the system could automatically track additional keywords on Twitter based on the subject of the demonstration. For example, during the exploratory analysis, we found that many tweets contained "Black lives matter" and "BLM" in the tweet text. By adding those keywords dynamically to the system per protest demonstration (e.g. by creating an integration with a system of the Dutch police), more tweets could be acquired and possible incidents could be identified sooner. Secondly, the designed system now only provides a warning when an incident is detected containing the tweet ids of the incident-related tweets. In the future, an information extraction module could be implemented that automatically extracts information from those tweets and uses that information when sending warnings. Due to time constraints, this module was not developed in this research project. Likewise, the introduction of a feedback module could be implemented. If humans could provide feedback to the system, for example when the system incorrectly detects an incident, the system could automatically learn from this feedback and improve itself. Furthermore, the system would benefit if humans could interact with it through a dashboard, which is covered in the work of Van der Plaat (2021), where a dashboard is developed for analysts of the Dutch police based on detected incident-related tweets. Finally, the system could benefit from a tweet acquisition module that adapts to the contents of the tweet. For example, if a hashtag is used often by Twitter users, all the tweets containing that hashtag in a specific time window could be acquired. Similarly, when someone responds with 'demonstratie' to a tweet not containing the term 'demonstratie', this last tweet could also be automatically acquired. During our exploratory analysis, we found some tweets that did not contain the term 'demonstratie', but were related

to the protest demonstration. In order to acquire those tweets, an adaptive tweet acquisition module could be implemented.

8.5 Limitations and future work

The largest limitation of this research project is the data. The complete tweets dataset did not contain the full text of a tweet and did not have all the relevant variables of a tweet. Therefore, the unique tweets dataset was acquired, resulting in a decrease of 2707 unique tweets (because this dataset was acquired approximately 6 months after the complete tweets dataset). Because of this decrease in unique tweets, the machine learning models were trained on less data, possibly negatively impacting the performance of those models. The same is true for the dataset used for system evaluation, which did not contain all the tweets that were sent on 3rd June 2020 between 14:00-18:15 (a decrease of 422 tweets), possibly resulting in more biased results. Moreover, the answers to our research questions are based on a dataset of one protest demonstration. Future research is necessary to determine whether other protest demonstrations show the same tendencies.

Another limitation of this study is the context of the protest demonstration. The protest demonstration has taken place during a COVID-19 pandemic with the incident of protest participants not keeping a distance from each other and being in groups larger than four. During "non-COVID" times, this would not be considered an incident. Therefore, the results of our study are mainly applicable to COVID-19 related incidents during protest demonstrations in a COVID-19 pandemic. Although the application scope could be observed as small, this research can be considered as a new step towards the automatic early warning of incidents during protest demonstrations.

Furthermore, as a result of our exploratory data analysis, we found a strong correlation between the use of media in a tweet and whether the tweet was incident-related (correlation score of 0.56). Therefore, it appears interesting to train a machine learning model on both the text of a tweet and the media contained in a tweet. Future work must show whether this improves the predicting performance of a machine learning model in the context of incident-related tweet prediction during protest demonstrations.

Likewise, a machine learning model could be trained on a combination of sentiment and incident-relatedness. From the findings of the semi-structured interviews, it was found that OSINT analysts are interested in the sentiment of a tweet. For example, if someone replies positively to an incident-related tweet, what does this imply? To further investigate this phenomenon, future work could be focused on the combination of sentiment in tweets and incident-relatedness.

Moreover, the semi-structured interviews with OSINT analysts revealed that the designed system could also be used by analysts of the Real-Time Intelligence Center (RTIC) at the Dutch national police force. Therefore, future work is necessary to identify the information need of RTIC analysts at national police forces when automatically detecting incidents using Twitter data.

Lastly, it was found that the location of an incident is especially important for OSINT analysts. However, only a small portion of tweets contains GPS coordinates of a Twitter user. Moreover, some Twitter users share a user-determined place, but this is not a reliable indicator, because it does not necessarily reflect the actual location of the tweet. Therefore, future research is required to automatically extract the

location of a tweet based on the tweet text. In the work of Vos (2021), the focus is on automatically extracting the location of a tweet based on the tweet text.

References

ACLED. (2020, Sep). *The rise of social media.* Retrieved March 20, 2021, from https://acleddata.com/acleddatanew/wp-content/uploads/2020/09/ACLED_USDataReview_Sum2020_SeptWebPDF_HiRes.pdf

AD. (2020). *Politie grijpt in na afbreken black lives matter-protest in rotterdam: 'je voelt de tranen en boosheid'.* Retrieved June 17, 2021, from https://www.ad.nl/nieuws/politie-grijpt-in-na-afbreken-black-lives-matter-protest-in-rotterdam-je-voelt-de-tranen-en-boosheid-br~a7483970/

Adedoyin-Olowe, M., Gaber, M. M., & Stahl, F. (2013). A survey of data mining techniques for social media analysis. *arXiv preprint arXiv:1312.4617*.

Agarwal, S., Godbole, S., Punjani, D., & Roy, S. (2007). How much noise is too much: A study in automatic text classification. In *Seventh ieee international conference on data mining (icdm 2007)* (pp. 3–12).

Aggarwal, C. C. (2018). An introduction to cluster analysis. In *Data clustering* (pp. 1–28). Chapman and Hall/CRC.

Aggarwal, C. C., & Zhai, C. (2012). A survey of text clustering algorithms. In *Mining text data* (pp. 77–128). Springer.

Alsaedi, N., Burnap, P., & Rana, O. (2017). Can we predict a riot? disruptive event detection using twitter. *ACM Transactions on Internet Technology (TOIT)*, *17*(2), 1–26.

Ansari, A. (2012). The role of social media in iran's green movement (2009-2012). *Global Media Journal-Australian Edition*, *6*(2), 1–6.

Anstead, N., & O'Loughlin, B. (2015). Social media analysis and public opinion: The 2010 uk general election. *Journal of Computer-Mediated Communication*, *20*(2), 204–220.

Armstrong, C. L., & Gao, F. (2010). Now tweet this: How news organizations use twitter. *Electronic News*, *4*(4), 218–235.

Atefeh, F., & Khreich, W. (2015). A survey of techniques for event detection in twitter. *Computational Intelligence*, *31*(1), 132–164.

Attia, A. M., Aziz, N., Friedman, B., & Elhusseiny, M. F. (2011). Commentary: The impact of social networking tools on political change in egypt's "revolution 2.0". *Electronic Commerce Research and Applications*, *10*(4), 369–374.

Bahrami, M., Findik, Y., Bozkaya, B., & Balcisoy, S. (2018). Twitter reveals: Using twitter analytics to predict public protests. *arXiv preprint arXiv:1805.00358*.

Becker, H., Naaman, M., & Gravano, L. (2011). Beyond trending topics: Real-world event identification on twitter.

Boulianne, S. (2015). Social media use and participation: A meta-analysis of current research. *Information, communication & society*, *18*(5), 524–538.

Bovet, A., & Makse, H. A. (2019). Influence of fake news in twitter during the 2016 us presidential election. *Nature communications*, *10*(1), 1–14.

Bruns, A., & Burgess, J. E. (2011). The use of Twitter hashtags in the formation of ad hoc publics. In *Proceedings of the 6th european consortium for political research (ecpr) general conference 2011*.

Bruns, A., & Liang, Y. E. (2012). Tools and methods for capturing Twitter data during natural disasters. *First Monday*.

Business of Apps. (2020, Nov). *Twitter revenue and usage statistics (2020).* Retrieved March 20, 2021, from https://www.businessofapps.com/data/twitter-statistics/

Carpenter, J. P., & Krutka, D. G. (2014). How and why educators use twitter: A survey of the field. *Journal of research on technology in education*, *46*(4), 414–434.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *Crisp-dm 1.0 user guide.* Retrieved March 20, 2021, from https://www.the-modeling-agency.com/crisp-dm.pdf

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321–357.

Cormode, G., & Krishnamurthy, B. (2008). Key differences between web 1.0 and web 2.0. *First Monday*.

Culnan, M. J., McHugh, P. J., & Zubillaga, J. I. (2010). How large us companies can use twitter and other social media to gain business value. *MIS Quarterly Executive*, *9*(4).

Dabiri, S., & Heaslip, K. (2019). Developing a twitter-based traffic event detection model using deep learning architectures. *Expert systems with applications*, *118*, 425–439.

Daft, R. L., & Lengel, R. H. (1986). Organizational information requirements, media richness and structural design. *Management science*, *32*(5), 554–571.

Demirhan, K. (2014). Social media effects on the gezi park movement in turkey: Politics under hashtags. In *Social media in politics* (pp. 281–314). Springer.

Diani, M. (1992). The concept of social movement. *The sociological review*, *40*(1), 1–25.

Dittrich, A., & Lucas, C. (2014). Is this twitter event a disaster?

Duan, L., & Da Xu, L. (2012). Business intelligence for enterprise systems: a survey. *IEEE Transactions on Industrial Informatics*, *8*(3), 679–687.

Edmond, C. (2013, 07). Information Manipulation, Coordination, and Regime Change*. *The Review of Economic Studies*, *80*(4), 1422-1458. Retrieved from https://doi.org/10.1093/restud/rdt020 doi: 10.1093/restud/rdt020

Edrington, C. L., & Lee, N. (2018). Tweeting a social movement: Black lives matter and its use of twitter to share information, build community, and promote action. *The Journal of Public Interest Communications*, *2*(2), 289–289.

Elsafoury, F. (2020). Teargas, water cannons and twitter: A case study on detecting protest repression events in turkey 2013. In *Text2story@ ecir* (pp. 5–13).

Enikolopov, R., Makarin, A., & Petrova, M. (2020). Social media and protest participation: Evidence from russia. *Econometrica*, *88*(4), 1479–1514.

Estellés-Arolas, E., & González-Ladrón-De-Guevara, F. (2012). Towards an integrated crowdsourcing definition. *Journal of Information science*, *38*(2), 189–200.

Ferrara, E. (2017). Disinformation and social bot operations in the run up to the 2017 french presidential election. *arXiv preprint arXiv:1707.00086*.

Forbes. (2020, Aug 17). *How much data is collected every minute of the day.* Retrieved June 30, 2021, from https://www.forbes.com/sites/nicolemartin1/2019/08/07/how-much-data-is-collected-every-minute-of-the-day/?sh=5a3f48e13d66

Gaby, S., & Caren, N. (2012). Occupy online: How cute old men and malcolm x recruited 400,000 us users to ows on facebook. *Social Movement Studies*, *11*(3-4), 367–374.

Gerbaudo, P. (2012). *Tweets and the streets: Social media and contemporary activism*. Pluto Press.

Gleason, B. (2013). # occupy wall street: Exploring informal learning about a social movement on twitter. *American Behavioral Scientist*, *57*(7), 966–982.

Goffman, E. (1949). Presentation of self in everyday life. *American Journal of Sociology*, *55*, 6–7.

Golafshani, N. (2003). Understanding reliability and validity in qualitative research. *The qualitative report*, *8*(4), 597–607.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., & Mikolov, T. (2018). Learning word vectors for 157 languages. In *Proceedings of the international conference on language resources and evaluation (lrec 2018)*.

HaCohen-Kerner, Y., Miller, D., & Yigal, Y. (2020). The influence of preprocessing on text classification using a bag-of-words representation. *PloS one*, *15*(5), e0232525.

Harlow, S. (2012). Social media and social movements: Facebook and an online guatemalan justice movement that moved offline. *New media & society*, *14*(2), 225–243.

He, W., Xu, G., Kim, Y., Dwivedi, R., Zhang, J., & Jeong, S. R. (2016). Competitive intelligence in social media twitter: iphone 6 vs. galaxy s5. *Online Information Review*.

Hootsuite. (2020a). *Digital in 2020.* Retrieved March 20, 2021, from https://p.widencdn.net/1zybur/Digital2020Global_Report_en

Hootsuite. (2020b). *Digital in 2020 - global report (q3 update).* Retrieved March 20, 2021, from https://hootsuite.com/pages/digital-2020

Hossin, M., & Sulaiman, M. (2015). A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining & Knowledge Management Process*, *5*(2), 1.

Howard, P. N., et al. (2011). The arab spring's cascading effects. *Pacific Standard*, *23*.

Hu, M., Liu, S., Wei, F., Wu, Y., Stasko, J., & Ma, K.-L. (2012). Breaking news on twitter. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 2751–2754).

Hua, T., Chen, F., Zhao, L., Lu, C.-T., & Ramakrishnan, N. (2013). Sted: semi-supervised targeted-interest event detectionin in twitter. In *Proceedings of the 19th acm sigkdd international conference on knowledge discovery and data mining* (pp. 1466–1469).

Hua, W., Huynh, D. T., Hosseini, S., Lu, J., & Zhou, X. (2012). Information extraction from microblogs: A survey. *Int. J. Software and Informatics*, *6*(4), 495–522.

Hwang, H., & Kim, K.-O. (2015). Social media as a tool for social movements: The effect of social media use and social capital on intention to participate in social movements. *International Journal of Consumer Studies*, *39*(5), 478–488.

Imran, M., Castillo, C., Diaz, F., & Vieweg, S. (2015). Processing social media messages in mass emergency: A survey. *ACM Computing Surveys (CSUR)*, *47*(4), 1–38.

Imran, M., Elbassuoni, S., Castillo, C., Diaz, F., & Meier, P. (2013). Practical extraction of disaster-relevant information from social media. In *Proceedings of the 22nd international conference on world wide web* (pp. 1021–1024).

Isa, D., & Himelboim, I. (2018). A social networks approach to online social movement: Social mediators and mediated content in# freeajstaff twitter network. *Social Media+ Society*, *4*(1), 2056305118760807.

Java, A., Song, X., Finin, T., & Tseng, B. (2007). 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* (pp. 56–65).

Kannan, S., & Gurusamy, V. (2014). Preprocessing techniques for text mining. *International Journal of Computer Science & Communication Networks*, *5*(1), 7–16.

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! the challenges and opportunities of social media. *Business horizons*, *53*(1), 59–68.

Klein, B., Laiseca, X., Casado-Mansilla, D., López-de Ipiña, D., & Nespral, A. P. (2012). Detection and extracting of emergency knowledge from twitter

streams. In *International conference on ubiquitous computing and ambient intelligence* (pp. 462–469).
Lasorsa, D. L., Lewis, S. C., & Holton, A. E. (2012). Normalizing twitter: Journalism practice in an emerging communication space. *Journalism studies*, *13*(1), 19–36.
Lee, S. (2018). The role of social media in protest participation: the case of candlelight vigils in south korea. *International Journal of Communication*, *12*, 18.
Little, A. T. (2016). Communication technology and protest. *The Journal of Politics*, *78*(1), 152–166.
Liu, I. L., Cheung, C. M., & Lee, M. K. (2010). Understanding Twitter Usage: What Drive People Continue to Tweet. In *PACIS 2010 Proceedings. Paper 92.* Retrieved from http://aisel.aisnet.org/pacis2010/92
Liu, X., Zhang, S., Wei, F., & Zhou, M. (2011). Recognizing named entities in tweets. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies* (pp. 359–367).
Lu, Y., Wang, F., & Maciejewski, R. (2014). Business intelligence from social media: A study from the vast box office challenge. *IEEE computer graphics and applications*, *34*(5), 58–69.
Marcus, A., Bernstein, M. S., Badar, O., Karger, D. R., Madden, S., & Miller, R. C. (2011). Twitinfo: aggregating and visualizing microblogs for event exploration. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 227–236).
Mathioudakis, M., & Koudas, N. (2010). Twittermonitor: trend detection over the twitter stream. In *Proceedings of the 2010 acm sigmod international conference on management of data* (pp. 1155–1158).
Meyer, D. S., & Whittier, N. (1994). Social movement spillover. *Social problems*, *41*(2), 277–298.
Mihailidis, P. (2014). The civic-social media disconnect: exploring perceptions of social media for engagement in the daily life of college students. *Information, Communication & Society*, *17*(9), 1059–1071.
Miles, J., & Gilbert, P. (2005). *A handbook of research methods for clinical and health psychology*. Oxford University Press on Demand.
Müter, L., Den Hengst, M., Van Nimwegen, C., & Veltkamp, R. C. (2021). *Anticipating the crowd using twitter: A case study.* (submitted)
Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*, *30*(1), 3–26.
Nguyen, H., Liu, W., Rivera, P., & Chen, F. (2016). Trafficwatch: real-time traffic incident detection and monitoring using social media. In *Pacific-asia conference on knowledge discovery and data mining* (pp. 540–551).
NOS. (2020). *Racismeprotest rotterdam afgebroken, twee arrestaties bij opstootjes.* Retrieved June 17, 2021, from https://nos.nl/artikel/2336102-racismeprotest-rotterdam-afgebroken-twee-arrestaties-bij-opstootjes
NPO Start. (2020, June 1). *Op1 - 1 juni 2020 gemist?* Retrieved July 2, 2021, from https://www.npostart.nl/op1-1-juni-2020/01-06-2020/POW_04675807
OECD. (2007). *Participative web and user-created content: Web 2.0 wikis and social networking*. Organization for Economic Cooperation and Development.
O'reilly, T. (2009). *What is web 2.0*. " O'Reilly Media, Inc.".
Ortiz-Ospina, E. (2019, Sep 18). *The rise of social media.* Our World in Data. Retrieved March 20, 2021, from https://ourworldindata.org/rise-of-social-media
Oxford Dictionary. (n.d.-a). *Definition of demonstration noun.* Retrieved March 20, 2021, from https://www.oxfordlearnersdictionaries.com/definition/

english/demonstration

Oxford Dictionary. (n.d.-b). *Definition of spillover noun.* Retrieved March 20, 2021, from https://www.oxfordlearnersdictionaries.com/definition/english/spillover

Piatetsky, G. (2014, Oct). *CRISP-DM, still the top methodology for analytics, data mining, or data science projects.* KDnuggets. Retrieved March 20, 2021, from https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html

Qian, Z. S., et al. (2016). *Real-time incident detection using social media data.* (Tech. Rep.). Pennsylvania. Dept. of Transportation.

Ranney, K. R. (2014). *Social media use and collective identity within the occupy movement* (Unpublished doctoral dissertation). [Honolulu]:[University of Hawaii at Manoa],[December 2014].

Rijksoverheid. (2020, Nov). *Demonstraties in coronatijd.* Our World in Data. Retrieved March 20, 2021, from https://www.rijksoverheid.nl/binaries/rijksoverheid/documenten/rapporten/2020/11/05/tk-bijlage-3a-rapport-demonstraties-in-coronatijd/tk-bijlage-3a-rapport-demonstraties-in-coronatijd.pdf

Robinson, B., Power, R., & Cameron, M. (2013). A sensitive twitter earthquake detector. In *Proceedings of the 22nd international conference on world wide web* (pp. 999–1002).

Russom, P., et al. (2011). Big data analytics. *TDWI best practices report, fourth quarter*, *19*(4), 1–34.

Saeed, Z., Abbasi, R. A., Maqbool, O., Sadaf, A., Razzak, I., Daud, A., ... Xu, G. (2019). What's happening around the world? a survey and framework on event detection techniques on twitter. *Journal of Grid Computing*, *17*(2), 279–312.

Salas, A., Georgakis, P., & Petalas, Y. (2017). Incident detection using data from social media. In *2017 ieee 20th international conference on intelligent transportation systems (itsc)* (pp. 751–755).

Sandoval-Almazan, R., & Gil-Garcia, J. R. (2014). Towards cyberactivism 2.0? understanding the use of social media and other information technologies for political activism and social movements. *Government information quarterly*, *31*(3), 365–378.

Settles, B. (2009). *Active learning literature survey* (Tech. Rep.). University of Wisconsin-Madison Department of Computer Sciences.

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, *19*(1), 22–36.

Singel, R. (2010, Aug 12). *Mastercard.com taken down by pro-wikileaks forces.* Wired. Retrieved March 20, 2021, from https://www.wired.com/2010/12/mastercard-anonymous/

Snow, R., O'connor, B., Jurafsky, D., & Ng, A. Y. (2008). Cheap and fast–but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 conference on empirical methods in natural language processing* (pp. 254–263).

Suh, C. S., Vasi, I. B., & Chang, P. Y. (2017). How social media matter: Repression and the diffusion of the occupy wall street movement. *Social science research*, *65*, 282–293.

Terpstra, T., Stronkman, R., de Vries, A., & Paradies, G. L. (2012). Towards a realtime twitter analysis during crises for operational crisis management. In *Iscram.*

The Guardian. (2010, Dec 08). *Mastercard site partially frozen by hackers in wikileaks 'revenge'*. Retrieved March 20, 2021, from https://www.theguardian.com/media/2010/dec/08/mastercard-hackers-wikileaks-revenge

Tremayne, M. (2014). Anatomy of protest in the digital era: A network analysis of twitter and occupy wall street. *Social Movement Studies*, *13*(1), 110–126.

Trump, D. (n.d.-a). *Tweet by Donald Trump.* Retrieved December 17, 2020, from https://twitter.com/realDonaldTrump/status/1336122354003537921

Trump, D. (n.d.-b). *Twitter profile of Donald Trump.* Retrieved December 17, 2020, from https://twitter.com/realDonaldTrump

Uysal, A. K., & Gunal, S. (2014). The impact of preprocessing on text classification. *Information Processing & Management*, *50*(1), 104–112.

Van der Plaat, J. (2021). *Smid - Social Media Incident Dashboard.* (unpublished thesis)

van der Velden, P., Nooy, K., & Boin, A. (2020). *Een verkeerde afslag, een analyse van de 1 juni demonstratie in amsterdam.* Retrieved June 30, 2021, from https://assets.amsterdam.nl/publish/pages/955579/rapport_1_juni_demonstratie.pdf

Van Laer, J., & Van Aelst, P. (2010). Internet and social movement action repertoires: Opportunities and limitations. *Information, Communication & Society*, *13*(8), 1146–1171.

Vijayarani, S., Ilamathi, M. J., & Nithya, M. (2015). Preprocessing techniques for text mining-an overview. *International Journal of Computer Science & Communication Networks*, *5*(1), 7–16.

Vos, D. (2021). *Linking tweets: Mapping locations to tweets using named entity recognition and textual similarity methods.* (unpublished thesis)

Wang, L., & Gan, J. Q. (2017). Prediction of the 2017 french election based on twitter data analysis. In *2017 9th computer science and electronic engineering (ceec)* (pp. 89–93).

Whiting, A., & Williams, D. (2013). Why people use social media: a uses and gratifications approach. *Qualitative Market Research: An International Journal*.

Wieringa, R. J. (2014). *Design science methodology for information systems and software engineering*. Springer.

Witten, I. H. (2004). *Text mining*. Retrieved March 20, 2021, from https://www.cs.waikato.ac.nz/~ihw/papers/04-IHW-Textmining.pdf

Wouters, J. (2021). *Early warning of incidents during protest demonstrations using twitter.* Retrieved from https://github.com/jorenman/early-warning-incidents-using-twitter

Xu, S., Li, S., Wen, R., & Huang, W. (2019). Traffic event detection using twitter data based on association rules. *ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences*, *4*.

Yang, L. C., Tan, I. K., Selvaretnam, B., Howg, E. K., & Kar, L. H. (2019). TEXT: Traffic Entity eXtraction from Twitter. In *Proceedings of the 2019 5th international conference on computing and data engineering* (pp. 53–59).

Appendix A

Timeline of events on 1st June 2020

This appendix describes the timeline of events related to the protest demonstration on 1st June 2020 in Amsterdam. The timeline was constructed based on the work of Müter, Den Hengst, Van Nimwegen, and Veltkamp (2021) and a research report commissioned by the municipality of Amsterdam (van der Velden, Nooy, & Boin, 2020).

Between 15:00-16:00, around 1-2 hours before the start of the demonstration, the protest organizers make their first preparations at Dam Square in Amsterdam. These preparations include painting crosses on the ground and preparing signs with calls to keep 1.5-meter distance from each other.

Just before the start of the demonstration, between 16:45 and 17:00, the number of people joining the demonstration increases fast. According to present police officers, they have never seen a crowd increase so fast in such a short time. At 17:00, around 1000 protesters are on Dam Square and are not keeping 1.5-meter distance from each other. At 17:49, approximately 50 minutes after the start of the demonstration, the number of protesters has increased to 5000.

One hour after the start of the demonstration, at 18:00, the mayor of Amsterdam is heading to Dam Square. There, she is interviewed and declares that people are responsible for their own safety. At this moment, there is not enough police force to disperse the demonstration.

Two hours after the start of the demonstration (19:00), the Telegraaf posts articles regarding the lack of compliance with the COVID-19 rules at Dam Square.

At 20:00, the Dutch police estimate the number of protesters between 10,000-14,000 at its peak.

At 22:15, a Dutch TV program starts with Femke Halsema (the mayor of Amsterdam), in which she explains the decisions made during the protest demonstration in Amsterdam (NPO Start, 2020).

Appendix B

Labeling process of users

This appendix describes the labels that were used in labeling Twitter users and the process of automatic and internal labeling.

B.1 Labels of users

Twitter users were labeled according to the following labels:

1. *Mass media*, this includes media organizations and programmes (including TV programmes and radio stations).
2. *Media people*, this includes people working at media organizations as presenters, radio DJs, journalists, columnists, editors and chiefs.
3. *Politician*, this includes politicians with various functions, such as members of political parties, (former) members of parliament (including Dutch and European parliament) and secretaries of state.
4. *Municipality*, this includes both Dutch municipalities and Dutch municipalities' local councils.
5. *Mayor*, this includes mayors of Dutch municipalities.
6. *Parliament*, this includes the parliament itself.
7. *Writer*, this includes the writer of books.
8. *Soccer clubs*, this includes Dutch soccer clubs, such as PSV, Ajax and Feyenoord.
9. *Government organization*, which refers to government organizations, such as the public prosecutor's office, the General Intelligence and Security Service and the tax office. This does not include the Dutch police.
10. *Part of government organizations*, refers to employees working at government organizations.
11. *Police*, this includes Twitter profiles related to the Dutch police.
12. *Political organization*, refers to organizations with a political goal, such as Greenpeace and Amnesty International.
13. *Province*, refers to Twitter profiles of Dutch provinces.
14. *Political party*, refers to Twitter profiles of Dutch political parties, such as the VVD.

15. *Political activist*, refers to Twitter profiles of activists with a political nature.
16. *Social networks*, refers to Twitter profiles of social networks, such as Facebook, LinkedIn and Twitter.
17. *Virologist*, refers to Twitter profiles of virologists.
18. *Cinema*, refers to Twitter profiles of Dutch cinemas, such as Pathé.
19. *Musician*, refers to Twitter profiles of Dutch musicians.
20. *Comedian*, refers to Twitter profiles of Dutch comedians.
21. *Actress*, refers to Twitter profiles of Dutch actresses.

The aforementioned labels were not determined beforehand and were decided iteratively during the labeling process. Also, if a user profile was related to multiple labels, one primary label was chosen.

B.2 Partly automated labeling

A slice of the users was automatically labeled based on the description of the Twitter profile or the screen name of the Twitter profile. From the 1475 labeled users, 81 users were automatically labeled by the process described below.

If the description of the Twitter profile contained "kamerlid", "lid tweede kamer", "member or european parliament", "member of the european parliament", "europarlementariër" or "lid europees parliament", the user was labeled as Politician (11 cases). Moreover, if the description of the Twitter profile contained "journalist", "nieuwschef", "verslaggever", "redacteur" or "columnist", the user was labeled as Media people (67 cases). Lastly, if the screen name of the Twitter profile contained "politie", the user was labeled as Police (3 cases).

B.3 User labels statistics

In total, 1475 users were labeled according to the labels described above. The resulted frequencies per user type are presented in Table B.1.

Type of users	**Frequency**
No type	1126
Media people	122
Politician	95
Mass media	53
Political party	9
Political organization	8
Government organization	8
Police	7
Musician	7
Soccer club	7
Writer	6
Political activist	6
Comedian	6
Municipality	3
Actress	3
Mayor	2
Part of government organization	2
Social network	2
Virologist	1
Province	1
Parliament	1

TABLE B.1: Frequencies of user labels according to user type.

Appendix C

Labeling statistics of tweets

In this appendix, the labeling statistics of the tweet labeling task are described.

The tweets of the unique tweets dataset were labeled by six labelers. Each tweet was labeled by multiple labelers using an in-house label tool (called Tweeti). This tool kept track of the labeling process and how many each tweet was labeled by whom. Tweeti automatically assigned a labeled tweet to the user that labeled the tweet for the last time. One might argue that this could give biased results, but this was not the case, because Tweeti labeling statistics were examined on a weekly basis. Figure C.1 shows the number of tweets per user. Figure C.2 shows the label rates per user in terms of what classes were provided to labeled tweets.

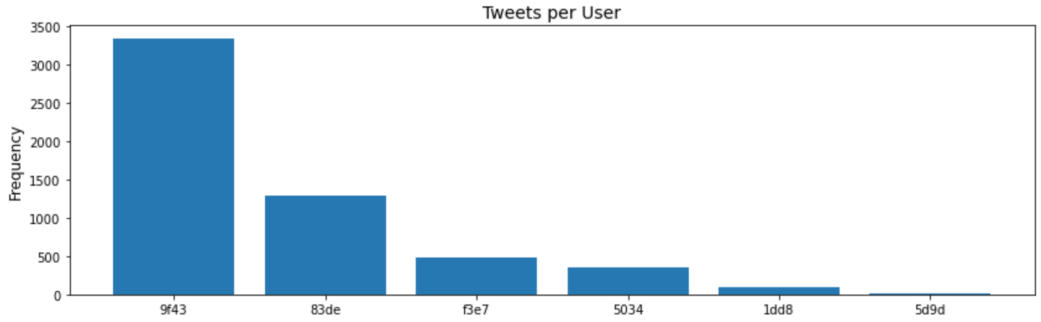

FIGURE C.1: Number of labeled tweets per labeler.

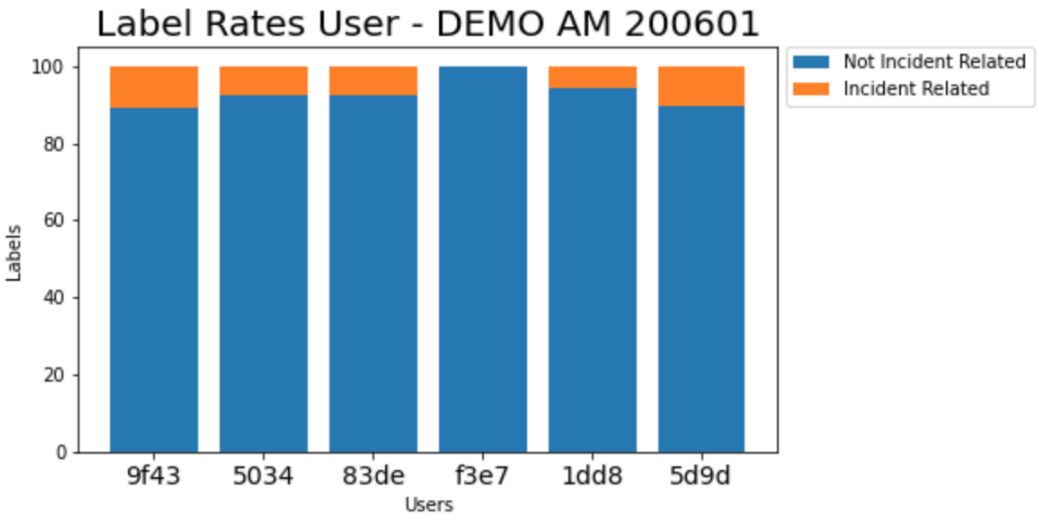

FIGURE C.2: Label rates per labeler in terms of labeled classes.

Appendix D

Interview questions

This appendix describes the questions that were used as a basis for the interviews with police officers of the Open-Source Intelligence Team of the Dutch national police force.

As described in Section 5.7, in the first phase of the interview, general questions were asked to the participants to get a general understanding of the way of working, the current process of social media data analysis and to identify possible problems when performing social media data analysis. The questions that were used as a basis for this phase are:

1. What is your position within the police?
2. In which department do you work? And what is the role of that department within the police?
3. What is your role within the department?
4. Is social media currently being analyzed by the police?
5. If yes, what is analyzed? And how is this performed?
6. What problems do you run into while analyzing social media?
7. What is the time span in which the analysis takes place?
8. How is the information resulting from the analysis used by the police?
9. When do you see something as an incident during a demonstration?
10. If we look back to the COVID-19 times. Is not adhering to the COVID-19 rules also an incident for you?
11. Do you currently detect incidents based on social media? If yes, do you do this in real-time?

Following, the incident-related tweets of the unique tweets datasets were presented one by one and the following questions were asked:

1. Do you want to receive a warning after observing this tweet?
2. If yes, why do you want the notification after seeing the tweet? Why not sooner/later?
3. Is that specific for this tweet? Of do you always want a notification after ... incident-related tweet(s)?

4. If you get a warning of an incident, what information would you like to receive?
5. Why would you want that information?
6. What is the time span in which you would like to receive that information?

Eventually, one additional question was asked:

1. Would you benefit from a system that automatically detects incidents based on tweets?

Appendix E

Interview Consent Form

Early warning of incidents during protest demonstrations

Consent form for participation in the Study.

Please complete the form below by ticking the relevant boxes and signing on the line below. A copy of the completed form will be given to you for your own record.

- ☐ I confirm that the research project ***"Early warning of incidents during protest demonstrations"*** has been explained to me. I have had the opportunity to ask questions about the project and have had these answered satisfactorily.

- ☐ I consent to the material I contribute being used to generate insights for the research project "***Early warning of incidents during protest demonstrations".***

- ☐ I am aware that the researcher will take a recording of the session. I understand that I can request to stop these recordings. I understand that I can ask for the recording to be deleted.

- ☐ I understand that my participation in this research is voluntary and that I may withdraw from the study at any time.

- ☐ I consent to allow the fully anonymised data to be used for future publications and other scholarly means of disseminating the findings from the research project.

- ☐ I confirm that I am 18 years of age or over.

- ☐ I understand that the information/data acquired will be securely stored by researchers, but that appropriately anonymised data may in future be made available to others for research purposes only.

- ☐ I understand that I can request any of the data collected from/by me to be deleted.

- ☐ I agree to take part in the above study on ***"Early warning of incidents during protest demonstrations"***.

Name of participant	Date	Signature

Appendix F

Pseudo code of designed system

Algorithm 1: Pseudo code of the designed system

```
mean = 0
mean_dev = 0
windows = DataFrame()
events = DataFrame()
n_windows = 5
length_window = 5
α = 0.125
queue = Queue()

Def on_new_tweet tweet:
    addTweettoQueue(tweet, queue)

Def process_tweet_from_queue tweet, queue:
    cleanTweet(tweet)

Def cleanTweet tweet:
    cleaned_tweet = clean(tweet)
    AddtoTimewindow(cleaned_tweet)

Def AddtoTimewindow cleaned_tweet:
    if len(windows) == 0 then
        windows.append(start_window = cleaned_tweet.created_at)
        window = getCurrentWindow()
    else
        window = getCurrentWindow()
    end
    n_tweets_in_window = getTweetsInTimeWindow(window)
    start_window = getStartofWindow(window)
    while True do
        if cleaned_tweet in start_window then
            addTweettoWindow(window)
            break
        else
            detectEvent(start_window)
            mean, mean_dev = update(mean, mean_dev,
              n_tweets_in_window)
            createNewWindow(start_window = start_window +
              length_window)
        end
    end
```

```
39 Def update mean, mean_dev, n_tweets_in_window:
40     if len(windows) == n_windows then
41         newmean = mean(n_tweets_last_n_windows)
42         newmean_dev = var(n_tweets_last_n_windows)
43     else
44         diff = | mean-n_tweets_in_window |
45         newmean = α*n_tweets_in_window + (1-α) * mean
46         newmeandev = α*diff + (1-α) * mean_dev
47     return newmean, newmeandev
48
49 Def detectEvent start_window:
50     if len(windows) > n_windows then
51         window = getCurrentWindow(start_window)
52         previous_window = getCurrentWindow(start_window)-1
53         n_tweets_current_window = getTweetsInTimeWindow(window)
54         n_tweets_previous_window =
             getTweetsInTimeWindow(previous_window)
55         if n_tweets_current_window > (mean + 2*mean_dev) and
             n_tweets_current_window > n_tweets_previous_window then
56             events.append(window)
57             detectIncident(start_window)
58
59 Def detectIncident start_window:
60     window = getCurrentWindow(start_window)
61     tweets_current_window = getTweetsCurrentWindow(window)
62     for tweet in tweets_current_window do
63         pred = predict(tweet)
64         if pred == Incident_Related then
65             setCurrentEventtoIncident(start_window)
66             alert("Incident detected")
```

Appendix G

Compare datasets

This appendix describes a comparison of the unique tweets dataset and two additional more balanced datasets.

Each of the datasets is divided into a training set (80%) and a test set (20%) by performing a randomized stratified split. The training set is used to train a simple Naive Bayes model upon (meaning that no hyperparameter optimization is performed). Consequently, the test set is used to evaluate the performance of the Naive Bayes classifier. The goal of the Naive Bayes classifier is to predict a tweet into two classes: [Incident-Related] and [Not Incident-Related]. Just as with the model evaluation task (described in 5.6.3), there is an emphasis on the F-Measure on the Incident-Related class.

Table G.1 presents the results of each performance metric for each dataset. For each metric, the highest scores have been printed in bold. Results show that the second balanced dataset provides the highest score on the F-Measure of the Incident-Related class. Therefore, this dataset will be used to train the machine learning models.

Dataset	F1 (Not-Related)	F1 (Related)	Precision (Not-Related)	Precision (Related)	Recall (Not-Related)	Recall (Related)	AUC	Accuracy
Unique tweets dataset	**0.955**	0.040	**0.915**	0.667	**0.999**	0.021	0.718	**0.914**
Additional dataset 1	0.907	0.501	0.844	0.843	0.981	0.357	0.911	0.844
Additional dataset 2	0.909	**0.675**	0.850	**0.896**	0.976	**0.542**	**0.933**	0.857

TABLE G.1: Performance metrics of a simple Naive Bayes model on each of the datasets. For each metric, the highest score are printed in bold. F1 relates to the the F-Measure. Related refers to the [Incident-Related] class, Not-Related refers to the [Not Incident-Related] class.

Appendix H

Most important feature words

This appendix describes the most important feature words of the Naive Bayes (model 1), Logistic Regression (model 1) and Gradient Boosted Decision Trees (model 1) algorithms. For all three algorithms hold that the values were calculated using the built-in functions of the *scikit-learn* Python package.

Table H.1 shows the most important feature words of model 1 of the Naive Bayes algorithm, in terms of the highest log probability on the Incident-Related class. From this table, it can be found that the feature words "dam", "ondernemers" and "verbijsterd" are the most important.

Feature word	Log Prob Not Incident-Related	Log Prob Incident-Related
dam	-5.305169984448026	-3.9022531479818374
ondernemers	-6.780558407277384	-4.345297956917787
verbijsterd	-7.247713951730142	-4.364634109447478
situatie	-6.903871446051294	-4.368145171426241
leg	-7.007199831928583	-4.36875042768475
politici	-7.148018673008963	-4.3778417830364065
burgemeester	-5.510693065818758	-4.409181429305725
demonstratie	-4.256838376455116	-4.419887135514203
telegraaf	-7.418050695845781	-4.47839953444457
via	-7.174556007401479	-4.7263020085048995

TABLE H.1: Most important feature words corresponding with the log probability of the Not Incident-Related class and Incident-Related class of model 1 of the Naive Bayes algorithm.

Table H.2 shows the most important feature words of model 1 of the Logistic Regression algorithm, in terms of the highest coefficient values. From this table, it can be noticed that the feature words "at5", "duizenden" and "telegraaf" are the most important.

Feature word	Coefficient
at5	8.594563661083185
duizenden	8.364111918288092
telegraaf	7.731129771243821
landinwaarts	7.491832410182088
dam	6.025687961778924
drukke	5.965523474879652
meeste	5.751295039890563
hielden	5.2777571713684965
meterregel	5.245201644926233
protest	5.067862288456788

TABLE H.2: Most important feature words corresponding with the coefficient values of model 1 of the Logistic Regression algorithm.

Table H.3 shows the most important feature words of model 1 of the Gradient Boosted Decision Trees algorithms, in terms of the highest feature importance values. From this table, it can be found that "telegraaf", "dam" and "at5" are the most important.

Feature word	Feature importance
telegraaf	0.09481483879633137
dam	0.050607337288610235
at5	0.04595351188554006
demonstratie	0.04191660942247982
duizenden	0.02542954750005819
afstand	0.012226108442003426
burgemeester	0.009094840752145948
gegooid	0.008810079707801217
uitgedeeld	0.008296253870716528
aanwezig	0.008114891104822686

TABLE H.3: Most important feature words corresponding with the feature importance values of model 1 of the Gradient Boosted Decision Trees algorithm.

In Table H.4, the most important feature words of all algorithms are presented together with their occurrence in the ten most important feature words of the corresponding algorithms (Naive Bayes, Logistic Regression and Gradient Boosted Decision Trees).

Feature word	Naive Bayes	Logistic Regression	Gradient Boosted Decision Trees
dam	✓	✓	✓
ondernemers	✓		
verbijsterd	✓		
situatie	✓		
leg	✓		
politici	✓		
burgemeester	✓		✓
demonstratie	✓		✓
telegraaf	✓	✓	✓
via	✓		
at5		✓	✓
duizenden		✓	✓
landinwaarts		✓	
drukke		✓	
meeste		✓	
hielden		✓	
meterregel		✓	
protest		✓	
afstand			✓
gegooid			✓
uitgedeeld			✓
aanwezig			✓

TABLE H.4: Most important feature words of all algorithms with occurrence in the ten most important feature words of the corresponding algorithms.