Detecting Rumors in Twitter for Humanitarian Activities

Dimitrios-Marios Vaporidis

Detecting Rumours in Twitter for Humanitarian Activities

Master thesis submitted to Delft University of Technology in partial fulfilment of the requirements for the degree of

MASTER OF SCIENCE

in **Engineering & Policy Analysis**

by **Dimitrios-Marios Vaporidis**

Student Number: 4623827

To be defended in public on January 24th 2019

Graduation Committee

Chairperson: Dr., M.E., Warnier, Systems Engineering and Simulation
First Supervisor: Dr., Y., Huang, Systems Engineering and Simulation
Second Supervisor: S., Cunningham, Policy Analysis
Advisor: A., Ebrahimi Fard, Policy Analysis

Acknowledgements

I am grateful to all of those with whom I have had the pleasure to work during this Master Thesis project. Each of the members of my Thesis Committee has provided me with extensive feedback and taught me a great deal about scientific research. I would especially like to thank my First Supervisor, Dr. Y. Huang. Her support and guiding in this project helped me more than I could give her credit.

Nobody has been more important to me in the pursuit of this project than my parents. I would like to thank my parents, whose love and support are with me in whatever I pursue. They are the ultimate role models. Additionally, I would to thank my friends, Panos, Thekla, Eirini, Antria, Aggelos, Makis Kostas, Alkis, for their support over the last 2,5 years of this Master's program.

Most importantly, I wish to thank my loving and supportive girlfriend, Rania, for all the years she is by my side and her help in understanding social media from a professional perspective.

Dimitrios-Marios Vaporidis

Table of Contents

List of Figures

List of Tables

Executive Summary

Social media information has increased over the last years. This new volume of information contains many event related facts. Due to the sheer amount of social media information, it is difficult to extract these facts with manual processes. Therefore, in order to extract these facts ITs are utilised in most sectors such as marketing and market analysis.

The driving force of this Master Thesis project is how can current IT methods and techniques be incorporated in the detection and verification of social media information in humanitarian crisis management. Social media information can potentially contribute significantly in humanitarian aid. Despite its potential contribution, social media information is used little to none in humanitarian crisis management and the detection and verification processes incorporate little automation. This constitutes the problem that is tackled by this Master Thesis project.

As this topic is broad, the research of this Master Thesis project focuses on the detection of humanitarian text-based rumours in Twitter. Twitter was chosen as a data source due to the news related nature of the posts in this social medium and the volume of tweets posted during a humanitarian crisis. These tweets can provide with critical information regarding aid operations, boost the moral and raise the awareness of the affected population.

The technology that was chosen to explore the solution space of this problem was Supervised Machine Learning using the Information Systems Design approach. The design of the artifact produced is created through iterative steps, i.e. build and evaluate, in order to align the objectives of the stakeholders of the problem with the capabilities of the technology explored. The model takes as input tweets and classifies them in rumours or not and relevant to humanitarian operations or not.

The model consists of 3 modules, the Preprocessing Module, the Rumour Detector and the Humanitarian Relevancy Classifier. The model is called Humanitarian Rumour Detector.

The performance of the model is satisfactory in terms of accuracy rates as well as operational time requirements. The Rumour Detector scored 75.8% in Recall which outperformed. The Humanitarian Relevancy Classifier scored 96.6%. Regarding their operational times, both classification modules are able to classify great volumes of tweets in under a second. The findings of this project indicate that Supervised Machine Learning can be utilised in humanitarian cri-

sis management to either detect text-based humanitarian rumours or filter the Twitter input and significantly reduce the volume of tweets that have to be manually reviewed.

The performance of the the model indicates that Supervised Machine Learning can be utilised as a technology to handle the volume of information produced in social media. This will lead to the incorporation of social media information in the crisis management decision making. Moreover, as the modules that were developed in this Master Thesis project can be used separately, humanitarian agencies can use parts of the model according to their preferences, which will result in a smoother transitioning period. Additionally, this could potentially boost the utilisation of computational approaches in the humanitarian sector as humanitarian organisations will gradually adjust to IT.

The analysis of the model provides critical insight on the inner workings of the model. The proper structuring of the dataset that trains the Humanitarian Relevancy Classifier directly affects the performance of the model in separating the tweets relevant to humanitarian activities. The dataset should be an anthology of humanitarian disasters in proportional ratios. The analysis of the Rumour Detector provides insight on what aspects of a tweet constitute whether a tweet is a rumour or not. The most important dimensions of detecting a rumour are user engagement and activity, text structure and propagation through retweeting.

The insight that was acquired by the analysis of the Rumour Detector can result in data driven decision making during humanitarian crises. This can make the manual review of tweets easier, faster and more accurate.

The findings of this project can be used in order to design more complex and autonomous Machine Learning models. This could be achieved by either a more sophisticated design of a model or by the combination of other technologies such as crowdsourcing or online platforms. In the case of online platforms, this could result in the affected population being directly and immediately informed about the on-going situation as crisis-mapping can visualise the available information. In the case of crowdsourcing, the output of the model could be evaluated directly by the affected population.

The most important implications of the incorporation of social media information and Machine Learning in the humanitarian crisis information retrieval process and decision making process are the need for a restructure of the hierarchy of humanitarian agencies and their resources allocation. IT departments would have to be increased in employee numbers as they are understaffed at the moment and in general IT would have to be a more significant department in comparison to its current status. Additionally, the allocation of resources would be quite different as a bigger portion of the budget would be spent in IT but much less volunteers would be required in the information retrieval process. One might say that this way humanitarian agencies will be introduced to the 21st century and the power of information.

The research and the findings of this Master Thesis project are an exploration of the Supervised Machine Learning as a solution to the automatic detection of humanitarian text-based rumours in Twitter. The findings of this Master Thesis

project can act as an introduction to the utilisation of automatic processes in the humanitarian crisis information retrieval. The aspiration of this project is that this research will lead in the development of an autonomous universal mechanism that could act as a real-time humanitarian information detector.

Chapter 1

Introduction

The first Chapter of this Master Thesis project explored the current Information Technologies that are used in the detection and verification of social media information for humanitarian activities. The objective of this Chapter was to present what constitutes a humanitarian crisis, how can social media contribute in humanitarian crises, what are the methods that are currently used to detect and verify social media information and what are the challenges in the integration of social media information in the humanitarian crisis management.

To achieve this objective, Chapter 1 focused on the potential contributions of social media information during a humanitarian crisis. Additionally, it was investigated why this information are still not incorporated in the decision making process of humanitarian crisis management and what are the available methods to detect and verify information in order to formulate the problem that was tackled in this Master Thesis project.

1.1 Social Media Information in Humanitarian Crisis Management

1.1.1 Social Media Information

Social media are computer-mediated technologies or platforms that let the users create information and share them amongst their network. This information can be ideas, opinions, facts, news or any form of expression.

The usage of social media has been intensified the last years. In 2010, the number of users was 0.97 billions, which has more than doubled till today reaching the number of 2.62 billion users. It is projected that by 2021 more than 3 billion users will be in social media platforms (Statista 2018b).

Another valuable point of the research of Statista (2018) is the correlation between the progress of mobile phone technology and the increase in the usage of social media platforms. This progress has expanded the capabilities of mobile social networks to create information, such as location-based services or

suggestion to the user given his or her preferences.

There is a great number of social media platforms offering a great variety of services. These services can be professional networking, for example LinkedIn, social networking, for example Facebook or Instagram, opinion and event sharing, for example Twitter and many more. Social media platforms offer possibilities for the creation and dissemination of user-generated content (UGC), like photos, videos, GPS location or news reporting.

The volume of social media information is not affected only by the number of social media users but also by the interactions between the users. Given the fact that the number of social media users has increased considerably over the last decade, the volume of social media information is bound to have increased as well. This translates to an abundant pool of information that can be utilised in various sectors. The analysis and utilisation of this information has already been initialised in several sectors such as marketing or market analysis (Kaplan 2012).

The propagation of social media information takes places in such a fast pace that conventional communication technologies' usage has started to decrease considerably (Kaplan and Haenlein 2010). Moreover, as indicated by a CNN research in 2010, many people are being informed about today's news from websites, web-magazines or micro-blog platforms such as Twitter (Gross 2010). Taking into account that the expansion and spreading of social media has only but increased since then, it is safe to assume that social media information play an even more crucial role in today's society.

As a result, the propagation of information through social media has introduced a big transition in the structure of information dissemination from a top-down to a bottom-up information flow. In the top-down information flow, a hierarchical structure filters the information and shares the one that they deem to fit the business model of the platform, newspaper or television channel whereas in bottom-up information flow, the sharing and the dissemination of information is result of the interaction between the interaction of the users of the platform (Fraser et al. 2006).

This project focused on Twitter as its main data source. The reasons for this choice are presented below:

- Twitter's Nature
- Volume of Information

In information related humanitarian research, most papers use as a data source Twitter (Granell and Ostermann 2016). Twitter is an American social media platform. Its scope is online news and social networking. Users interact with each other. The statements that are posted in a user's profile are called *Tweets*. The combination of news propagation and human interaction makes Twitter very suitable to be used as a data source for humanitarian research.

Furthermore, Twitter is one of the most popular social media at the moment, as it is ranked 11th with 336 million active users (Statista 2018a). This engagement translates to 350000 tweets per minute on average (Stats 2018).

The complete and correct utilisation of the available tweets could result in an unprecedented volume of information for humanitarian crisis management.

1.1.2 The potential contribution of social media information to humanitarian operations

Since 1950, the frequency of natural disasters, the damage they cause and the number of people affected by them have increased exponentially (Özdamar and Ertem 2015). This has led to the increase of humanitarian crises and the need for efficient crisis management. A humanitarian crisis is a single or a sequence of events jeopardising the well-being and security of the population of a community (Fottrell and Byass 2009).

Recent humanitarian crises, such as the Egyptian revolution of 2011 and the earthquake in Haiti in 2010, have involved the contribution of social media platforms. For example Twitter, Facebook and other online platforms have facilitated the dissemination of information much more efficiently than conventional communication channels (Norheim-Hagtun and Meier 2010; Hermida, Lewis, and Zamith 2014). Conrado et al. (2016) have shown in their research that citizens and authorities can make quicker and safer decisions based on real-time information available on social media.

Social media information can contribute greatly to a quick and accurate humanitarian aid (Panagiotopoulos et al. 2016). Not only does the operational efficiency of the crisis management increases but also the moral of the citizens (Haworth 2016). The involvement of locals as volunteers in aid operations and more specifically in the dispersion of information results the following (Haworth 2016):

- Increased reach of communications between the affected population and humanitarian agencies inside and outside the affected area
- Individual empowerment for the affected populations
- Increased spatial awareness for the people conducting relief operations

1.2 Social Media Rumouring during Humanitarian Crises

Online rumouring behaviour during times of crisis is perceived as a collective sense-making process, which occurs during uncertain situations or events (Caplow 1947; DiFonzo and Bordia 2007). Under this perception, rumouring can be regarded as a mechanism to cope with the anxiety imposed to the affected individuals by the crisis events (Shibutani 1966).

Twitter has a plethora of rumours that are spread intentionally or unintentionally (X. Liu, Li, et al. 2016). Thus, it is important to be able to classify information to rumour and non-rumour. The volume of information currently

available on the internet and especially in social media platforms is unprecedented. In order to acquire information regarding the event that concerns the relief operations, irrelevant information should be filtered out.

1.2.1 The Challenges of integrating Rumouring in Crisis Management

The utilisation of social media information in disaster management creates many challenges (Granell and Ostermann 2016). These challenges hinder the integration of this massive pool of data in relief operations. The obstacles that have been studied most and categorised as the most critical are the following (Haworth 2016; Conrado et al. 2016; Anson et al. 2017):

- Quality of information
- Trust Issues

Information Quality Assessment in Crisis Management

The quality of information is a major challenge that affects both public and humanitarian agencies (Haworth 2016). The vast amount of information that circulates in social media includes relevant and useful information but also inaccurate information (Anson et al. 2017). This challenge acts as an obstacle for the quality assessment of social media information and for the detection of relevant information (Conrado et al. 2016).

Relief operations in a natural disaster requires facts regarding the event and the current situation (Yuan and R. Liu 2018; Popoola and Krasnoshtan 2013). The dissemination of social media information could contribute significantly in the efficient and timely response (Takahashi, Tandoc, and Carmichael 2015). The internet and the progress of Information Technology (IT) have unlocked a vast amount of information (Watson and Rodrigues 2017; Sump-Crethar 2012). As the size of information networks has been growing, so does the amount of information available (Schifferes et al. 2014). This information can be utilised to boost the accuracy of relief operations (Carley et al. 2016).

In order for this information to be utilised, it has to be of good quality. Zubiaga et al. (2018) developed a framework to evaluate the quality of rumours. The stages of this framework are presented below:

1. Rumour Detection: A rumour classification system that identifies whether a shared information is a rumour or not (Zubiaga et al. 2018).
2. Rumour Tracking: A tracking system that collects posts and reaction to the rumour which is investigated.
3. Stance Classification: A system that analyses the posts collected in Rumour Tracking to determine the stance of the reader of the statement is supporting, denying, querying or commenting on the content of the statement (Zubiaga et al. 2018).

4. Veracity Classification: This stage tries to ascertain the truth value of the post.

In this framework, an information in the beginning of the process is considered a rumour. Rumour is defined as an unverified information statement circulating the news at a time-period close to the event/information entailed in the statement (DiFonzo and Bordia 2007). Therefore, in this project a tweet is considered as rumour when it fulfils the following criteria:

- Being an information statement
- Being unverified
- Being related to an event
- Circulating social media at the time of the event it relates to

This Master Thesis project focused on the first stage of the information quality assessment which is rumour detection.

Twitter has a plethora of rumours that are spread intentionally or unintentionally (X. Liu, Li, et al. 2016). Thus, it is crucial to be able to classify information to rumour and non-rumour. The volume of information currently available on the internet and especially in social media platforms is unprecedented. In order to acquire information regarding the event that concerns the relief operations, irrelevant information should be filtered out.

In order for the rumour detection to be beneficial to humanitarian crisis management, the rumours detected have to be relevant to humanitarian context. An information is relevant when it is useful or legitimate to the subject of investigation (Zimmer et al. 2010).

Social Media Utilisation Mistrust

In order for a technology to be integrated in the activities of an organisation or company, it has to be trusted by the people that is going to be used. The majority of the agencies in humanitarian organisations are doubtful about the incorporation of social media information in their decision-making (Wendling, Radisch, and Jacobzone 2013).

Tapia et al. (2011) have analysed the use of social media information by Humanitarian Non-Governmental Organisations (NGOs). The authors identified the bottlenecks of integrating social media as a source of information regarding the operational activities of these agencies. Nevertheless, several of the statements presented in their work indicate that the the roots of the problem are the following:

- Understaffed IT Departments
- Prejudice against Computational Approaches

These reasons were identified as roots of the problem according to the following statements:

- *"Lives are on the line. Every moment counts. We have it down to a science. We know what information we need and we get in and get it... We would never try out anything new at this point."*
- *"Look. It's just not going to happen. We have to have certain kinds of information from the field and some random crowd of Twitters are not going to give it to us."*
- *"We aren't an IT company, we do relief, and IT folks are the last to get hired around here and the first laid off...They are way over worked and way under paid... There is no way we can get more out of them to do Twitter stuff. They are barely keeping our computers working as it is."*

In many cases, NGOs do not have employees with a strong background in Information Systems and Information Analytics, as the IT departments of these agencies are under-stuffed. Over-complicated models that act as black boxes, create suspicion and mistrust amongst the agencies, especially when their accuracy is not up to par with the methods that they have developed and used for years.

Additionally, a major factor for not incorporating social media information in crisis management is prejudice against this approach. Several statements find the scenario of utilising social media information in such cases completely unfeasible. This opinion in combination with a slow or even stationary stance regarding the incorporation of new methods in their operations makes the clear communication of the tool even more urgent and essential.

1.2.2 The need for social media information detection and quality assessment methods

Social media contain a vast amount of information (Diakopoulos, De Choudhury, and Naaman 2012) but only information of good quality should be utilised in the decision making of operational activities. Due to the questionable quality of User Generated Content (UGC), humanitarian agencies are still doubtful about its integration into the decision making process, thus leaving this pool of available data unexploited (Callaghan 2016;Watson and Rodrigues 2017).

Three major text-based social media information detection and verification methods were identified in the literature:

- Cross-validation
- Expert Opinion
- Crowdsourcing

Cross-validation with other data sources a.k.a. triangulation is the process where personnel from a humanitarian agency utilises additional data sources in order to validate the truthfulness of information extracted from social media (Crowley, Dabrowski, and Breslin 2013). The main limitations of such methods are the required manual input of users for validation and the direct dependence between the skill of the user and the implementation of the method (Daume, Albert, and Gadow 2014).

Expert opinion is the process when experts or people of authority utilise their expertise or authoritative sources to validate the truthfulness of social media information (Martin 2016). Such methods are restricted by the knowledge and the skill of the user, the required manual input of the user and the limited number of experts in the field (Martin 2016).

Crowdsourcing is the utilisation of Internet platforms in combination with the input of social media in order to validate the truthfulness of information harnessed from social media (Riccardi 2016). The users of the platform, verify whether the information presented is of good quality or not. Such methods are limited by the required user input to the method and the great number of users required by the platform to determine the truthfulness of the information (Basu, Bandyopadhyay, and Ghosh 2016).

This introduction to the detection and verification methods has identified a major limitation for the methods developed or utilised in the detection and verification of social media rumours. This limitation is the excessive dependence to human input and total lack of any autonomous and automatic detection tools. As timely response during humanitarian crises requires information as fast as possible this lack of autonomy creates a problem.

1.3 Problem Definition

Social media offer a vast pool of unexploited information. As discussed in Section 1.1, this data is only going to increase. This information can provide decision-makers of humanitarian crises management with valuable information and boost the confidence of the affected population. At the moment this information go to waste as they are not properly utilised.

As quality and trust issues are the biggest challenges for the utilisation of social media as a source of information and the existing work in detecting relevant social media information does not tackle these issues, establishing a method to detect relevant information is of critical importance in the field of humanitarian aid (Meier 2011). This Master Thesis project focused on detecting text-based rumours in Twitter for humanitarian activities.

As discussed in Subsection 1.2.1, a statement is considered a rumour when is unverified to be true or false at the moment of its circulation and useful to humanitarian crisis management when it is relevant to the context of humanitarian activities. As discussed in in Subsection 1.2.2, the existing detection methods are highly dependent to human input which consequently leads to slower rumour detection and verification which results in slower aid, high demand in

volunteers who could be utilised in other relief activities and scalability issues which means that the detection and verification system has a finite and limited input of information it can process. The problem defined in this Master Thesis project is the lack of automation in the detection of rumours in humanitarian activities.

In Chapter 2, a literature review has taken place in order to investigate the existing literature and explore which methods and techniques could be utilised so that the detection of rumours relevant to humanitarian activities could be automatic. After having established the qualitative background and the knowledge gap of this Master Thesis project, research questions were formulated. These research questions facilitated the research that took place in this project.

Chapter 2

Literature Review & Research Questions

2.1 Master Thesis Literature Review Scope

The objective of this Master Thesis project was to tackle the problem that was defined in Section 1.3. The problem is the automatic detection of rumours in Twitter that are relevant to humanitarian activities.

To tackle this problem, this project focused on developing a model that classifies whether a tweet constitutes a rumour or not and if it is relevant to humanitarian activities or not. In order to develop a model that would be able to execute automatically this process two literature reviews had to take place first.

The scope of the first literature review was to study the technologies that have already been researched in the humanitarian context for the detection and verification of social media information. This review indicated that out of the existing publications only Supervised Machine Learning can be used to automatically classify tweets to rumours or non-rumours and relevant to humanitarian activities or not. Thus, a more in-depth literature review took place in order to identify techniques and methods to approach the design of this tool.

In the second literature review the scope was narrowed down even more as it focused on Twitter and only in the detection of information in Twitter and not in the verification process using Supervised Machine Learning. Due to the fact that the existing publications in humanitarian research did not explore this solution, a broader perspective was obtained in order to accumulate knowledge on the field of information analytics. This broader perspective was acquired by the liberation of this literature review of the context restriction. This way articles that studied information detection regardless the field of application were reviewed too.

The knowledge background, which was acquired after conducting these literature reviews, is combined with the problem that was defined in Section 1.3 and

the Knowledge Gap of this project is defined. After acquiring the problem that this Thesis tackled and the gap of knowledge that exists in the field, research questions were formulated that facilitated the research that took place in this Master Thesis project.

2.2 Humanitarian Literature Review

As discussed in Section 1.3, the need to detect text-based rumours relevant to humanitarian activities is of great importance. This Chapter explored the existing relevant research to find technologies, methodologies and techniques to tackle the problem at hand.

In order to acquire the required qualitative background on the field of rumour detection, a literature review had taken place. As an introductory literature review indicated that there is no automatic approach for the detection of rumours in humanitarian crisis management, this literature review examined the utilisation of IT in humanitarian activities and how they contribute in solving the problem defined. Therefore, this review is not limited in the detection of relevant rumours in Twitter but in the detection of relevant information in social media in general as its objective is to create a clear depiction of the technologies utilised in humanitarian activities currently.

The literature, which was reviewed, is presented in Table 2.1. Even though, researchers point out the boost that social media information can provide to humanitarian aid, as discussed on Subsection 1.2.1 a literature review on the field indicated that the detection processes are still kept manual. The findings of this literature review indicated that most papers focus on the utilisation of crowdsourcing as means to detect relevant information. Alternative options to detect information are online platforms. None of these methods has an automatic information retrieval mechanism built in it. The only technology that showed that it can be used to develop an automatic process was Supervised Machine Learning. Supervised Machine Learning is the least researched technology out of the ones that were reviewed.

2.2.1 Crowdsourcing

This part has investigated how the existing literature for Crowdsourcing in humanitarian research can contribute to automatically detect relevant information in social media. The papers mentioned in Table 2.1 have been examined on terms of topic, approach and their contribution to the literature of the field.

Crowdsourcing has been studied to a great extend as indicated by Table 2.1. Additionally, crowdsourcing as a solution has been explored in different scenarios and approaches. This thechnology has been studied from a high-level perspective till its application and case studies.

On a high-level, the advantages and disadvantages of crowdsourcing have been explored (Gao, Barbier, and Goolsby 2011; Riccardi 2016). The accuracy of the method and the involvement of the affected population have been

#	Social Media Information Detection Method	Research Papers
1	Crowdsourcing	Gao et al. 2011 Riccardi 2016 Basu et al. 2016 Norheim-Hagtun et al. 2010 Soden et al. 2014 Carley et al. 2016 S. E. Middleton et al. 2014 Ludwig et al. 2017
2	Online Platforms	Sump-Crethar 2012 Maresh-Fuehrer et al. 2016 Goolsby 2010
3	Machine Learning	X. Liu et al. 2016 Andrews et al. 2016 Starbird et al. 2014 Hung et al. 2016

Table 2.1: Humanitarian Literature Review Overview

recognised as advantages of the method (Riccardi 2016). The disadvantages of crowdsourcing were the excessive manual input (S. E. Middleton, L. Middleton, and Modafferi 2014; Gao, Barbier, and Goolsby 2011; Ludwig et al. 2017), the disconnection between information and solution (Gao, Barbier, and Goolsby 2011; Soden and Palen 2014), the speed of the method to evaluate an information and its limitation to detect new information (Basu, Bandyopadhyay, and Ghosh 2016; Norheim-Hagtun and Meier 2010). The limitation to detect new information has been explored though in order to be tackled. Interactive crowdsourcing can be used to retrieve new information and not just classify already published information by offering the people that evaluate information the opportunity to provide the system with additional data (Basu, Bandyopadhyay, and Ghosh 2016).

The earthquake in Haiti in 2011 (Norheim-Hagtun and Meier 2010; Soden and Palen 2014; Meier 2011), the fires in Russia in 2010 (Meier 2011), the civil war in Libya in 2011 (Meier 2011), the drought in Somalia in 2011 (Meier 2011), the earthquake in Indonesia in 2009 (Carley et al. 2016) and the Matthew Hurricane of 2016 (Yuan and R. Liu 2018) were case studies that crowdsourcing has been implemented. Not only that but its combination with crisis mapping has been explored too (Meier 2011). Crisis mapping is the real-time collection, presentation and analysis of data during a crisis (S. E. Middleton, L. Middleton, and Modafferi 2014). Given that there has been studies that investigated the application of crowdsourcing to a real crisis, researchers also explored ways on how to facilitate the process in a more efficient way (Gao, Barbier, and Goolsby 2011; Ludwig et al. 2017). An valuable approach was that crowdsourcing would be more efficient if it was embedded in the everyday life of individuals (Ludwig et al. 2017).

Crowdsourcing is a valuable technology and has a lot to offer to the humanitarian operations. The fact that it involves the affected population directly to the evaluation of information boosts their confidence and raises their awareness. Even though, its disadvantages have been addressed in relevant research, the fact that its speed as a technology does not really differ from a manual process remains unanswered. Therefore, the papers that were reviewed in this section and the field in general had little impact in the design of the model of this Master Thesis project.

2.2.2 Online Platforms

The findings of this literature review indicated online platforms as an alternative method of quality assessment in the humanitarian sector other than crowdsourcing. This part examined the alignment of the technologies and methodologies utilised in online platforms to the objectives of this literature review.

Many online platforms have been introduced to the people such as BBC UGC, CNN i-Report, Storyful and ReliefWeb (Sump-Crethar 2012; Maresh-Fuehrer and Smith 2016; Goolsby 2010). For example BBC UGC and CNN i-Report are mainly used to investigate news that include visual aid in order to determine whether the image uploaded was photoshopped or not. On the other hand, ReliefWeb is mostly utilised for long-term decision making and not so much for providing information during the response phase of a crisis.

Online platforms can be of great benefit to the humanitarian agencies and the affected population as they can facilitate crisis mapping which can provide a visual presentation of the on-going event and help in the dissemination of information. The problem with this approach is that it cannot detect information on its own, which makes manual input a prerequisite in the process. Online platforms do not provide insight on how someone can create an rumour detection model. Therefore, the existing work on online platforms in the humanitarian sector can not be utilised as knowledge base for this Master Thesis project.

2.2.3 Supervised Machine Learning

Machine Learning is the technology of getting computer systems to act without being explicitly programmed (Michalski, Carbonell, and Mitchell 2013). This is achieved through data training (Alpaydin 2014). Machine learning is applied in many fields such as voice and image recognition, financial predictions, information verification and many other fields (Patterson 2010). Decision makers are still hesitant to use such methods due to accuracy rates and the fact that this method is a black box system. A black box system is a system with known input and output but no knowledge of its internal working (Saleh et al. 2016).

This part explores the potential contribution of papers in the humanitarian sector that used Machine Learning in their research. This part explored how the existing work can add to the research framework of this project. Even though Artificial Intelligence and more specifically Machine Learning is not yet

widely researched in this specific field, computational journalism investigated if Machine Learning could outperform conventional detection and verification methods which gave positive results (X. Liu, Li, et al. 2016). A quick and automated process that is not bound by the limitations of human fatigue can review significantly more entries to the system in comparison to manual browsing and filtering. This provides the decision makers of the humanitarian activities with a broader spectrum of information regarding the event.

Conventional detection methods are regarded as cumbersome and demanding when it comes to human resources (X. Liu, Li, et al. 2016). Additionally, the fact that their speed in information detection can not compete the speed of a computational approach means that the scenario of additional damage to the affected area, additional crisis aid expenditure or even additional humans deaths might be unavoidable due to time constraints.

Machine Learning and in general purely computational approaches are constantly avoided by humanitarian agencies as discussed in Section 1.2.1. Even though the accuracy rate of the Machine Learning method is still not up to par with the other available methods that does not mean that its potential contribution is lesser or that it does not posses the capability to reach these levels. The topic is not researched enough so that this knowledge gap would close.

Furthermore, in many cases that it has been studied, the researchers did not focus on the accuracy rates or providing solution to the problem which has been discussed in Section 1.3 which was the automatic rumour detection during humanitarian activities in Twitter. On the contrary, many researches focused on the analysis of social media utilisation during crisis or the network dynamics (Andrews et al. 2016; Starbird et al. 2014). Nevertheless, Machine Learning was investigated in order to assess the credibility of information posted during a flood (Hung, Kalantari, and Rajabifard 2016). The research focused on a probabilistic model and the spatial distribution of volunteer generated information. The limitation of this research is that it was explicitly developed for cases of flood. Despite, their significant overall contribution to the field, these papers do not explore the solution space of the problem that was described in Chapter 1.

Conclusion

All the technologies that were reviewed in this Section offer some level of automation but they can not facilitate an autonomous process except from Supervised Machine Learning. Even in the case of Supervised Machine Learning information detection has not been studied in the context of Humanitarian activities.

The articles associated with Machine Learning which were presented in this literature review provided with some insight and ideas on how to develop a model that detects rumours related to a humanitarian crisis but they can not act as a base for this Master Thesis project. Ideas such as the increased influence and credibility of official news channels or accounts and the dissemination of false information through big networks contributed on the design of the rumour

detection tool. Thus, the next Section explored explicitly how can existing research on Supervised Machine Learning assist in rumour detection in Twitter for Humanitarian activities.

2.3 Literature Review in Supervised Machine Learning

As discussed in the previous section, the research of the field that addresses the problem defined in Chapter 1 is limited. For this reason the literature review was liberated by the restriction of context, which was humanitarian crises. The problem defined is composed by two dimensions, rumour detection and topic classification. In the same way, the works reviewed are separated in two classes given their topic. This categorisation is presented in Table 2.2.

The selection of these works was based on their alignment with this Master Thesis project, which were to find ways to classify tweets according to their context and to classify whether it constitutes a rumour or not as it was discussed in Section 1.3.

Supervised Machine Learning is considered by experts to be highly suitable for classifying information (Brownlee 2018b). Therefore, the articles that were reviewed had applied Supervised Machine Learning. Supervised learning maps an input to an output based on sample input-output pairs (Russell and Norvig 2016). The mapping is done through the training process of the algorithm. Sample data are injected into the algorithm in order to correlate the properties of the input to the paired output and predict the output. In Supervised Machine Learning the training data have to be annotated. The annotation has to be on the property that the prediction is going to be about.

#	Machine Learning Problem Dimension	Research Papers
1	Rumour Detection	Hamidian et al. 2015 Castillo et al. 2011 X. Liu et al. 2015 Kwon et al. 2013
2	Topic Classification	K. Lee et al. 2011 Hamidian et al. 2015 Suh et al. 2017 2017

Table 2.2: Supervised Machine Learning Literature Review Overview

The next two parts examined the published work on the field of rumour detection and topic classification. The works presented in Table 2.2 were reviewed regarding the objective of their research, the sets of features that they utilised in their research, the Supervised Machine Learning algorithms they used, the datasets that they used to train the algorithm and the performance of their model. This has provided an overview of the published work on the field and

assisted significantly in the formulation of the knowledge base of this Master Thesis project.

2.3.1 Rumour Detection

The objectives of the articles in Table 2.2 are similar to the objectives of this Master Thesis project. Even though the authors explored if the tweet was a rumour or not, some of them implemented static analysis and a static predictor (Hamidian and Diab 2015; Castillo, Mendoza, and Poblete 2011), whereas some investigated the case of a dynamic predictor (X. Liu, Nourbakhsh, et al. 2015; Kwon et al. 2013). The difference is that the dynamic predictor monitors the tweet and adjusts its prediction given the dissemination of the information and other users' endorsement.

Another deviation from the objective of this Master Thesis project is that the papers that were reviewed even though they claim to study rumour detection they actually focused on veracity classification (Zubiaga et al. 2018), which is the last stage of the rumour quality assessment framework introduced in Section 1.2.1. Only Kwon et al. (2013) have contacted rumour detection. Even though their research primarily focused on the analysis of data, they also investigated classification algorithms to detect rumours automatically. Nevertheless, the approach that they have followed and the techniques that they have utilised are of great value to the objectives of this Master Thesis project.

Features are a vital component in a Machine Learning classifier. In Machine Learning, a feature is a quantifiable property or characteristic of the phenomenon that is being studied (Bishop et al. 2006). A feature is an explanatory variable used in complicated statistical techniques such as regression which is utilised in Machine Learning. The features that are used in such algorithms should be informative, discriminating and independent from each other. Feature engineering is a crucial part of developing an effective model in pattern recognition (H. Liu and Motoda 1998).

Feature engineering is the process of using expert knowledge of the domain in order to create features that will maximise the utility and efficiency of the Machine Learning algorithm (Gen and Cheng 2000). Even though it is essential, it is expensive and time-consuming.

As the only research that could be set as benchmark for this Master Thesis project is the work of Kwon et al. (2013) this article was reviewed and presented separately. The objective of the paper is to identify the driving forces of text-based rumours of generic topic. Furthermore, they applied classification algorithms and the results of their performance metrics varied from 60 to 80%. These results varied as the performance of their model did not perform well for small observation windows. The features that were used in this project were divided in the categories User, Linguistic, Network, Temporal. The weaknesses of this scientific article is that it did not incorporate real-time dimensions of the situation but focused on more mathematic and statistical features.

The works of Castillo, Mendoza, and Poblete (2011) and Hamidian and Diab (2015) mainly focused on content features. These features describe the

content of the tweet in explicit characteristics. Other feature categories that were explored were Twitter meme features, network features and social features (Hamidian and Diab 2015; Castillo, Mendoza, and Poblete 2011). Twitter meme features refer to properties that are available in Twitter and annotate the context of the tweet such as hashtags or emoticons. Network features describe the network of the user that has posted the tweet that is being analysed. Social features describe the account of the user that has posted the tweet at question.

An alternative approach was the investigation of verification and belief features (X. Liu, Nourbakhsh, et al. 2015). Given the relatively different objectives of this project the authors consulted journalists and investigated if the simulation of the cross-validation done by a journalist could be applied in a computational method. The verification features refer to static analyses which can be done by a journalist, such as whether the profile of the user has a name or not and whether the user that posted the tweet was an eye-witness of the event or not. The belief features apply a sentiment analysis on the users' reaction on the tweet to acquire insight on how a rumour is spread.

The Supervised Machine Learning classification algorithms that were investigated in the papers showed consistency as the same algorithms were investigated by multiple researchers. The most common algorithm was the Decision Tree classifier or variations of this algorithm (Castillo, Mendoza, and Poblete 2011; Hamidian and Diab 2015). Other options that were explored were Support Machine Vector, Random Forest and Multinomial Naive Bayes (Castillo, Mendoza, and Poblete 2011). Some researchers developed their own classifier by applying Bayesian statistics to their case (X. Liu, Nourbakhsh, et al. 2015).

The data that was used to train, validate and test the Information System products developed was annotated data. The annotation of the data was whether the tweet was truthful or not. The collection and the annotation of the data took place either manually (Hamidian and Diab 2015) or by the utilisation of online rumour detection platforms such as Twitter Monitor and snopes.com (Castillo, Mendoza, and Poblete 2011; X. Liu, Li, et al. 2016).

The efficiency of the sets of features that were explored indicated positive results. The user features performed more than the average (Castillo, Mendoza, and Poblete 2011; X. Liu, Nourbakhsh, et al. 2015), while network and belief features seemed to be outperformed (X. Liu, Nourbakhsh, et al. 2015; Hamidian and Diab 2015). Content-based features' performance varied from paper to paper.

The works that were reviewed provided ideas on how to deal with a rumour detection problem. A logic on how to approach the creation and formation of features was acquired by the review of these articles. Most researchers have developed an extended set of features that describe the digital profile of the user that is spreading an information (Castillo, Mendoza, and Poblete 2011; Hamidian and Diab 2015). Furthermore, these papers offer several innovative ideas regrading feature engineering, for example sentiment analysis on the positive and negative words of the tweet, the account's age (Castillo, Mendoza, and Poblete 2011), if there is a name in the account details of the user and if the user was an eye-witness to the event that he has posted (X. Liu, Li, et al. 2016).

On the other hand, these papers had weaknesses too. Some researchers investigated a rather limited set of features (Hamidian and Diab 2015). Some of the features could have been explored in a bigger depth, for example the feature hashtags could have been a numerical feature providing the model with an increased insight in comparison to the binary feature that it was (Castillo, Mendoza, and Poblete 2011; Hamidian and Diab 2015). Even though the journalistic approach in the detection of rumour was a innovative and intriguing idea, some of the features were difficult to verify such as if the source is a satirical account or not and if the tweet entails satirical urls (X. Liu, Li, et al. 2016).

These papers have provided information and insight on how to structure and design a rumour detection classifier. This review has boosted the technical knowledge base of this Master Thesis project and on the same time has set the performance benchmark that the Information System product developed in this project has to be compared to.

2.3.2 Topic Classification

In this part of the literature review, the objectives of the papers reviewed presented bigger diversity than the field of rumour detection. The objectives varied from classifying rumours per event (Hamidian and Diab 2015), classifying trending tweets in 18 major categories such as sports, music or politics (K. Lee et al. 2011) and exploring the impact of oversampling to a topic classifier's performance (Suh et al. 2017). Even though the work of Suh et al. is diverging from the objectives of this Master Thesis project, the paper goes into depth regarding the analysis and the topic classification process.

The feature categories that were utilised in these papers was Term Frequency - Inverse Document Frequency (TF-IDF) (Hamidian and Diab 2015; K. Lee et al. 2011; Suh et al. 2017) and network analysis (K. Lee et al. 2011). TF-IDF is a word vectorisation methodology that calculates the importance and the weight of each word within a text in order to assign this text to a class.

The Supervised Machine Learning classification algorithms that were used do not differ from the classifier used in Rumour Detection. As in rumour detection, the most widely used or at least experimented algorithm was Decision Tree (Suh et al. 2017; Hamidian and Diab 2015; K. Lee et al. 2011). Other algorithms that were explored by multiple researchers were Multinomial Naive Bayes and Logistic Regression (K. Lee et al. 2011; Suh et al. 2017). Additional Machine Learning classifier that were investigated were Support Vector Machine, K-Nearest-Neighbours (Suh et al. 2017).

The data collection and annotation were manually performed by all researchers. The researchers utilised the Twitter Application Platform Interface in order to download their datasets. The data were labelled to the topic or event that they belonged.

Even though the objective of this Master Thesis project is much simpler than that the ones presented in the literature review of topic classification, the articles reviewed have provided critical insight to the tweet topic classification knowledge base. The researchers explored valuable ideas. Even though the most

commonly implemented method in their work was TF-IDF, which already is a well-known method in topic classification, its implementation in the context of Twitter offered critical insight in classifying text-based information originating from Twitter (K. Lee et al. 2011; Hamidian and Diab 2015; Suh et al. 2017).

The research that has been used as a benchmark for this Master Thesis project is the work of K. Lee et al. (2011). This paper conducted a multi-class classification. More specifically, they classified the tweets in 18 categories with a 70% performance level.

On the other hand, there are significant differences between the problem that was dealt by the authors of the papers reviewed on topic classification and the classification problem tackled in this Master Thesis project. This creates doubts whether the techniques implemented in those cases would perform at the same level. The major differences is the fact that in this project, the classification takes place between two classes in comparison to multiple classes (K. Lee et al. 2011; Hamidian and Diab 2015). Further, the categories that the information are being classified are more explicit and static, for example sports, music etc. (Suh et al. 2017; K. Lee et al. 2011) in contrast to the broad concept of humanitarian activities relevant.

Conclusion

This literature review examined the two dimensions which were identified for the solution of the problem defined in Section 1.3. Therefore, this literature review was divided into two parts. The first part investigated papers relevant to rumour detection and the second part examined papers relevant to topic classification.

Regarding rumour detection, it was found that most papers that were reviewed focused on information verification even though they defined it as rumour detection. The only paper that was actually about rumour detection was the work of Kwon et al. (2013), which set the performance benchmark of this Master Thesis project's tool to 60 to 80%. Nevertheless, insights from all the papers reviewed were utilised in the design of the tool of this Master Thesis project.

One of the points that was derived from this part was that the features used in these papers were about the account of the user, the content of the tweet and the propagation of the tweet. Furthermore, the Supervised Machine Learning algorithms that were used mostly in the papers reviewed were Decision Tree, Random Forest and Support Vector Machines. Another major point was the requirement for a dataset that has the desired annotation. For example, the dataset that trains a classifier that classifies tweets to rumours and non-rumours has to have a dataset annotated in this specific dimension.

Regarding topic classification, the papers that were reviewed were investigating the classification of tweets according to their content. The techniques that were utilised during their research were text analysis techniques. The most widely used technique is Term Frequency-Inverse Document Frequency (TF-IDF) and in some cases the technique Bag of Words was also used. The classification algorithms that were most used in the papers reviewed in this Sec-

tion were Decision Tree, Random Forest and Multinomial Naive Bayes. The work of K. Lee et al. (2011) has been used as a performance benchmark for the model developed in this Master Thesis project. The performance of the model developed in the work of K. Lee et al. (2011) was at 70%.

Topic classification of tweets is already a thoroughly researched scientific field. Many techniques and methods are well-known for their performance and capabilities but also the task that they conduct best. TF-IDF is such a technique when it comes to classifying text-based tweets according to their content.

2.4 Knowledge Gap

The points derived by these reviews contributed to the formulation of the knowledge gap between the existing literature of this Master Thesis project.

The need to detect relevant rumours regarding an event during a humanitarian crisis is urgent and of great importance as it was discussed in Section 1.3. The field of data related humanitarian research is still young and has not explored all the available detection methods at the same depth. Despite the urgency of the issue, current literature has not explored automatic ways to detect rumours. Most efforts were focused on the people in affected areas through crowdsourcing or the utilisation of existing online platforms.

From the findings of the literature review that took place in Section 2.2, it can be concluded that the technology that could provide an automatic solution to the problem defined in Section 1.3 is Supervised Machine Learning. According to the literature review that took place in Section 2.3 Supervised Machine Learning has not been investigated yet to detect text-based rumours in social media for humanitarian activities but also in general the published research is rather limited. The knowledge gap of this Master Thesis project is the usage of Supervised Machine Learning in detecting text-based rumours in Twitter for humanitarian activities.

Outside the field of data related humanitarian research, rumour detection has been studied. This provides great assistance in addressing the issue at hand. Nevertheless, this research needed to be adjusted so that its implementation fits the needs of humanitarian crisis management. Most of the articles that have investigated rumour detection have focused on the last stage of the framework introduced in Section 1.2.1 which is veracity verification and not rumour detection. The work of Kwon et al. (2013) studied the first step of the framework which is rumour detection and offered a performance benchmark for rumour detection.

2.5 Research Questions

2.5.1 Principal Research Question

Integrating social media information in relief operations faces challenges that need to be answered in order to incorporate it in crisis management (Popoola

and Krasnoshtan 2013). These challenges were discussed in Section 1.2.1. This Master Thesis focuses on the following research question:

"How can Supervised Machine Learning be used to detect text-based rumours relevant to humanitarian activities in Twitter?"

The objective of this research is to acquire a solution to the problem of the detection of text-based rumours in Twitter for humanitarian activities. The problem has been identified as socio-technical as it incorporates the social aspects of humanitarian activities and the technical aspects of information systems. Technology and human behaviour are inseparable in information systems (A. Lee 2000), thus, acquiring this knowledge has required two complimentary paradigms: behavioural science and design science. Therefore, in order to answer to the principal research question a design approach has been adopted.

The design approach is at the same time a process and a product (artifact) (Walls, Widmeyer, and El Sawy 1992). The work of March and Smith (1995) identifies two processes. The processes are *build* and *evaluate*. The Build process offers the development of the artifact. This does not mean that the artifact is complete yet. The Evaluate process facilitates the assessment of the artifact but most importantly refines the developed artifact.

Both build and evaluate processes are vital to the design of an artifact. The build process allows the development of the artifact and the evaluate process allows the alignment of the design to the identified operational needs. Pragmatists argue that scientific research should always be evaluated regarding its implications to the real world (Aboulafia 1991). Therefore, the loop between build and evaluate should iterate several times to produce the final version of the artifact (Markus, Majchrzak, and Gasser 2002).

2.5.2 Sub-questions

To reach the answer to the principal research question the following sub-questions have to be answered:

1. Which methods or techniques can be used to design a model to detect text-based rumours using Supervised Machine Learning?

2. What is the performance of the Supervised Machine Learning Classifier that is used to detect whether a tweet is a rumour or not?

3. What is the performance of the Supervised Machine Learning Classifier that is used to classify whether a tweet is relevant to humanitarian activities or not?

4. Which are the most effective features for the Supervised Machine Learning Classifier that detects whether a tweet is a rumour or not?

Chapter 3

Model Design

3.1 Introduction

In Chapter 3, the design of the model developed for this Master Thesis project is presented. The model was first developed on a conceptual level, then its structure was developed. This conceptual model was validated in Chapter 4.

Chapter 3 presents the conceptual design of the model developed, the feature engineering required for the Supervised Machine Learning modules that classify the tweets to rumours and non-rumours and relevant to humanitarian activities or not and the selection of the classification algorithms that are used in the these modules.

3.2 Model Description

The objective of the model designed is to predict the classes that the input belongs to. Classification predictive models use statistics to map function (f) from input variables (X) to discrete output variables (y) (Asiri 2018). Predictive modelling can have a wide range of applications.

In this section the functions of the model are described on a conceptual level. The model developed in this project has as its objective to detect rumours relevant to humanitarian activities in Twitter. Thus, the model conducts three separate processes. The processes are the following:

- Represent the properties of the tweets into measurable, quantitative values that can be used in a Machine Learning algorithm.
- Classify tweets whether they are relevant to humanitarian activities or not.
- Classify tweets whether they are rumours or not.

Predictive Machine Learning models are composed by two groups of modules. The first group transforms the input of the model to features suited to a

Machine Learning algorithm. The second group conducts the predictive part of the model. The model developed in this project addresses the processes mentioned above. The model developed in this Master Thesis project is named *Humanitarian Rumour Detector*.

Each process was conducted by a different module. The module responsible for extracting and formatting the features from the Twitter feed is the *Preprocessing Module*. The module responsible for classifying the information gathered from Twitter for relevance is the *Humanitarian Relevancy Classifier* and the module responsible for detecting if the tweet is a rumour is the *Rumour Detector*.

Each module performs one of the processes defined by the objectives of the project. The modules work in sequence, providing the user with an estimation regarding the relevant rumours posted in Twitter. An alternative approach would have been to combine the Humanitarian Relevancy Classifier and the Rumour Detector. This would have resulted in a classifier that would have been conducting a multi-class classification.

This approach was not explored due to the fact that such a Supervised Machine Learning classifier would require multi-dimensional datapoints in order to be trained and to be tested. By extension, this would also apply in the real-time usage of such a classifier, which might complicate things in certain circumstances.

The input of the model is the Twitter feed at question. The only information that is required is the tweet ids of the Twitter feed. Tweet id is a unique identification number that is assigned to each tweet by Twitter. The output of the model is a list of tweets that have been deemed as rumours relevant to humanitarian activities. The input of the model first goes through the Preprocessing Module, in order for the features to be extracted. More details in the feature extraction process are provided in the next subsection. The model designed for this Master Thesis project is presented in the Figure 3.1.

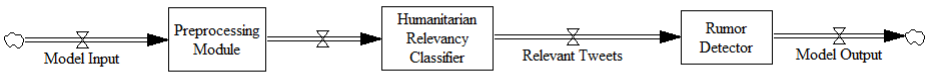

Figure 3.1: Humanitarian Rumour Detector Presentation

As presented in Figure 3.1, the flow of data pass from the Preprocessing Module to the Humanitarian Relevancy Classifier. The exclusion of the datapoints tha are not of use is done to enhance the performance of the classifier (X. Liu, Li, et al. 2016). Then, these tweet ids are matched with the extracted features from the Preprocessing Module in order to act as the input in the Rumour Detector. The output of the last module is also the output of the model.

3.3 Feature Engineering in Preprocessing Module

The Rumour Detector and the Humanitarian Relevancy Classifier perform classifications of different context. Thus, the features that are used in the Rumour Detector are different from the ones that are used in the Humanitarian Relevancy Classifier. In the following part a detailed presentation and explanation of the selection and extraction process of the features utilised in this Master Thesis project is provided. First the features from the Rumour Detector are presented followed by the presentation of the features utilised in the Humanitarian Relevancy Classifier.

3.3.1 Features in Rumour Detector

The feature engineering of this Master Thesis project was based on the papers that were reviewed in Section 2.3. These papers represented the account of the user, the content of the tweet and the propagation of the tweet to quantitative properties that can be used by a Machine Learning classifier. This aligned with the objective of the Preprocessing Module that was presented in Section 2.1.1.

In essence, the authors of major papers on the field such as the ones that were discussed in Section 2.3.1 attempted to create a digital profile of the person that has posted the tweet at question and not just an analysis of the text of the tweet (Castillo, Mendoza, and Poblete 2011; Qazvinian et al. 2011; X. Liu, Li, et al. 2016). By creating a digital DNA, a researcher is able to accumulate information and proceed to classifications.

The features of this research were divided in three categories: *1) User features, 2) Linguistic features and 3) Meta-content features.* This categorisation was based on the papers reviewed in Section 2.3.1. These categories were the ones that were used most often and showed the most promise according to their results. A brief explanation for the categories is provided below:

- **User Features:** the objective of these features is to describe the user that has posted the tweet from the properties that can be provided from his or her Twitter account.
- **Linguistic Features:** the objective of these features is to describe the content of the tweet by translating non-quantitative properties into quantitative ones so that they can be utilised by a Machine Learning classifier.
- **Meta-content Features:** the objective of these features is to represent the propagation of the tweet at question in Twitter.

Through the extraction of information from Twitter API this project aspired to create as many features as possible to describe the above-mentioned properties to the model developed in this project. This is shown as all the available information which are presented in Figure 3.3 are transformed into features

that can be used in a Supervised Machine Learning classification module. As it is the first time that such a computational approach has been implemented in rumour detection, this project covered a broad spectrum of features in order to investigate their performance. The extraction process is presented in Figure 3.2. A detailed presentation of the features is presented in Appendix A. Additionally, in order to point out the contribution of this Master Thesis project to the computational approach of rumour detection an additional Table is presented in Appendix B that classifies the features using the following columns:

- *Original:* A binary column that indicates whether a feature is an original product of this Master Thesis project or not.
- *Adjusted:* A binary column that indicates whether a feature already exists in published literature but is adjusted to the needs of this Master Thesis project.
- *Existing:* A binary column that indicates whether a feature already exists in published literature and is used in this Master Thesis project as it is.
- *Reference:* This column reports the paper that either an Adjusted or an Existing feature originates from.

The input to this module is the tweet_id of the tweet at question. Providing this information to the Twitter API, it is possible to extract an object which entails all the available information in the Twitter database regarding the tweet and the account that posted the tweet. An object in computer science can be a variable, a data structure or a method. This object is a value in the memory of the operating system referenced by an identifier (Fleming, Von Halle, et al. 1989).

All the necessary information is then extracted from this object and is used for the creation and formation of the features. Some of the tweets included in the dataset had been deleted and no information could be extracted for these tweets. Figure 3.3 explains in more detail the information retrieval step of this module and presents the information extracted by Twitter API. Tweepy was utilised in order to connect to Twitter and to extract all the available information. Tweepy is an open-sourced Python library that facilitates the communication between Python and Twitter.

When the information regarding the tweet and user are available, the formulation of features begins. The features can be formulated in the following stages:

1. Quantitative Information directly extracted from the Twitter Object.
2. Qualitative information turned into quantitative information.
3. Combination of extracted information in order to create a new quantitative feature.

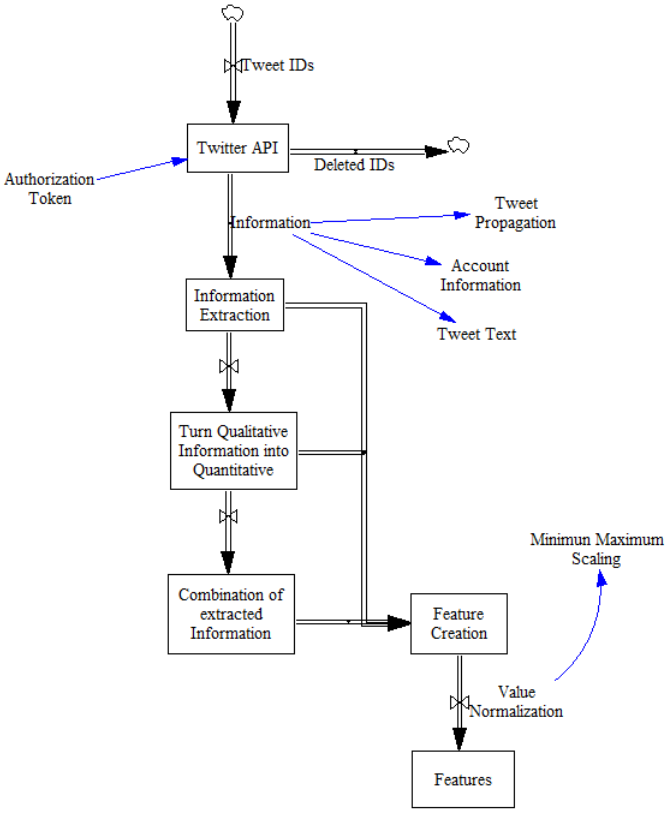

Figure 3.2: Feature Extraction Model Diagram

Different features are created in each stage. The first stage formulates numerical features that can be directly extracted from the Twitter API. A novelty of this Master Thesis project, apart from the creation of original features, is the utilisation of the numerical values of many features instead of a binary interpretation. For example, hashtags, urls, retweets were investigated as binary values in the papers reviewed in Section 2.3.1 while in this project they were investigated as numerical features. Numerical features incorporate an increased volume of information in comparison to binary features as they describe not only if an element exists or not but also what is its value.

Features that are created in this stage are straightforward as they already have a numerical value to be extracted and utilised by the Humanitarian Rumour Detector. Figure 3.4 describes the first stage of the features formulation and categorises the formulated features to their corresponding group.

The only features that require some additional work instead of a simple extraction are tweet_age and profile_age which indicate how old is the tweet and the profile of the user respectively.

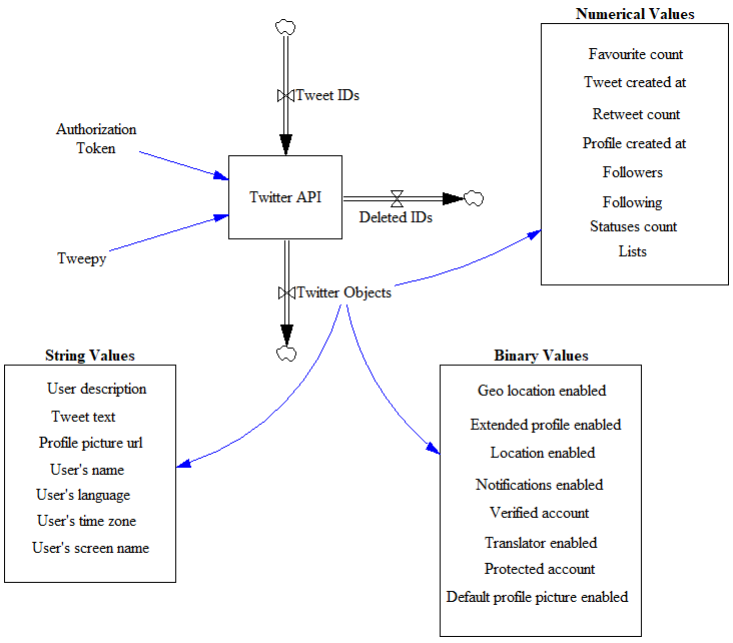

Figure 3.3: Information Retrieval Step

As some of the characteristics of the user or the content of the tweet can only be expressed in a qualitative way, therefore string and binary (True or False) values are transformed into numerical values. The model transforms them in quantitative elements so that they can be utilised by a Machine Learning classification module, which in essence is a statistical model. The features created can either be numerical or binary (0 or 1).

Figure 3.5 describes the second stage of the features formulation and categorises the formulated features to their corresponding group.

In User Features, binary values of True and False are represented as 1 and 0 such as if the account has enabled the tracking of his geographical position or if the user has verified his account.

Qualitative values such as the name of the user and whether the user has a profile picture or not are also investigated. The features default_profile (Table A.1, #13) and face_in_picture (Table A.1, #14) investigate how much did the user proceed in customising his or her account. The feature default_profile indicates whether the user has changed or not the default profile picture provided by Twitter and the face_in_picture examines if there is a discernible human face in the profile picture or not. These two features demonstrate the dedication of the user to his or her account and if they feel represented by this account or not. The steps that are followed for the extraction of these binary features are the following:

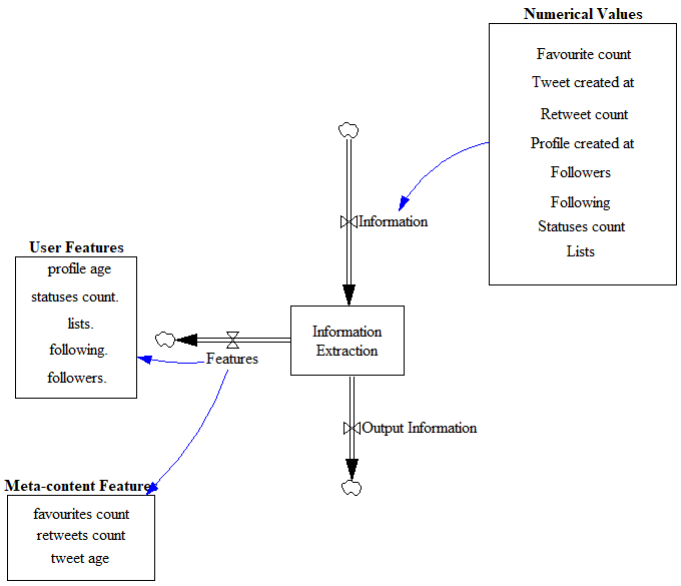

Figure 3.4: Feature Creation First Stage

1. Extract the url of the profile picture of the user
2. Download the profile picture of the user
3. Scan the picture using the Face_recognition Python Library, which is a face recognition library
    - Face_recognition recognises and manipulates faces from Python
    - The model has an 99.38% accuracy (Geitgey 2018)
4. The output of the face recognition library is a string containing the positioning of the face in the picture

The features name_sex (Table A.1, #15) and male_name (Table A.1, #19) examine the name that the user has provided Twitter with. Name_sex investigates whether the name provided corresponds to either a male or female name. Male_name attempts to take this a step further by looking into the name and matching it to a specific sex. The steps that are followed for the extraction of these binary features are the following:

1. Extract the name of the user
2. Examine whether the name of the user includes a person's name using the TextBlob library

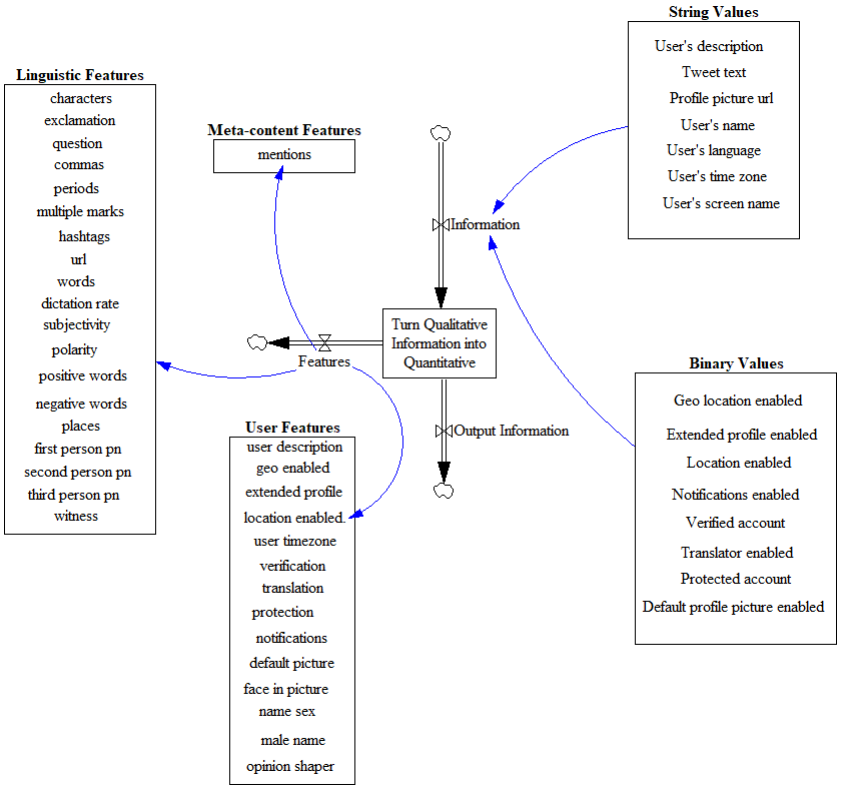

Figure 3.5: Feature Creation Second Stage

- TextBlob is a Python library for text analysis. An API conducts Natural Language Processing (NLP) tasks.

3. The name of the user is examined whether it is male or female

The feature opinion_shaper (Table 1, #17) investigates whether the screen-name of the user that posted the tweet is an opinion-shaper. Andrews et al. (2016) in their research argue that major news-reporting accounts had a significant impact to the dispersion of information in social media and more specifically in Twitter. They identified that a particular set of words if included in the screen name of the user indicate that this account could potentially be a news account or an opinion shaper. According to X. Liu, Li, et al. (2016), this has an impact in the credibility and propagation of a tweet. Thus, the steps that are followed for the extraction of this binary feature are the following:

1. Extract the screen name of the user/ text the tweet
2. Check if any of the words included in the following lists is entailed in the corresponding element

- opinion=['news','breaking','news','report','report','daily','times', 'feed', 'radar', 'net']

Most of the Linguistic Features are created in this stage. This is expected as most of the Linguistic Features originate from analysis of the content of the tweet.

By applying sentiment analysis to the content of the tweet, it is possible to calculate the positive and negative words in the tweet. Moreover, by utilising the TextBlob library it is possible to further analyse the sentiment of the content by calculating the polarity and the subjectivity of tweet.

By text analysis of the tweet, it is possible to count the number of hashtags, characters, question-marks, commas or periods. Additionally by utilising the TextBlob library it is possible to calculate the dictation rate of the tweet.

The feature dictation_rate examines how accurate is the dictation of the tweet that the model is classifying. In order to create the feature, these steps are followed:

1. Extract the text of the tweet
2. Correct the dictation of the text using the TextBlob Python library, which is a text analysis library
    - Spelling correction is based on the work of Norvig (2007)
3. Calculate the similarity of the original of the tweet and the text with the corrected dictation using a string sequence matcher

The features that are created in the third stage of the Preprocessing Module are features that derive from the combination of extracted information, the combination of already created features or the combination of extracted information with already created features. The novelty of this project is that even though this model is a static classifier, it incorporated features that simulate temporal features as it inserted the parameter of time in crucial dimensions such as the number of retweets, followers of the account that posted the tweet and the total number of tweets published by the user. This approach expands the scope of the features that are used in this model as temporal features are a major dimension of features as discussed in Section 2.3.1. Figure 3.6 describes the third stage of the features formulation and categorises the formulated features to their corresponding group.

In this stage, the already acquired information were explored in order to create insightful combinations so that the model could cover a broader and more complete spectrum of information and quantify the digital DNA of the user, the content of the tweet and its propagation.

The features favourite_per_second (Table A.3, #4) and retweet_per_second (Table A.3, #5) indicate not only how many accounts have favoured or retweeted the tweet but also the speed that this information has spread through the digital

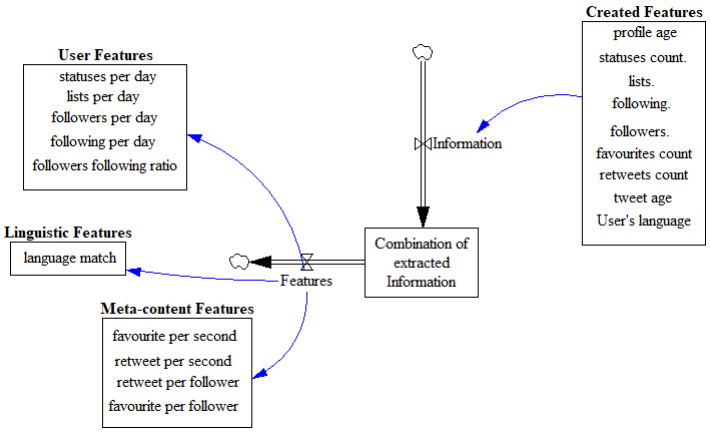

Figure 3.6: Feature Creation Third Stage

world. In a way, one could say that it is an indicator whether this tweet has become viral or not for the time being.

The features followers_per_day (Table A.1, #22) and following_per_day (Table A.1, #23) indicate the speed at which the network of the account has spread. These features have as their objective to support accounts that are still new in Twitter and still developing. Accounts that exist in Twitter for extended periods of time could possibly amass great amounts of followers or post many statuses without being particularly popular or active. These features can detect accounts that are trending or discover tweets that have become viral. This could prove to be vital during a humanitarian crisis that a rather small or unknown account could post an information of great importance.

The features retweet_per_follower (Table A.3, #6) and favourite_per_follower (Table A.3, #7) indicate the engagement of the followers of the account in the tweet at question. This does not simply characterise this individual tweet but also the network of the account in general, as popular and reliable accounts are engaged by their followers usually.

The feature followers_following_ratio (Table A.1, #24) indicates ratio between the people that follow this account and the accounts the user follows. Given the fact that some users might attempt to build a network by following a great number of accounts without really being active in Twitter, this ratio is an important indicator of the account at question.

The features statuses_per_day (Table A.1, #17) and lists_per_day (Table A.1, #18) attempt to capture how active is the account of the user that has posted the tweet. This way, the frequency at which the user has posted tweets can be discovered and the engagement of the user can be quantified.

The features favourite_per_second, retweet_per_second, followers_per_day, following_per_day, followers_following_ratio, retweet_per_follower, favourite_per_ follower, statuses_per_day and lists_per_day could be characterised as ratio features.

Even though Linguistic Features were exhausted in the second stage of the feature creation, the combination of information led to the creation of the feature lang_match. The feature lang_match (Table A.2, #21) examines whether the language of the tweet at question matches the language of the Twitter account of the user. In order to create this binary feature, these steps need to be followed:

1. Extract the the language of the account of the user from Twitter API
2. Identify the language of the text by using TextBlob library
    - Language detection is powered by the Google Translate API (Loria et al. 2014).
3. Compare the values of these two properties

The last step of the Preprocessing Module is the normalisation of the values of the features. Normalised values train Machine Learning classifiers more efficiently, which results in better performance levels (Ioffe and Szegedy 2015). Standardisation of datasets is a common requirement for many Machine Learning estimators. They might behave badly otherwise (Pedregosa et al. 2011).

In this case, the normalisation that has been utilised was based on the minimum-maximum scaling. This normaliser is scaling features to lie between a given minimum and maximum value. This method was chosen due to its robustness in small standard deviations (Pedregosa et al. 2011).

This model in combination with the dataset that is introduced and explained in Chapter 4 were utilised in order to create a training dataset for the Rumour Detector. This module defines the features and the way that these features are created from a given dataset. This process is only a part of the whole process as the injection of data in the model facilitates the training of the Machine Learning classifier. Nevertheless, the design of the model is one of the most crucial parts of this Master Thesis project.

3.3.2 Features in Humanitarian Relevancy Classifier

In order to classify the tweets that are used as input to the Humanitarian Rumour Detector, the analysis has focused on identifying the context of the tweet. In the conceptualisation and creation of the features utilised in the Rumour Detector, the objective was to create a digital DNA of the proprieties and values of the network, the account settings and the text of the tweet posted by the user. In this classifier, the objective is to represent in a quantitative manner the context of the tweet at question. Therefore, an analysis of the text and more specifically an analysis of the words in the tweet was required.

Word frequency and word importance are key parameters in determining the context or even better the topic of a given tweet or any generic piece of text. In order to accomplish this objective, Term Frequency-Inverse Document Frequency (TF-IDF) was utilised. This technique was chosen because of the literature review of major articles of the field that took place in Section 2.3.2. In

most cases, TF-IDF was the only option that was explored for topic classification (Hamidian and Diab 2015; Suh et al. 2017).

This method is usually used in information classification and text mining. During this method a weight is assigned to each word in the dataset. This weight is a statistical measure to evaluate the importance of the word in a dataset of strings. The relative importance increases by the number of times it appears in a specific tweet but decreases by the frequency it appears in the whole dataset (Leskovec, Rajaraman, and Ullman 2014). In this technique the selection of the dataset that is training the algorithm is crucial as it directly affects the performance of the model. To formulate these weights, two components are required:

- Term Frequency
- Inverse Document Frequency

Term Frequency (TF) measures the frequency of a term appearing in the tweet. Due to the fact that the length of a tweet varies, this value is normalised by dividing the frequency by the total number of words in the tweet. The value of TF ranges between 0 and 1. TF takes the value 0 when the term is not included in the tweet and the value 1 when it is the only word in the tweet. For example, the value of this variable for a term x in a tweet i is:

$$TF(x,i) = \frac{Number\ of\ times\ term\ x\ appears\ in\ tweet\ i}{Total\ number\ of\ terms\ in\ tweet\ i}$$

Inverse Document Frequency (IDF) quantifies the importance of a term. During the calculation of TF, the terms are considered of equal importance. Therefore, there is a need to differentiate common terms that are used and rare ones. IDF's value ranges from 0 when the term is in every tweet of the dataset and could possibly reach infinite values if the volume of the dataset is too great and the term is included in a very limited amount of tweets. To achieve this, the following value is computed:

$$IDF(x) = log_e \frac{Total\ number\ of\ tweets}{Number\ of\ tweets\ with\ term\ x\ in\ it}$$

Having these two values, it is possible to calculate the TF-IDF(x,i) for the term x in the tweet i. The formula is presented below:

$$TF - IDF(x,i) = TF(x,i) * IDF(x)$$

By combining these two values, a composite weight for each term in each tweet is created. This value is at its highest when a term appears many times in a specific tweet and only in a few tweets. The value is lesser when this term appears fewer times in a tweet or appears in many tweets. Finally, the value is at its lowest when the term appears almost in every tweet.

This results in the creation of a vector for each tweet. This vector contains one component corresponding to each term in the dataset and a weight for each component that is given in the last equation (Christopher, Prabhakar, and Hinrich 2008).

Data has to be injected into the algorithm in order for the weights mentioned in the TF-IDF process to be defined. These features in combination with the dataset that is presented in Chapter 4 determined how the frequency of words in a tweet assisted to its topic classification into relevant and non-relevant to humanitarian crises.

3.3.3 Feature Engineering Conclusions

The Preprocessing Module extracts the required information for the tweet at question, formulates and creates the features tha are used in the Rumour Detector and in the Humanitarian Relevancy Classifier.

The Rumour Detector uses three sets of features, User Features, Linguistic Features and Meta-content Features. The features that are used in Rumour Detector attempt to classify the tweets at question given real-life properties of a tweet in comparison to the work of Kwon et al. (2013) that used statistical properties of the tweets at question. The features that are used in Rumour Detector are 53 and are divided into Original features, Adjusted features and Existing features. The Existing features that are used are 17, the Adjusted features are 9 and the Original features are 27. The original features were created and tested in this Master Thesis project for the first time in order to classify tweets to rumours and non-rumours. The major hurdle in the design of this classifier was the limited pool of published works in the field.

The Humanitarian Relevancy Classifier uses the text analysis technique TF-IDF in order to create features to classify the tweets at question according to their content to relevant to humanitarian activities and not relevant. This technique calculates the importance of each term and the frequency of each term in the dataset. TF-IDF is widely considered to be highly suited for topic classification and text categorisation, which was also indicated by the literature review that took place in Section 2.3.2.

When comparing the two classification modules, the design of the Rumour Detector was a much more challenging endeavour due to the lack of published literature on the matter. Moreover, the task of rumour detection was a much more complicated one in comparison to topic classification, as the first required socio-technical approach and perception in comparison to topic classification that required detailed text analysis. In the case of the Humanitarian Relevancy Classifier, the research of this Master Thesis project sufficed in validating the approach of applying the TF-IDF technique and to test its performance.

3.4 Algorithms for Classification Modules

The Humanitarian Relevancy Classifier and the Rumour Detector are two classifiers that make estimations regarding a certain property of the tweet based on a fitted statistical model. In the case of Humanitarian Relevancy Classifier, the property is if the tweet is relevant to humanitarian crisis or not and in the case of the Rumour Detector is whether the tweet is a rumour or not. The algorithms for the classifiers were not developed in this Master Thesis project but already developed algorithms were utilised. The algorithms were from the Python Library *scikit-learn*.

This Master Thesis project investigated which classification Machine Learning algorithms are best suited for the model that has been designed. The classification algorithms that have been utilised in order to classify the tweet to categories are the following:

- Decision Tree
- Support Vector Machines (SVM)
- Multinomial Naive Bayes (MNB)

These algorithms were investigated due to the literature review of the field of text-based topic classification and rumour detection presented in Section 2.3.

Due to the time limitations of a Master Thesis project, only two classification algorithms were chosen to be investigated in the experiments of this Master Thesis project per classification module. The rational behind the selection was to choose one rather simplistic and straight-forward classification algorithm and one more complicated and sophisticated classification algorithm. The Decision Tree Classification algorithm was chosen as the more simplistic algorithm for both classifiers. As for the more sophisticated algorithms, SVM was chosen for Rumour Detector and MNB was chosen for Humanitarian Relevancy Classifier.

Decision Tree as a Machine Learning algorithm utilises the Decision Tree Framework from decision analysis in order to draw a conclusion for the value of a variable. In this tree structure, the leaves of the tree represent class and branches represent conjunctions of features that lead to those class (Rokach and Maimon 2008).

Support Vector Machine is a Machine Learning algorithm that can be utilised either for classification or regression. This algorithm assigns new examples randomly to the categories available, which makes SVM a non-probabilistic classifier (Cortes and Vapnik 1995). SVM depicts the datapoints as points in space in order for them to have a distinct gap between them. Predictions are made by printing the new examples to the space developed during training and classify them to the corresponding category (Friedman, Hastie, and Tibshirani 2001).

The Multinomial Naive Bayes Classifier is regarded as a rather popular choice when it comes to classifying text-based information. This classifier is

suggested in the book of James et al. (James et al. 2013). Furthermore, this algorithm and its performance were investigated in the work of Imran, Elbassuoni, et al. with positive results (Imran, Elbassuoni, et al. 2013).

Naive Bayes classifiers are a family of probabilistic classifiers based on applying Bayes' theorem with independence assumptions between the features (Russell and Norvig 2016). These classifiers are a popular method to categorise text or documents in general utilising word frequency as a feature (Rennie et al. 2003). In Multinomial Naive Bayes Classifier, samples represent the frequencies of certain events that have been generated by a multinomial vector (McCallum, Nigam, et al. 1998).

The methods described above have been investigated regarding their suitability to the model developed in this Master Thesis project. Details regarding the inner workings of the these methods are presented in Chapter 4 along with the experimentation set-up and the experiments.

Even though the design of the Humanitarian Rumour Detector was completed at this point of the process, the model itself was not complete. In order for the model to be completed and validated, data points had to be injected so that the statistical model entailed within the model was fitted to the needs of the model. After applying data to the model, the model was evaluated by being compared to benchmarks that have been set by similar existing model retrieved by the literature. All this process has acted as a validation of the model and its utility.

The next Chapter completes the presentation of the model delivering the final validated version of the model developed in this Master Thesis project.

Chapter 4

Model Validation & Performance Analysis

The main focus of Chapter 4 is the validation of the model and the analysis of the performance of the model.

The model that was designed in Chapter 3 that classifies tweets to rumours or non-rumours and relevant to humanitarian activities or not relevant, is a conceptual model. In order to validate this model an annotated dataset is required. The datasets that were used for the training, testing and validation of the model are presented in Section 4.1. These datasets were used in the experiments.

The experiments of this Master Thesis project were conducted in order to tune the hyperparameters of the classification algorithms that are used in the Rumour Detector and the Humanitarian Relevancy Classifier, to find which algorithms perform best and which features are the most influential in the Rumour Detector. The experiment design is presented in Section 4.2.

The model was validated with the method 5-fold cross-validation. The results of the 5-fold cross-validation were then compared to the performance benchmarks that were set in Section 2.3. Additionally, the operational times of the model are evaluated for their capacity to handle the volume of information produced during a humanitarian crisis.

The last part of Chapter 4 focuses on analysing the results of the experiments. The analysis investigated how the tuning of the hyperparameters of the classification algorithms affect the performance of the model and the influence of individual features in the decision making process of the Rumour Detector.

4.1 Experimentation Dataset

4.1.1 Dataset Overview

The Humanitarian Relevancy Classifier and the Rumour Detector are both major components of the model developed in this project. As both these modules are basically Supervised Machine Learning classifiers, they required annotated data in order to be trained. The datasets that have been used for the training of the Humanitarian Relevancy Classifier and the Rumour Detector originate from the works of Kwon et al. (2013) and Imran, Mitra, and Castillo (2016). The Rumour Detector was explicitly trained by the dataset of Kwon et al. (2013), while for the training of the Humanitarian Relevancy Classifier both datasets were used.

Kwon et al. (2013) have accumulated a great quantity and variety of rumour and non-rumour tweets. Their dataset includes 140.000 datapoints corresponding to 110 events. Out of the 110 events, 60 are rumours and 50 are non-rumours. None of these events is related to humanitarian activities. Therefore, they can be used in order to check and validate both classifiers. More details regarding the events that are included in the dataset used in this project are provided in Appendix C. The datapoints were collected from online platforms such as snopes.com and archives.com for rumours and cnn.comn and times.com for non-rumours. This dataset was manually annotated by experts to rumours and non-rumours (Kwon et al. 2013).

The data points were annotated whether they were a rumour or not on the same criteria that were set in this Master Thesis project in 1.2.1. In order for a tweet to be considered a rumour it has to be an information statement, be unverified, related to an event and circulate Twitter around the time of the event. The additional annotation do not affect the rumour annotation as a rumour being proven to be false does not change the fact that it is a rumour.

This dataset was selected to be used in this Master Thesis project due to its variety of 110 events, the great number of 140.000 datapoints that it includes and its annotation to rumour and non-rumour. As the dataset entails a great number of topics that were not of particular relevance to the scope of this project, the naming that was chosen for it, was *Generic Tweets*.

This dataset was used to train both the Humanitarian Relevancy Classifier and the Rumour Detector. The great number of events that is included in this dataset assisted in the generalisation of the model in classifying tweets that were not used during its training. In order to avoid any bias either towards rumours or non-rumour the number of datapoints was reduced to the point that the training dataset of rumour tweets equals the number of non-rumour tweets, which allowed the utilisation of 84.000 datapoints in the training of the Rumour Detector. Given the fact that the ratio of the dataset was 2:1 in favour of the rumour tweets the model would develop a bias towards classifying a tweet as a rumour which would distort the results of the model. Therefore, the dataset was limited to 84.000 data points so that the ratio of rumour to non-rumour tweets was 1:1. As Imran, Mitra, and Castillo (2016) collected more than 50.000.000

tweets related to humanitarian crises, all the available tweets from the work of Kwon et al. (2013) were utilised in the training of the Relevancy Classifier.

As mentioned in the previous paragraph, Imran, Mitra, and Castillo (2016) have collected over 50.000.000 tweets relevant to humanitarian activities. This dataset was selected due to its quantity but also the great variety that it incorporates as it studied 19 different humanitarian crises in order to train the Humanitarian Relevancy Classifier.

The data was collected by Imran, Mitra, and Castillo (2016) by AIDR, which is a information retrieval tool for Twitter developed by the same authors (Imran, Castillo, et al. 2014). After collecting relevant tweets for the 19 humanitarian crises that they were studying, annotation took place. The annotation was performed by volunteers. The tweets were annotated on whether they were reporting injuries or deaths, missing individuals, infrastructure damage, donation offers or emotional support (Imran, Mitra, and Castillo 2016).

The datasets of this Master Thesis project were structured in a way that there are enough datapoints and variety of datapoints so that the Machine Learning algorithm is generalised. Additionally, as the modules of the model are performing binary classification, the datasets were structured in a way that the algorithm did not acquire any bias towards any of the classes. As in the case of the Rumour Detector, the ratio of humanitarian relevant to non-humanitarian relevant tweets was 1:1 so that the model would not develop a bias towards any of the classes. Therefore, 140.000 datapoints were utilised in the training of the Humanitarian Relevancy Classifier from the work of Imran, Mitra, and Castillo (2016) in order to match the available datapoints acquired from the work of Kwon et al. (2013).

4.1.2 Training Dataset Structure

Training Dataset for Humanitarian Relevancy Classifier

As discussed before, the objective of the Humanitarian Relevancy Classifier is to classify text-based rumours according to their relevance to humanitarian activities relevant or not. Therefore, the dataset utilised for the training should be a compilation of annotated tweets of generic topic and a collection of tweets from various humanitarian crises. These crises should be both man-made disasters and natural disasters in order to train the model in a wider range of cases. An overview of the 140.000 tweets relevant to humanitarian activities are presented in Figure 4.1.

As mentioned before, the dataset utilised to train the Humanitarian Relevancy Classifier should be an anthology of events. For this reason, in order to find the proper ratios between the different events included in this dataset, humanitarian crises were separated in two categories, natural disasters and man-made disasters or biological crises. The dataset was structured in a way that the ratio between natural disasters and man-made disasters or biological crises is 1:1 and the ratio between man-made disasters and biological crises is 1:1.

In order to structure the training dataset of the Humanitarian Relevancy

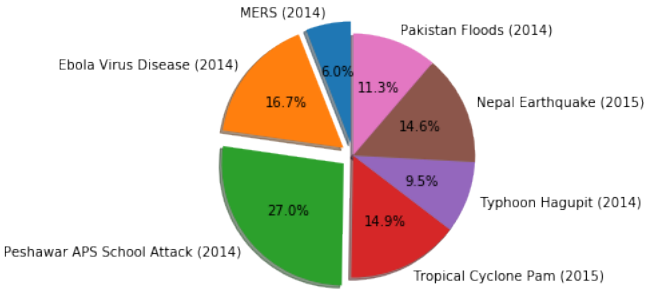

Figure 4.1: Humanitarian Dataset Overview

Classifier, the humanitarian relevant datapoints have to be collected first. The tweets that were utilised in this project are presented below:

- The Nepal Earthquake of 2015 (23.487 tweets)
- The Tropical Cyclone Pam 2015 (20.863 tweets)
- The Severe Tropical Cyclone Pam of 2015 (20.863 tweets)
- The Pakistan Floods of 2014 (15.834 tweets)
- The Peshawar APS School Attack of 2014 (37.943 tweets)
- The Ebola Virus Disease of 2014 (23.487 tweets)
- The Middle East Respiratory Syndrome (MERS) (8.392 tweets)
- Generic Tweets (140.910 tweets)

The dataset selected includes not only both man-made or a single a natural disaster but a variety of natural disasters. Even though the Peshawar APS School Attack of 2014 is a security issue due to the fact that it was not possible to acquire datapoints relevant to a war or an attack that could be classified as a humanitarian event, these data points were utilised so that the module could be trained in this aspect too. This Master Thesis project utilised the dataset acquired from the work of Imran, Mitra, and Castillo (2016) in a different way than it was used in the original paper. This project explored how the structure of the dataset can contribute in the identification of humanitarian related text-based tweets.

Training Dataset for Rumour Detector

The objective of the Rumour Detector is to be able to classify if a text-based information posted in Twitter is a rumour or not. Thus, a compilation of rumours and non-rumours were required for the training of this module. For

this Master Thesis project the dataset from the work of Kwon et al. (2013) has been utilised.

The dataset provided in the work of Kwon et al. (2013) is composed by over 110 events. This diversity facilitates high hopes that the classifier has not overfit to the dataset of its training. An overfitted model is a statistical model that been fitted too much to the datapoints of the training dataset. This model may ultimately fail to fit additional data or predict reliably (Pressr 2018).

The dataset used to train the Rumour consists of 84.000 datapoints extracted from the dataset Generic Tweets. This Master Thesis project utilised the dataset acquired from the work of Kwon et al. (2013) in a different than it was used in the original paper. This project explored how this dataset can contribute in the identification of text-based rumours in Twitter using a broad spectrum of Machine Learning features.

4.2 Experimentation Design

4.2.1 Experimentation Set-up

In Chapter 4, this project explored which of the Supervised Machine Learning algorithms are most efficient for the Classifiers of this model. According to Section 3.3, the classification algorithms that were chosen to be investigated are the following:

- Decision Tree Classifier (Rumour Detector & Humanitarian Relevancy Classifier)
- Support Machine Vectors Classifier (Rumour Detector)
- Multinomial Naive Bayes Classifier (Humanitarian Relevancy Classifier)

Apart from exploring which algorithm performs best, the hyperparameters of these algorithms have to be tuned in order for the algorithm to perform optimally (Claesen and De Moor 2015). In Machine Learning, a hyperparameter is a parameter of the algorithm that has been set before the training process, whereas the value of a parameter has been derived through training (Bergstra and Bengio 2012). For example, a parameters of an algorithm is the weight of the features or the TF value of a term whereas a hyperparameter is the number of the leaves used in the Decision Tree Classifier or the value of C in the SVM Classifier. Different algorithms have different hyperparameters as their design is based on different statistical equations and hypotheses (Thornton et al. 2013). By having the hyperparameters of the algorithms tuned, it is possible to compare the available algorithms amongst each other and decide which one is the most suitable.

There are three cases when it comes to the range of a hyperparameter. The first case is that the minimum and the maximum value of the hyperparameter are already set, given the nature of the hyperparameter. The second case is that either the minimum or the maximum value of the range is already set given the

nature of the hyperparameter. Last case is that there is not a minimum or a maximum value but a range of values that needs to be explored. This project investigated the hyperparameters for the algorithms that were explored in order to choose the best solution.

The experiments found the optimal values for the hyperparameters of each classifier. Each hyperparameter had a set of values that were explored in order to discover the optimal one. Each set of values has two attributes: *1) Range, 2) Number of values within this range.*

Depending on which category the range of the hyperparameters falls into, the number of values entailed within this range is affected too or at least the process to find this number. In general, the number of values within the tuning range is minimised in to order to save time and computational power (Duan, Keerthi, and Poo 2003). Nevertheless, this number should be enough so that the experimentation can provide sufficient insight regarding the inner workings of the classifier. As a result, in the first case, as the extreme values of a hyperparameter contribute to the detail of a classifier, these extremes are chosen so that they offer a meaningful contribution to the performance of the classifier. In essence, this a trade-off between the level of detail presented in the model and saving time and computational power (Bechhofer 1995). On the second case, an iterative process is followed in order to acquire a pattern explaining the correlation between the hyperparameter and the performance of the classifier (Bechhofer 1995).

After having discovered the most suitable algorithm for each module and having tuned the hyperparameters of these algorithms, the validation process was completed by comparing these modules to other similar state-of-the-art Information System models. The model validation process in software engineering is checking if the model satisfies the use, it was designed and created for. For example questions such as whether the user requirements are fulfilled but also the stakeholders' requirements should be satisfied too (Rakitin 2001). There are two ways to validate the model: *1) internal validation, 2) external validation.*

In internal model validation, it is assumed that the stakeholders' objectives have been understood and are expressed in the design of the model completely and precisely. As a consequence, the validation of the model is achieved if the model meets the required performance specifications (Hevner et al. 2004). On the other hand, external validation requires the active feedback and input of the stakeholders of the issue that the model solves (Albrecht and Gaffney 1983). In this Master Thesis project, the validation that has been implemented to the modules developed was internal.

The modules were validated by applying 5-fold cross-validation and comparing the model to similar state-of-the-art models on their performance levels. K-fold cross-validation in Machine Learning is a sampling process utilised to evaluate algorithms on a limited training dataset (Kohavi et al. 1995). The only parameter in this process is the parameter k, which refers to the number of groups that the training dataset is split into.

In this project, the value of k was chosen to be 5 so that in the cross-validation process the ratio of training data points to testing datapoints was

4:1. It is a popular method used in many major papers of the field (Castillo, Mendoza, and Poblete 2011; Qazvinian et al. 2011; X. Liu, Li, et al. 2016) as it generally provides a less biased estimate of the model's performance methods (Braga-Neto and Dougherty 2004). The sequence of this process follows the steps below (Brownlee 2018a):

1. Shuffle the dataset
2. Divide the training dataset into k groups
3. Each group is used as the testing sample of the run
4. The evaluation scores of each run is kept
5. The performance of the model is evaluated given the evaluation scores of all the runs of the model

This way the dataset has been divided into a training and testing dataset 5 times. The ratio of the training to the testing dataset was 4:1 as mentioned before. This method decreases the chance of having biased performance results as the model is trained and tested in every point of the dataset at some point of this process. The performance metrics of the model are calculated as the mean of the performance metric of each iteration.

4.2.2 Supervised Machine Learning Algorithm Grid-search

A Supervised Machine Learning Algorithm Grid-search is the experimental exploration of the values of the hyperparameters of a Supervised Machine Learning algorithm.

Decision Tree Hyperparameter Tuning

The hyperparameters that have been used for the optimal fitting of the algorithm in the case of the Decision Tree classifier are presented below (Pedregosa et al. 2011):

- Maximum Depth of the Tree : The maximum number of layers the tree possesses in order to divide the dataset into classes. [10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, None]
- Minimum Sample Number per Leaf : The minimum number of samples require in a leaf node of the tree. [1, 2, 4, 8]
- Minimum Sample Number per Split : The minimum number of samples required to split an internal node of the tree. [2, 4, 6, 8, 10]

Support Vector Machine Hyperparameter Tuning

The hyperparameters that have been used for the optimal fitting of the algorithm in the case of the Support Vector Machine classifier are presented below (Pedregosa et al. 2011):

- Kernel : A kernel can be seen as a similarity function. Given two options the kernel calculates a similarity score. This function can several forms. The ones that have been explored in this Master Thesis project are : *1) Radial Basis, 2) linear*
- Gamma : Gamma is the kernel coefficient for the Radial Basis kernel. [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0]
- C : C is a penalty parameter of the error term. [1, 10, 100, 1000]

Multinomial Naive Bayes Hyperparameter Tuning

The hyperparameters that have been used for the optimal fitting of the algorithm in the case of the Multinomial Naive Bayes classifier are presented below (Pedregosa et al. 2011):

- Alpha : Alpha is a smoothing parameter. The value of this parameters ranges from 0, where there is no smoothing, till 1, where the effect is maximised. [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0]
- Priors : This hyperparameter takes two values. The value can be either True or False. When the hyperparameter takes the value True the classifier learns the class prior probabilities, whereas when the hyperparameter takes the value False a uniform distribution is assumed as class prior probability.

4.3 Model Validation Results

In this Section, the results from the experiments are presented. The results that are presented correspond to the best performing hyperparameter fittings. The results of the model were evaluated in two dimensions timeliness and performance. This two dimensions were chosen for the validation of the model as they are the most crucial dimensions as timely and accurate information is of critical importance in aid operations.

In order to evaluate the timeliness of the model, its performance was evaluated on the premise of dealing with the volume of information that are posted during a humanitarian crisis. To evaluate the performance of the model, performance metrics were compared with performance benchmarks that have been set in Section 2.3.

In order to explain the performance metrics the results of the experiments were reported on, four fundamental terms of binary classification have to be explained. The terms are the following:

- True Positive(TP): TP is an element when it is classified by the model as a positive element correctly.
- False Positive(FP): FP is an element when it is classified by the model as a positive when it is not.
- True Negative(NP): TN is an element when it is classified by the model as a negative element correctly.
- False Negative(FN): FN is an element when it is classified by the model as a negative element when it is not.

The performance metrics the results of the experiments were reported on are the following:

- Accuracy: It is the percentage of correct classifications over the sum of elements.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

- Precision: It is the percentage of the elements that were classified as positive correctly.

$$Precision = \frac{TP}{TP + FP}$$

- Recall: It is the percentage of the positive elements that were retrieved.

$$Recall = \frac{TP}{TP + FN}$$

As mentioned in Section 3.3, the results that are presented in this Master Thesis project are the mean values of the performance metrics as the iterative process of training and testing was repeated 5 times following the 5-fold cross-validation method.

4.3.1 Rumour Detector Results

The results of the Rumour Detector are presented in Table 4.1. The performance metric that was chosen to define the best performing hyperparameter tuning was Recall. Recall was chosen due to the fact that it is the performance metric that measures the percentage of the positive elements that were retrieved out of the sample.

The figures that are presented in the Table 4.1 are the results of the 5-fold cross-validation of the Rumour Detector that took place during the training of the model.

Method	Accuracy	Precision	Recall
Decision Tree User	0.567	0.55	0.718
Decision Tree Linguistic	0.675	0.669	0.695
Decision Tree Metacontent	0.516	0.515	0.607
Decision Tree All	0.664	0.655	0.732
SVM User	0.524	0.627	0.729
SVM Linguistic	0.7	0.701	0.698
SVM Metacontent	0.526	0.532	0.425
SVM All	0.697	0.697	0.758

Table 4.1: Rumour Detector 5-fold Cross-validation Results

As the results presented were chosen on their Recall, the results from different algorithms or sets of features were not compared on Accuracy or Precision as this does not match the best performing scores of that algorithm or set of features. For example, even though the Accuracy performance of the Linguistic set of features with the Decision Tree Classifier outperformed the User set of features with the same Classifier, they can not be compared as these models were chosen because of their Recall performance.

In both algorithms, the model performed best when it utilised all sets of features with 73.2% in the Decision Tree Classifier and 75.8% in the SVM Classifier. This indicates that the combination of the features used in this project assists in detecting rumours in Twitter. The features that performed best on their own were the User Features with 71.8% in the Decision Tree Classifier and 72.9% in the SVM Classifier. Linguistic Features were the other set of of features that performed well with 69.5% in the Decision Tree Classifier and 69.8% in the SVM Classifier. Metacontent Features were significantly outperformed by the other sets of features scoring 60.7% in the Decision Tree Classifier and 42.5% in the SVM Classifier.

No relevant research was found that implemented an automatic rumour detection approach according to the literature review of Section 2.2 in the humanitarian context. The research of Kwon et al. (2013) was the only paper similar to the objectives of this Master Thesis project. The performance of the model developed in this project is up to par with the performance of the model developed in the work of Kwon et al. (2013).

User Features might have performed best in the Recall but in both Classifiers their performance in the other metrics was not up to par. On the other hand, Linguistic features performed well in all metrics. The whole set of features performed on good levels in all performance metrics. It could be deduced that User Features are detecting more rumours than the other sets of features but are making more mistakes in the process in comparison to the Linguistic Features. Therefore, their combination might be the reason why the whole set of features outperformed the other features and additionally scored well in all metrics.

Metacontent features were consistently and significantly outperformed by all the other sets of features. This might have been due to the fact that many of

the Metacontent features attempted to simulate temporal features of a dynamic classifier when the input of the model was not real time data. This must have created difficulties to the model assigning proper weights to the features in order to classify the tweets into rumours and non-rumours.

The performance of the sets of features that were developed and used in this Master Thesis project have similar performance with the corresponding sets of features in the works of Castillo, Mendoza, and Poblete (2013), X. Liu, Nourbakhsh, et al. (2015) and Hamidian and Diab (2015).

Out of the two classification algorithms that were investigated, SVM performed better in identifying the tweets that were rumours out of the sample. This is not that unexpected as SVM is considered to be a more sophisticated classification algorithm than Decision Tree.

Table 4.2 presents the time that was required for the training of the Rumour Detector but also the time this module requires approximately for the classification of unknown input in the module. The operational times that are presented in Table 4.2 correspond to hyperparameter tuning that are presented in Table 4.1. The numbers that are presented in Table 4.2 are in seconds and they correspond to 67.200 training datapoints and 16.800 testing datapoints. Additionally, the operational times that are presented are the mean values as in Table 4.1.

Method	**Fit Time**	**Score Time**
Decision Tree User	0.257	0.029
Decision Tree Linguistic	0.176	0.029
Decision Tree Metacontent	0.268	0.034
Decision Tree All	0.338	0.033
SVM User	630.426	155.879
SVM Linguistic	5212.43	174.328
SVM Metacontent	488.886	181.493
SVM All	16468.7	314.435

Table 4.2: Rumour Detector Operational Times

The operational time difference between the Decision Tree algorithm and the SVM algorithm is more than significant. Rumour Detector is able to classify 16.800 tweets in less than half a second. Using the volume of tweets during the events of Black Saturday as a standard, a typical inflow of Tweets would be 9.000 tweets per minute (O'Brien 2011). Under these circumstances, Rumour Detector would be more than capable to deal with this kind of input with the Decision Tree classification algorithm.

Rumour Detector did not perform that well when it used the SVM classification algorithm. When using all the available features, it required around 5 minutes to classify 16.800 tweets. This performance is far from bad as it can approximately classify more than 3.000 tweets in a minute with high performance levels but it could not cope with an inflow of 9.000 tweets per minute.

Nevertheless, with a more careful and sophisticated design of the kernel or just by increasing the computational power, this problem could be solved.

4.3.2 Humanitarian Relevancy Classifier Results

The figures that are presented in the Table 4.3 are the results of the 5-fold cross-validation of the Humanitarian Relevancy Classifier that took place during the training of the model.

Method	**Accuracy**	**Precision**	**Recall**
Decision Tree	0.88	0.941	0.84
MNB	0.95	0.937	0.966

Table 4.3: Humanitarian Relevancy Classifier 5-fold Cross-validation Results

The results of this classifier can be compared to the results of papers reviewed in Subsection 2.3.2. The Humanitarian Relevancy Classifier outperforms the Classifier developed in the work of K. Lee et al. (2011). This was not unexpected as in the case of K. Lee et al. (2001), a multi-class classification was conducted whereas in this Master Thesis project a binary topic classification was conducted.

As indicated by the figures in Table 4.3, the Multinomial Naive Bayes Classifier outperformed the Decision Tree Classifier.

Table 4.4 presents the time that was required for the training of the Humanitarian Relevancy Classifier but also the time this module requires approximately for the classification of unknown input in the model. The operational times that are presented in Table 4.4 correspond to the hyperparameter tunings that are presented in Table 4.3. The numbers that are presented in Table 4.4 are in seconds and they correspond to 224.000 training datapoints and 56.000 testing datapoints. Additionally, the operational times that are presented are the mean values as in Table 4.3.

Method	**Fit Time**	**Score Time**
Decision Tree	50.939	0.85
MNB	0.769	0.314

Table 4.4: Humanitarian Relevancy Classifier Operational Times

The operational times of the Humanitarian Relevancy Classifier were remarkable. Either using the Decision Tree classification algorithm or the MNB classification algorithm, the module was able to classify 56.000 tweets in under a second. This is more than sufficient in order to classify the input of 9.000 tweets per minute. Therefore, in the case of the Humanitarian Relevancy Classifier, there is no trade-off between performance and operational time as in the case of Rumour Detector.

4.4 Rumour Detector: Performance Analysis

In this Section, the model and the results of the model from the experiments that took place in order to tune the hyperparameters of the Supervised Machine learning algorithms are analysed. From the analysis of the model and the results of the model, this project acquired insight on the inner workings of the model and what are the important elements in detecting a rumour.

4.4.1 Rumour Detector Hyperparameter Tuning Analysis

As the features sets that performed best were All Features the analysis of the hyperparameter tuning that was investigated was the experiments for the model that used all the features developed for this Master Thesis project.

Decision Tree Classifier

In Figure 4.2, it is presented how much fitting time is required for the training of the Decision Tree Classifier, how does this affect the Recall performance metric and what is the correlation to the maximum depth of the decision tree.

Figure 4.2: Rumour Detector Decision Tree: Mean Recall to Fit Time

From this Figure, 3 insights can be acquired for the fitting of this Decision Tree Classifier. First, the fitting of the classifier is very quick as it fits the dataset to the Decision Tree Classifier in less than 1 second in all cases. Secondly, the time that is required for fitting of the data to the module increases as the maximum depth of the tree increases which is not unexpected. The last and most important note of the three is that the performance is not affected by the fitting time of the algorithm but from the maximum depth of the tree. Moreover, these two variables are inversely correlated. The algorithm fits optimally if the depth of the tree does not exceed the value of 10.

In Figure 4.3, it is presented how much time is required for the algorithm to classify new information after being trained, how does this affect the Recall performance metric and what is the correlation to the maximum depth of the decision tree.

Figure 4.3: Decision Tree: Mean Recall to Score Time

The insights that can be acquired from Figure 4.3 are the same as the ones acquired from 4.2. The Decision Tree Classifier is extremely quick in classifying information after being trained as it was in its training too. The correlation between the variables of the figure are same as in Figure 4.2. The only difference that can be noted is that the difference in scoring time is smaller than in the case of fitting time which is not unexpected as the ratio of the input of the training to the input of the testing is 4:1 as mentioned in Section 4.2.1.

In the previous graphs, it was explored how the maximum depth of the tree affects the performance of the model. In Figure 4.4, it is explored how the Minimum Sample Number per Split and the Minimum Sample Number per Leaf affect the performance of the model. Figure 4.4 presents only the results for the decision trees that their depth did not exceed the value of 10. These results are presented as they were the best performing.

For each value of the Minimum Sample Number per Leaf the performance of the model follows a certain pattern. This pattern is the same for different values of the maximum depth of the tree. The best performing combination is the value 2 for Minimum Sample Number per Leaf with low values of Minimum Sample Number per Split as for values higher than 6 the performance of the model decreases. If Minimum Sample Number per Leaf take the value of 8 the performance of the model becomes independent from the value of Minimum Sample Number per Split.

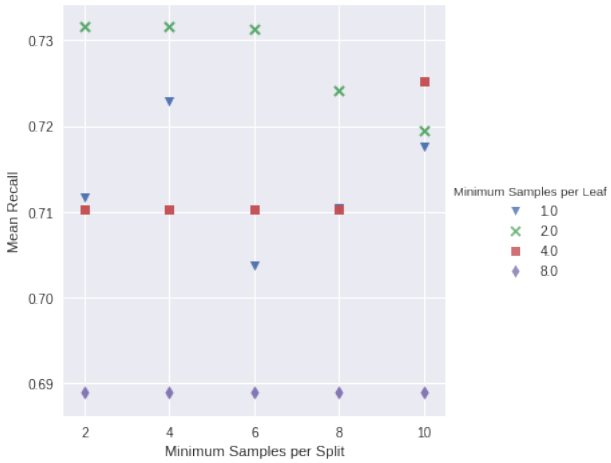

Figure 4.4: Decision Tree: Algorithm Hyperparameter Tuning

SVM Hyperparameter Tuning

In Figure 4.5, it is presented how much fitting time is required for the training of the SVM Classifier, how does this affect the Recall performance metric and what is the correlation to the hyperparameter C of the algorithm. Moreover, the results are divided into columns given the kernel that was used. The left graph corresponds to the rbf kernel and the right graph to the linear kernel.

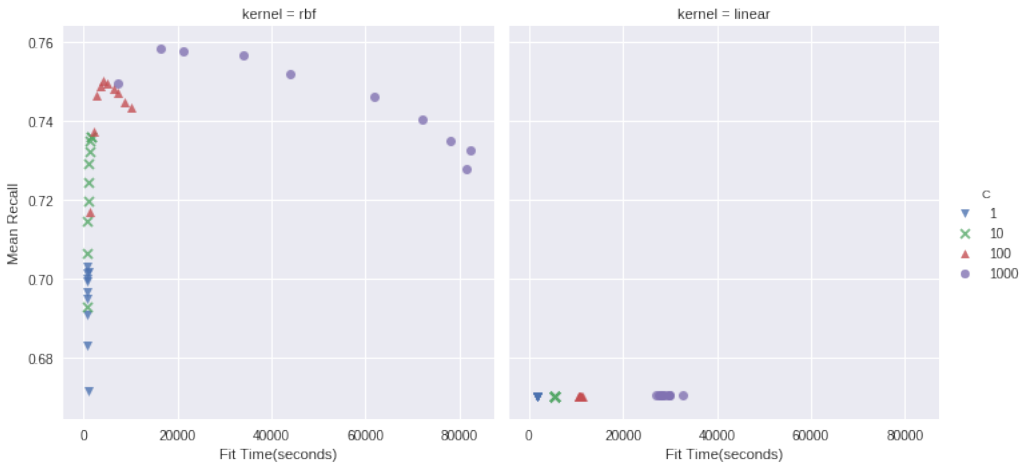

Figure 4.5: SVM: Mean Recall to Fit Time

From Figure 4.5 the major insight that can be acquired is that the performance of the module is highly correlated to the hyperparameter C of the algorithm. As the value of the hyperparameter C of the algorithm increases so

does the performance of the classification module. Moreover, the value of the hyperparameter C affects the required fitting time of the module significantly as indicated in Figure 4.5.

As indicated in Figure 4.5, for every value of the hyperparameter C the performance of the classifier follows a certain pattern. The different values correspond to different values of the hyperparameter gamma. The pattern is that it increases until a certain point and then it starts decreasing. The fact that the increase in the case of C having the value 1000 is slight means that further increase of the value would not benefit the model.

The rbf kernel outperforms the linear kernel in rumour detection. Additionally, when the algorithm used a linear kernel the module was independent to the variations of the hyperparameters of the algorithm. As the same behaviour is noticed in all the graphs regarding the linear kernel, in the next figures the results of the linear kernel are only presented and no further commenting takes place as it is not beneficial. This means that SVM can not separate the dataset in a better way linearly. This does not mean though that it can not provide with important insight on the driving forces of the decision process of the algorithm.

In Figure 4.6, it is presented how much time is required for the algorithm to classify new information after being trained, how does this affect the Recall performance metric and what is the correlation to the kernel used by the module. As in Figure 4.5, the results were divided into columns given the kernel that was used, the same is applied in Figure 4.6. The left graph corresponds to the rbf kernel and the right graph to the linear kernel.

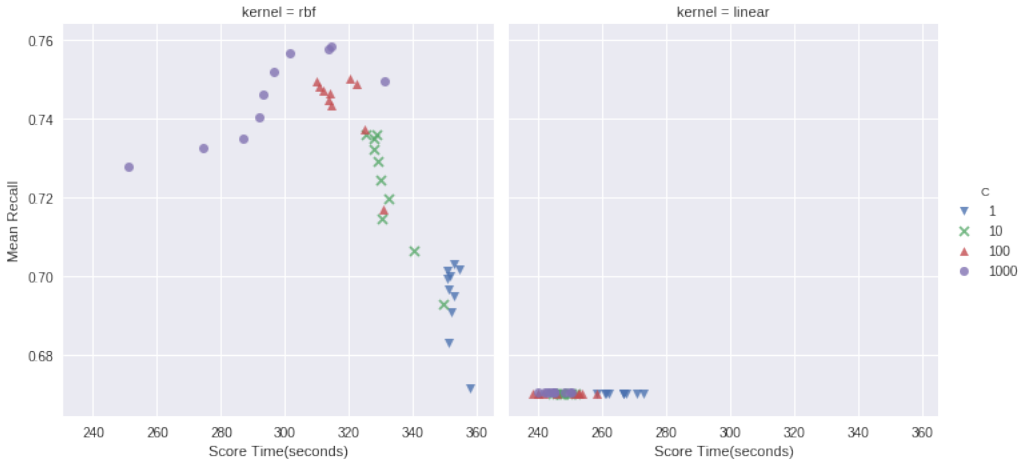

Figure 4.6: SVM: Mean Recall to Score Time

As in Figure 4.5, the results are separated according to the value of the hyperparameter C. The points that have the same value follow again a certain pattern as in Figure 4.5. Points with value 1000 seem to follow the same pattern as in the previous graph where the the increase is followed by a decrease. In the

cases of the other values of the hyperparameter the performance of the model decreases while the scoring time increases.

A contradiction to the Figure 4.5 is the fitting time required and the scoring time are in reverse analogy given the value of the hyperparameter C. Even though high values of the hyperparameter C required more fitting time, in the case of the scoring time the opposite occurs.

In Figures 4.5 and Figure 4.6, it was explored how hyperparameter C and kernel affect the performance of the model. In Figure 4.7, it is explored how the hyperparameter gamma affects the performance of the model. As gamma is a coefficient that applies only in the case of the rbf kernel in Figure 4.7 only the performance of the rbf SVM is presented.

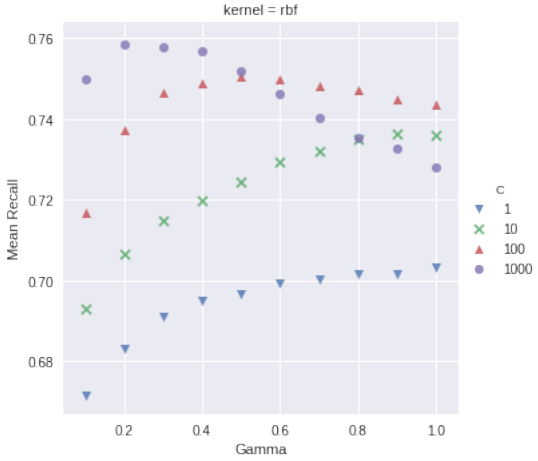

Figure 4.7: SVM: Hyperparameter Tuning

Figure 4.7 supports the findings of Figures 4.5, 4.6 and explains the direction of the slopes for its value of the hyperparameter C. The direction of the slope in the previous figures is due to the value of the hyperparameter gamma. As gamma increases each line follows the same pattern as in Figures 4.5 and 4.6. For high values of C low values of gamma are beneficial, whereas for low values of C, increasing gamma benefits the performance of the model.

Classification Module Comparison

In the previous parts of this Section, the analysis of the results of the training of the Classification Modules explored was presented. From the Figures 4.2, 4.3, 4.5 and 4.6, it was concluded that the fitting and the scoring time required by each algorithm does not affect the performance of the model. Nevertheless, these Figures are crucial for comparing these classifiers.

Even though the SVM Classifier outperformed the Decision Tree Classifier, its operational time is significantly larger. The Decision Tree Classifier is trained

and classifies information almost instantly in comparison to the SVM Classifier that required significant training and scoring time. The most critical comparison is the difference in scoring time as timeliness during crisis is one of the motivating dimension of this research. In the case of the Decision Tree Classifier the classification was instant when the SVM Classifier required around 5 minutes at its optimal fitting.

Nevertheless, this is a trade-off of model performance to operating time. The time required to train and to score could decrease with more computational power or an improved design of the Classifier. Even in this case the complexity of the SVM Classifier and the increased computational requirements of the algorithm compared to the Decision Tree Classifier would remain.

4.4.2 Feature Importance for Rumour Detector

This part investigates the importance of each feature in the Rumour Detector. Feature importance is the contribution of each feature to the prediction or classification made by a model.

Feature Importance assists significantly in the Feature Selection process. Feature Selection is the process of selecting the most useful features of the model as it benefits the model in : *i) easier interpretation, ii) shorter training times, iii) enhanced generalisation* (James et al. 2013). This Master Thesis project did not do a Feature Selection. This project investigated the importance of the features of the model for an easier interpretation and a deeper understanding of the model.

The problem tackled in this project is a socio-technical problem as it incorporates both aspects. As discussed in Section 1.2.1, one of the challenges met in the integration of rumouring in crisis management is trust issues. A deeper understanding in how the features of the model contribute to model prediction can help resolving this issue.

The method that was used in the investigation of the importance of the features of this module, was feature ranking with correlation coefficients. The algorithm that was used in this analysis was Linear SVM classification algorithm. Linear SVM offers the option to retrieve composite weight of each feature in the classification of the tweets to rumours and non-rumours. Additionally, the SVM algorithm is particularly convenient for this analysis as it tries to minimise the contribution of non-important features (Guyon et al. 2002). These composite weights are directly extracted from the functions of a trained Supervised Machine Learning module from the Python scikit library. This analysis is available and possible in SVM only when using the Linear kernel due to the way that the problem is solved (James et al. 2013).

As discussed in Section 4.4.1, Linear SVM is outperformed by rbf SVM due to the non-linear dataset that has to be separated. The Linear SVM model performed 68% in recall, 64% in accuracy and 65% in precision. This can be translated as that these are the performance levels that this dataset can be classified with a linear method. Even though outperformed by the rbf SVM algorithm, Linear SVM still performed within the range of the performance of

the model developed in the work of Kwon et al. (2013). This can provide with a deeper understanding of the individual contribution of each of the features of the module in the classification of a tweet to a rumour or not.

Figure 4.8 presents the weight of the 10 most dominating features. This selection of features corresponds to the 84% of the contribution of features in the classification process of the Linear SVM algorithm.

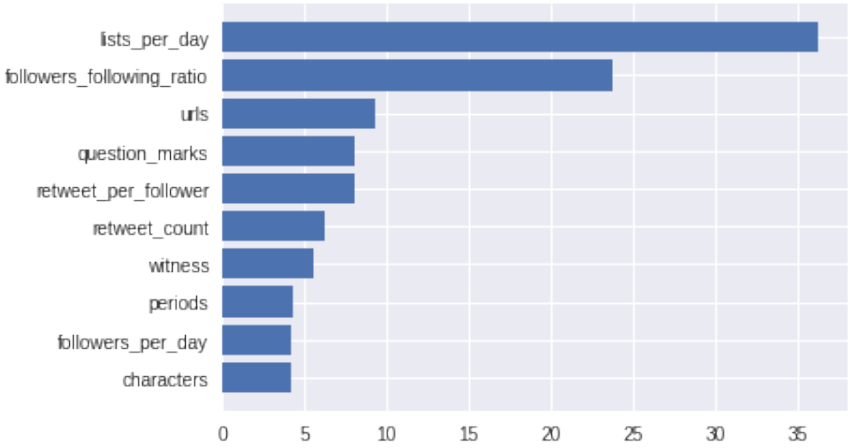

Figure 4.8: Feature Importance of Linear SVM

In the 10 individual features that are contributing the most in this classification model, there are 5 features from the Linguistic Features, 3 features from the User Features and 2 features from the Meta-content Features. This is rather unexpected as the Meta-content Features set did not perform that well in classifying the tweets to rumours and non-rumours. Nevertheless that does not exclude individual features from being useful for this classification model. Even though the Linguistic Features outnumber the User Features in this selection, the most influential features are lists_per_day and followers_following_ratio, which belong to the User Features. The selection of these features can provide information on which aspects are the most influential when it comes to classify whether a tweet is a rumour or not.

The User Features that are in this selection, indicate that the most influential features are the ratio features that were created during the third stage of feature creation as described in Section 3.2.1. The features that contributed the most described how active is the account of the user that posted the tweet in relation to the age of his account (lists_per_day, followers_per_day). Additionally, the ratio of followers of this account to the number of accounts that this user follows was an important feature.

Most of the Linguistic Features describe the structure of the text of the tweet. The features question_marks, periods and characters may be the most fundamental features in describing the structure of a piece of text. Furthermore,

two more dimensions are added with the features urls and witness that belong in this selection. The urls feature indicates a tweet that includes other information than just the text as it provides links to other sources either to support its claims or provide additional information. The witness feature indicates whether the user possibly is a witness of the event at question through text analysis.

The Meta-content Features in this list focus on the propagation of the tweet. More specifically, they focus on the retweets of this tweet. It is not only the number of retweets but also the ratio of retweets to the number of followers of this account. These two features not only show the engagement of this tweet in general but also the relative engagement.

To sum up, the dimensions that influence the decision making of this module are whether the account of the user that posted the tweet at question is active and engaging in relation to its age, a structured text including a link that either supports the claim or provide additional information, an indication that the user might be an eye-witness of the event that he is describing and propagation metrics in relation to his followers. This analysis boosted the contribution and the novelty of this Master Thesis project as 4 out of the 10 most influential features were features that were created in this project.

4.5 Humanitarian Relevancy Classifier: Performance Analysis

In this Section, the results from the hyperparameter tuning experiments of the Humanitarian Relevancy Classifier are presented and analysed. TF-IDF, which was the technique that was used in this classifier, can not provide insight on how and why the tweets are classified the way they do. Therefore, the analysis is only about the effect of the hyperparameters in the performance of the model and the time required for its operation.

4.5.1 Decision Tree Classifier

In Figure 4.9, it is presented how much fitting time is required for the training of the Decision Tree Classifier, how does this affect the Recall performance metric and what is the correlation to the maximum depth of the decision tree.

Figure 4.9 indicates that the required fitting time and the performance of the model are highly correlated to the maximum depth of the Decision Tree. As the value of the maximum value of the depth of the tree increases so does the performance of the model and the training time required. Additionally, the smaller the value of the hyperparameter minimum leaf samples the more time of training is required. The most notable insight from the results of the experiment with this algorithm is that the performance of the model is maximised when the limit of the depth of the decision tree is removed, which means that the depth of the tree exceeds the value of 110. These values are presented in the corresponding figures with the value of 0.

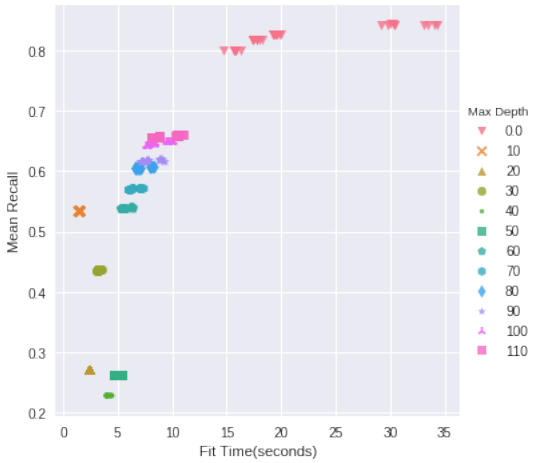

Figure 4.9: Humanitarian Relevancy Classifier Decision Tree: Mean Recall to Fit Time

In Figure 4.10, it is presented how time is required for the algorithm to classify new information after being trained, how does this affect the Recall performance metric and what is the correlation to the maximum depth of the decision tree.

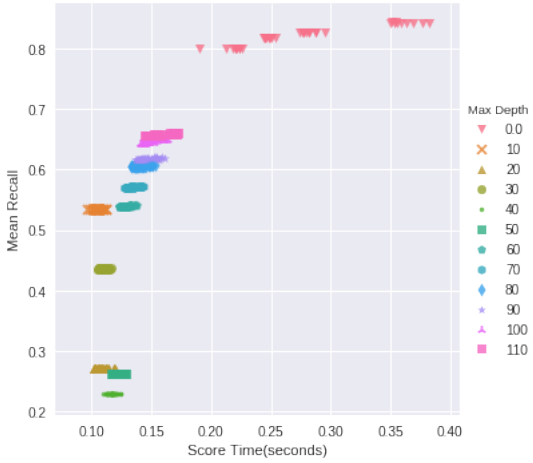

Figure 4.10: Humanitarian Relevancy Classifier Decision Tree: Mean Recall to Score Time

Figure 4.10 offers the same insights as the Figure 4.9. Furthermore, in both cases the Decision Tree Classifier is extremely fast. It requires in most cases less than half a second in order to be trained and it classifies an input of 56000

tweets in less than half a second.

In Figure 4.11, it is presented how the hyperparameters of this algorithm affect the required training time of the model.

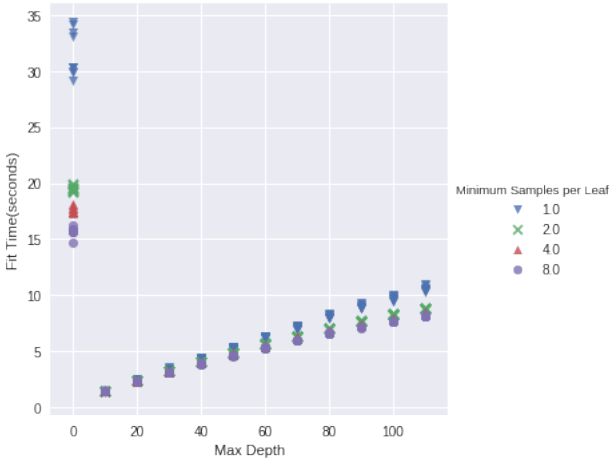

Figure 4.11: Humanitarian Relevancy Classifier Decision Tree: Hyperparameter Tuning

Figure 4.11 explains and supports the claim that the required training time of the model and the value of the minimum number of leaf samples are negatively correlated. By combining the insights acquired from Figures 4.9 and 4.11, it is indicated that the structure of the Decision Tree that is most beneficial for this model is a deep tree with a small number of samples per leaf.

4.5.2 MNB Classifier

In Figure 4.12 (and 4.13), it is presented how much fitting (scoring) time is required for the training of the MNB Classifier, how does this affect the Recall performance metric and what is the correlation to the hyperparameter alpha of the algorithm. Moreover, the results are divided into columns according to the distribution of class prior probabilities. In the left graph the class prior probabilities are utilised whereas in the right graph a uniform distribution is assumed.

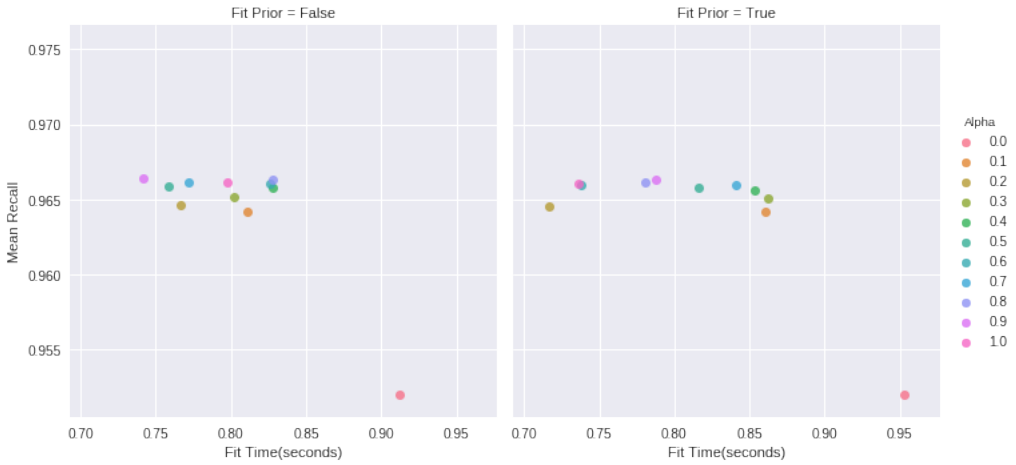

Figure 4.12: Humanitarian Relevancy Classifier MNB: Mean Recall to Fit Time

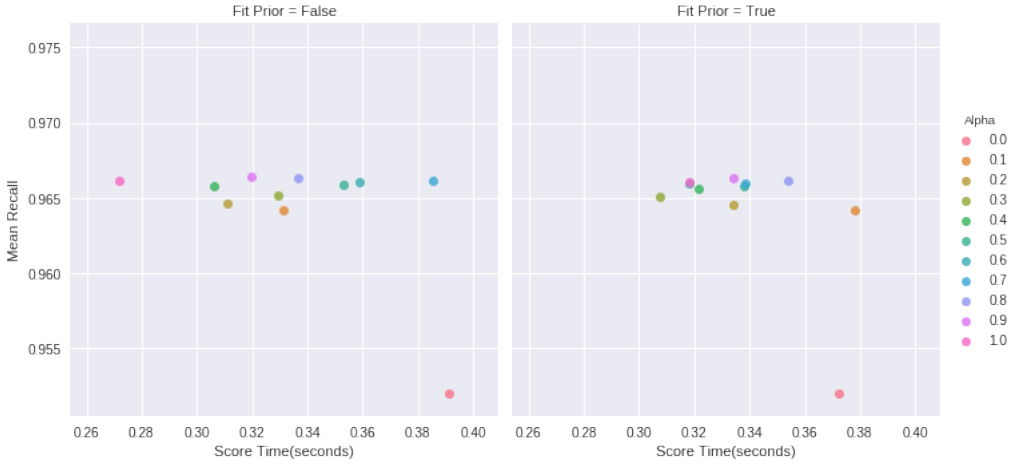

Figure 4.13: Humanitarian Relevancy Classifier MNB: Mean Recall to Score Time

Figures 4.12 and 4.13 indicate that the hyperparameters of the algorithm do not affect the performance of the model to a significant degree. The hyperparameter alpha affects the performance of the model to some extend which is even more visible in Figure 4.14, especially when it takes the value 0.0. Figure 4.14 presents the explicit impact of the value of the hyperparameter alpha to the performance of the model.

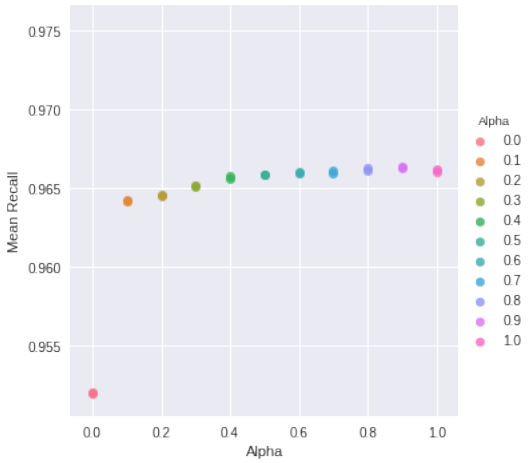

Figure 4.14: Humanitarian Relevancy Classifier MNB: Mean Recall to Score Time

4.5.3 Model Comparison

By comparing the results from the two classification algorithms available, the MNB classifier has outperformed the Decision Tree Classifier both in terms of performance but also of minimising the operational time as it is significantly faster.

In contrast to the findings of the Rumour Detector that there were trade-offs between model performance and model operational times, in the case the Humanitarian Relevancy Classifier the experiments show that the MNB Classifier is the most beneficial solution.

As mentioned in Section 2.3.2 and Section 3.3.3, classifying tweets according to their content is a thoroughly researched field. This did not leave much room for contribution in the field. This Master Thesis project focused on structuring a dataset that incorporated a wide range of humanitarian events in order to develop an efficient classifier.

The logical behind the structure of the dataset that trained tha Humanitarian Relevancy Classifier, was validated as the classification module scored 96.6% in Recall with very fast operational times also. The technique that was chosen is not useful when it comes to acquiring a deeper understanding of which terms

in a tweet define its topic.

4.6 Validation & Performance Analysis Conclusions

In this Section the most important insights derived from the training, testing, validation and analysis of the model are presented. This happened in order to gather all the knowledge acquired from this project and draw conclusions regarding the performance and the utility of the model and the tool developed in this Master Thesis project.

The conceptual model that was designed in Chapter 3 was validated in Chapter 4. The performance of the model showed that this model can assist in the solution of the problem that was defined in Section 1.3. This can be supported by the performance of the model as the classification modules, Rumour Detector and Humanitarian Relevancy Classifier, scored 75.8% and 96.6% in Recall respectively. Furthermore, the operational times of the model showed that it can handle the volume of information that is created during a humanitarian crisis.

Regarding rumour detection, the findings of this Master Thesis project agree with the papers that were reviewed in Section 2.3.1 as the usage of all the available features was the features set that performed best. Moreover, User Features(72.9%) and the Linguistic features(69.8%) were the individual sets that performed best, which comes close to the findings of the papers reviewed in Section 2.3.

According to the analysis that took place in Section 4.4.2, the detection of a rumour is highly related to the continuous activity of the user, the retweeting of the tweet at question, the structure of the tweet, the progress of the network of the user and whether the text of the tweet indicate that the user was an eyewitness of the event. This can provide with great insight not only in the inner workings of the model but also how one can predict whether a tweet constitutes a rumour or not.

In topic classification when applying text analysis, the performance of the model is related to the structure and consistency of the dataset. Using an anthology of events and in specific ratios showed validity as the Humanitarian Relevancy Classifier performed at 96.6%. This would require further investigation but the results of this Master Thesis project on the topic are more than promising.

Chapter 5

Discussion & Future Research

5.1 Conclusions & Limitations

5.1.1 Conclusions

Accurate and timely information is very important for humanitarian activities. Social media incorporate a vast amount of information. Till this point this pool of information has remained unexploited. With the current detection and verification methods, the volume of information produced in social media cannot be handled. Thus an automatic detection method is needed to detect text-based rumours in Twitter relevant to humanitarian activities.

From the technologies that have been researched in the humanitarian context the only technology that provides a solution to the problem mentioned is Supervised Machine Learning. An in-depth literature review showed that rumour detection using Supervised Machine Learning has not been thoroughly researched as a subject not only in the humanitarian sector but in general. On the other hand classifying a tweet according to its content has been researched to a great extend.

The research sub-questions that were formulated in Section 2.5.2 are presented and answered below. Answering the sub-question results in answering the principal research question formulated in Section 2.5.1 and by extension a presentation of the conclusions of this Master Thesis project.

Which methods or techniques can be used to design a model to detect text-based rumours using Supervised Machine Learning?

A conceptual model was developed in this Master Thesis project that used techniques and methods that were reviewed in Section 2.3. These techniques were the creation of a digital DNA of the user and the tweet at question regarding rumour detection and the utilisation of TF-IDF regarding topic classification.

Creating a digital DNA of the user and the tweet at question incorporated features that were created by information extracted either directly from Twitter API or from this information after being analysed.

The information extracted was referring to account properties, such as account age or followers of the account, profile completion, such as if the profile of the user has a description or not, account's permissions, such as if the account has enabled the translation of his posts or not, the activity of the user, such as the number of posts of the user and information propagation, such as the number of retweets.

The analyses that were implemented in the information extracted from Twitter API involve an analysis of the structure of the tweet, such as how many words and letters are included in the text of the tweet posted, sentiment analysis, such as polarity or subjectivity, twitter features detection, such as hashtags and urls, name entity recognition, such as places and explicit analyses of increased weight. The features that were created from the explicit analyses were witness and opinion_shaper. Both features were binary features and assumed values if specific words were included in the text of the tweet at question.

Regarding topic classification, TF-IDF was used in the researches that were reviewed in Section 2.3.2. TF-IDF was commented on its accuracy performance and its compatibility in being used with Bayesian probabilistic classification classifiers. These comments were validated by this research project too.

What is the performance of the Supervised Machine Learning Classifier that is used to detect whether a tweet is a rumour or not?

Rumour Detector scored 75.8% in Recall when using the SVM classification algorithm and 73.2% in Recall when using the Decision Tree classification algorithm. These scores are achieved when all the available features are used. This means that Rumour Detector can retrieve 75.8% of the tweets that are rumours out of the sample that is classified to rumours and non-rumours.

Regarding the required operational times, Rumour Detector classifies 16.800 tweets in 0.033 seconds when using the Decision Tree classifier while it classifies 16.800 tweets in 314 seconds using the SVM classification algorithm. This means that when using the Decision Tree classifier, Rumour Detector can almost instantly classify 16.800 tweets to rumours and non-rumours while scoring 73.2% in Recall. As 350.000 tweets are approximately posted in Twitter per minute, Rumour Detector has the capability of classifying this volume of information, thus reviewing all the available Twitter feed in real time when using the Decision Tree classification algorithm.

This creates a trade-off between the accuracy performance of the Rumour Detector and the time requirements of the classifier. Even though the SVM classification algorithm outperforms the Decision Tree classification algorithm, the operational times of the SVM classification can not be compared to the almost instantaneous classification of the Decision Tree classification algorithm.

What is the performance of the Supervised Machine Learning Classifier that is used to classify whether a tweet is relevant to humanitarian activities or not?

The Humanitarian Relevancy Classifier scored 96.6% in Recall when using

the MNB classification algorithm and 84% when using the Decision Tree classification algorithm. This means that the Humanitarian Relevancy Classifier can retrieve 96.6% of the tweets that are relevant to humanitarian activities out of the samples that is classified. These high performance percentages indicate that the Humanitarian Relevancy Classifier can detect the tweets that are relevant to humanitarian activities especially when taking into consideration that in Accuracy it scored 95% when using the MNB classification algorithm and 88% when using the Decision Tree classification algorithm.

Regarding the required operation times, the Humanitarian Relevancy Classifier can classify 56.000 tweets in 0.85 seconds when using the Decision Tree classification algorithm and 0.314 seconds when using the MNB classification algorithm. As in the case of the Rumour Detector, the Humanitarian Relevancy Classifier can handle the volume of information posted in Twitter. This means that the Humanitarian Relevancy Classifier has the capability of detecting the tweets relevant to humanitarian activities out of the tweets posted in Twitter in real-time with high performance percentages regarding Recall and Accuracy.

Which are the most effective features for the Supervised Machine Learning Classifier that detects whether a tweet is a rumour or not?

The analysis of the results of the performance of the model provided with critical insights regarding the inner workings of the model but more importantly what properties of the user and the content of the tweet at question indicate that a tweet constitutes a rumour. The User Features and the Linguistic Features are the most defining features sets in this classification. More specifically, the continuous activity of the account, the retweeting of the tweet, the structure of the tweet and whether the content of the tweet indicates that the user was an eye-witness of the event are the major dimensions that affect the decision making of this classification.

This insight can provide with valuable information in formulating data driven decision making in the manual review of a tweet to decide whether it is a rumour or not. This way volunteers can have guidelines that are based in historic data. This data can be acquired from similar analyses as the one that took place in Section 4.4.2, which highlights the most crucial and influencing quantitative dimensions of a tweet.

How can Supervised Machine Learning be used to detect text-based rumours relevant to humanitarian activities in Twitter

The conclusion that can be drawn from all this information is that Supervised Machine Learning can be used in various ways in the detection of text-based rumours in Twitter and by extension in the decision making of humanitarian activities. It can be used in order to develop a model that can classify tweets to humanitarian rumours either as a definite decision or as a preliminary tool that can help significantly decrease the volume of tweets that have to be manually reviewed. Moreover, the analysis of such a model can provide critical insight on what constitutes a rumour or not in order to decrease the time of the manual review of a tweet.

The accuracy, the speed and the autonomy of the model that was developed in this Master Thesis project indicate that Supervised Machine Learning is a

technology that should be utilised in humanitarian crisis management. Supervised Machine Learning can be used as a tool to detect humanitarian rumours posted in Twitter or as a technology for a deeper understanding of what constitutes a tweet to be rumour or related to humanitarian activities. Similar analyses to the one carried out in Section 4.4.2 could be utilised to develop more data-driven policies in manual information retrieval process or decision making process.

Apart from the model developed in this Master Thesis project used as a whole, the individual modules can also be used by humanitarian agencies in order to handle social media information.

The Preprocessing Module can be used to extract and formulate quantitative information regarding a tweet so that the manual review of a tweet would be more spherical and the reviewer can have a more holistic image of the tweet at question instantly. Especially when compared to the triangulation process that is time-consuming and limited by the skills of the reviewer, who could be a volunteer, meaning that he might not be a specialist in reviewing manually a tweet. This can empower volunteers to help in more critical processes reducing the burden of agents so that they can act in other operations.

The Rumour Detector on its own can be used to detect whether a tweet constitutes a rumour or not. This can be applied outside the context of humanitarian activities offering a wider field of application for the Rumour Detector.

The Humanitarian Relevancy Classifier can be used in general in the information retrieval process of humanitarian agencies. This module can scan the available Twitter feed and retrieve the tweets that are relevant to humanitarian activities. Additionally, if using a different technique than TF-IDF, a similar analysis as the one that took place in Section 4.4.2 can provide critical insight regarding which words indicate high relevance to humanitarian activities. This can be used to develop data-driven guidelines regarding the manual review of tweets or any kind of text-based information regarding its relevance to humanitarian activities.

Taking into account all the findings and conclusions that are drawn in this Master Thesis project Supervised Machine Learning can be used in two ways in detecting text-based rumours relevant to humanitarian activities in Twitter. It can be used as a detection or filtering tool of the available Twitter feed or as an analysis tool that can provide critical insight on what constitutes a rumour and which words add humanitarian relevancy to a piece of text.

EPA Relevance

This Master Thesis project focused on a grand challenge that is faced, which is the detection of information in the humanitarian sector.

This project analysed the current situation and pinpointed the obstacles that are faced regarding the incorporation of social media information in the information retrieval process in the humanitarian sector. This project incorporated the perspective of the main stakeholders in information systems, analyse the situation and develop a model to either assist in the decision making process of

humanitarian activities or inform on the drivers of what constitutes a rumour.

Supervised Machine Learning was explored as a solution to the problem that was defined in Section 1.3 and how this technology can be used in the humanitarian context. This modelling approach enveloped the capabilities of the technology of Machine Learning with the stakeholders perspective also in consideration.

5.1.2 Limitations

In this section the limitations of this Master Thesis project are presented. The limitations can be about the Preprocessing module, the Rumour Detector, the Humanitarian Relevancy Classifier or for both of the classification modules.

The limitation of the Preprocessing Module is that it can process static data and not real-time data. This means that it can not be part of a process that streams live data. This inability to process live feed constitutes one of the most significant limitations of this Master Thesis project.

A limitation of the Rumour Detector is the fact that even though feature selection was initialised by calculating the importance of the features used, it was not completed. Feature selection could greatly benefit the performance of the model.

The Humanitarian Relevancy Classifier classifies tweets to tweets which are relevant to humanitarian activities and tweets that are not relevant to humanitarian activities. As expressed in the previous sentence, this leads to two major limitations. The first limitation is that the classifier can not differentiate for which phase of the humanitarian crisis a tweet refers to. The second limitation is that the classifier can not differentiate between specific natural or man-made humanitarian disasters. Another limitation of this classifier is a result of the technique that was used. Even though TF-IDF is a widely known and used technique in topic classification, it can not provide insight regarding the inner workings of this model.

Apart from the limitations that were mentioned in the two previous paragraphs, the Rumour Detector and the Humanitarian Relevancy Classifier have two more common limitations. The first limitation is that even though the literature review in Section 2.3 offered many options regarding the algorithms that could carry out the classifications only three of them were explored in the experiments of this Master Thesis project. The second limitation is that the classifiers have not been tested by a multi-dimensional dataset that could test the combined performance of both classifiers.

5.2 Discussion

In this Section, a discussion about what are the implications of the developed model implementation in this Master Thesis project as well as the implementation of the Machine Learning technology in the humanitarian sector is carried out. The implications are divided into the ones that are going to affect the

information retrieval process of humanitarian crisis management and the ones that affect the humanitarian agencies in general.

Regarding the implications that affect the information retrieval process, the most important fact is that Machine Learning given its automation, speed and scalability can handle the input of social media. This means that humanitarian organisations can now exploit data extracted from Twitter information which are relevant to their activities. This broadens the capabilities of humanitarian agencies to acquire information during a humanitarian crisis. Apart from the increased volume of available information, this will also boost the confidence and the awareness of the affected population as discussed in Section 1.1.2.

Limitations of the Humanitarian Rumour Detector could be dealt with the pairing of the tool with other technologies that are utilised in the humanitarian sector. The limitation of the Humanitarian Rumour Detector to detect specific events and disasters or the mistakes that might be included in the output of the model could be corrected by the pairing of the tool with a crowdsourcing platform. This way, instead of information that has been discovered by humanitarian agents and then published to a crowdsourcing platform, the output of the Humanitarian Rumour Detector can be posted in a crowdsourcing platform and evaluated by the affected population.

Additionally, the output of the Humanitarian Rumour Detector can be used as input in a Unsupervised Machine Learning model that solves a clustering problem. This way, the output of the Humanitarian Rumour Detector can be separated and only the tweets of use to explicit events can be extracted.

The sorted output, either in the case of the Humanitarian Rumour Detector being paired with a crowdsourcing platform or an Unsupervised Machine Learning model, can be plotted in an online platform or a web application. This will empower affected populations as people can be informed in real-time regarding the on-going situation. Raising awareness will not only increase the safety of the affected population but also motivate people to join aid operations as a more clear picture of the situation will be depicted.

A computational approach in exploiting the information available in social media will deal with privacy issues too. In comparison to manual processes contacted by individuals, a computational approach can protect the personal information of people that post in social media. A computational approach can be designed in order to be impersonal and keep no information regarding individual action. Moreover, Twitter API provides information that the user has agreed to be shared publicly. Therefore, if not completely solving all privacy issues, computational approaches provide ways to deal with such issues.

The utilisation of the Humanitarian Rumour Detector or any Machine Learning model will not only increase the operational effectiveness of humanitarian agencies but also introduce them to ITs that are avoided at the moment, due to under-stuffed IT departments and prejudice. The utilisation of such tools will make the transition from processes that are purely manual to processes that incorporate computational approaches smoother, thus dealing with the prejudice issues.

The most important and complex implication of the implementation of such

a tool would be the under-stuffed IT departments in humanitarian agencies. In order for humanitarian agencies to be able to implement, maintain and function such a tool their IT departments have to be strengthened. This requires the rearrangement of the structure of the humanitarian agencies as IT department would have a more influencing role in comparison to the on-going situation.

Another dimension that will be affected will be the way that aid operations are contacted. As the information retrieval process will be drastically altered with the implementation of such a tool, this will affect the way that aid operations are conducted. The budget allocation and the volunteers allocation will be altered. A bigger budget will be required in the retrieval of information not only regarding the human resources of the IT department but also regarding its infrastructure. On the other hand, more volunteers will be available to take part in other processes such as logistics.

In my opinion, the humanitarian sector does not use ITs in their full potential at the moment. The private sector has embraced the power of information over the last years and social media information is used in applications that their usefulness is much less apparent than in the case of humanitarian disasters such as the performance of a stock based on Twitter posts. The humanitarian organisations should come up to speed with current technological breakthroughs and unlock the power of information that resides in social media.

This transition is not going to be easy as it requires changing structures and mechanisms that have been used for extended periods of time. And although social media have positively contributed in connecting people around the world, a failure to utilise social media information will eventually come to the expense of the very people humanitarian agencies aspire to help.

5.3 Future Research

The aim of future research would be to tackle the limitations of this project.

The additional investigation of techniques and algorithms would be required. For example, Random Forest Classifier could be explored as an option as it was explored in most articles reviewed in Section 2.3 and the fact that Decision Tree Classifier showed positive results supports this investigation even more.

In the case of the Rumour Detector, the feature selection could have been completed. This would have boosted Rumour Detector not only regarding its performance but also its operation time. Additionally, it would improve the generalisation of the classifier.

In the case of Humanitarian Relevancy Classifier, other techniques than TF-IDF could be explored. For example, Bag of Words would be one of the technique that could be investigated as it also provides with insight regarding the most influential words in the dataset.

Furthermore, Humanitarian Relevancy Classifier could be designed and trained in order to make more explicit classifications. Depending on the time, resources and which dimension is more important, a future project could explore how the classifier could conduct multi-class classification. This means that it would

differentiate between the phases of humanitarian crisis management or the type of humanitarian disaster that is at hand.

A future step of this research could be to either acquire or create multi-dimensional data in order to test the model designed in this Master Thesis project as a whole. This would provide a more holistic image of the performance of the model and provide a more concrete validation of the model.

Last but not least, having acquired all this knowledge and experience on the subject, the research could be developed in a dynamic predictor that would be able to detect text-based rumours in Twitter for humanitarian activities in real-time.

Appendix A

Table A.1 presents the User Features that were created for this Master Thesis project.

#	Name	Description	Type
1	profile_age	This feature indicates how old is the account of the user that has posted the tweet.	numerical
2	user_description	This feature indicates the user has added a description to his profile or not.	binary
3	geo_enabled	This feature indicates if the user has enabled the geographical tracking of Twitter or not.	binary
4	extended_profile	This feature indicates if the user has activated an extended profile account in Twitter or not.	binary
5	location_enabled	This feature indicates if the user has authorized Twitter to store his location when posting a tweet.	binary
6	statuses_count	This feature indicates how many statuses were posted by the user that posted the tweet.	numerical
7	user_timezone	This feature indicates if the user has authorized Twitter to store his timezone or not.	binary
8	verification	This feature indicates if the user has verified his Twitter account or not.	binary
9	translation	This feature indicates if the user has enabled Twitter's translator to translate his tweets or not.	binary

#	Name	Description	Type
10	lists	This feature indicates the number of lists that the user has subscribed to.	numerical
11	protection	This feature indicates if the user has provided Twitter enough credential so that his account is categorised as protected or not.	binary
12	notifications	This feature indicates if the user has enabled the notifications from Twitter or not.	binary
13	default_profile	This feature indicates if the user is using the default Twitter profile picture or not.	binary
14	face_in_picture	This feature indicates if there is a discernible face in the profile picture of the user or not.	binary
15	name_sex	This feature indicates if the user's name in Twitter shows his sex or not.	binary
16	opinion_shaper	This feature indicates if the user given his screen name in Twitter is an opinion-shaper or not.	binary
17	statuses_per_day	This feature indicates how many statuses were posted by the user per day.	numerical
18	lists_per_day	This feature indicates how many lists were followed by the user per day.	numerical
19	male_name	This feature indicates if the user's name in Twitter indicates that he is male or not.	binary
20	followers	This feature indicates how many accounts are following the user that posted the tweet	numerical
21	following	This feature indicates the number of accounts that the user that posted the tweet is following	numerical
22	followers_ per_day	This feature indicates the number of that the user starts following per day on average.	numerical
23	following_ per_day	This feature indicates the number of accounts that start follow the user per day on average.	numerical
24	followers_ following_ratio	This feature indicates ratio between the followers of the user and the accounts the user follows.	numerical

Table A.1: User Features

Table A.2 presents the Linguistic Features that were created for this Master Thesis project.

#	Name	Description	Type
1	characters	This feature indicates the number of characters that are used in this tweet.	numerical
2	exclamation	This feature indicates the number of exclamation marks that are used in this tweet.	numerical
3	question	This feature indicates the number of question marks that are used in this tweet.	numerical
4	commas	This feature indicates the number of commas that are used in this tweet.	numerical
5	periods	This feature indicates the number of periods that are used in this tweet.	numerical
6	multiple_marks	This feature indicates if the tweet contains more than one dictation mark or not.	binary
7	hashtags	This feature indicates the number of hashtags that the tweet contains.	numerical
8	url	This feature indicates the number of urls that the tweet contains.	numerical
9	words	This feature indicates the number of words that this tweet contains.	numerical
10	dictation_rate	This feature indicates the percentage of the tweet that has correct dictation.	numerical
11	subjectivity	This feature indicates how subjective is the writing of the user in the tweet at question.	numerical
12	polarity	This feature indicates how polarised is the writing of the user in the tweet at question.	numerical
13	positive_words	This feature indicates the number of positive words that the tweet contains.	numerical
14	negative_words	This feature indicates the number of negative words that the tweet contains.	numerical
15	uppercase_letters	This feature indicates the number of uppercase case that the tweet contains.	numerical
16	places	This feature indicates the number of places that are mentioned the tweet.	numerical
17	first_person_pn	This feature indicates the number of first person pronouns that are used in the tweet.	numerical
18	second_person_pn	This feature indicates the number of second person pronouns that are used in the tweet.	numerical
19	third_person_pn	This feature indicates the number of third person pronouns that are used in the tweet.	numerical

#	Name	Description	Type
20	witness	This feature indicates if the user that has posted the tweet was a witness or not.	binary
21	language_match	This feature indicates if the language that has been registered in the Twitter account matches the language the tweet was written.	binary
22	mentions	This feature indicates the number of mentions that the tweet contains.	numerical

Table A.2: Linguistic Features

Table A.3 presents the Meta-content Features that were created in this Master Thesis project.

#	Name	Description	Type
1	favourite_count	This feature indicates how many accounts have favoured the posted tweet	numerical
2	retweet_count	This feature indicates how many accounts have retweeted the posted tweet.	numerical
3	tweet_age	This feature indicates how many seconds have passed since the tweet has been posted	numerical
4	favourite_ per_second	This feature indicates the number of favourite of the tweet at question per second.	numerical
5	retweet_ per_second	This feature indicates the number of retweet of the tweet at question per second.	numerical
6	retweet_ per_follower	This feature indicates the ratio of the retweets of the tweet given the followers of the account.	numerical
7	favourite_ per_follower	This feature indicates the ratio of the favourites of the tweet given the followers of the account.	numerical

Table A.3: Meta-content Features

Appendix B

Table B.1 presents which of the features that were used in this Master Thesis project are original products of this research, adjusted features from literature or existing features from published articles. The column reference refers to the article that the adjusted and existing features originate from.

#	Feature	Original	Adjusted	Existing	Reference
1	profile_age	✗	✗	✓	Castillo et al. (2013)
2	user_description	✗	✓	✗	Castillo et al. (2013)
3	geo_enabled	✗	✗	✓	X. Liu et al. (2016)
4	extended_profile	✓	✗	✗	-
5	location_enabled	✓	✗	✗	-
6	statuses_count	✗	✗	✓	Castillo et al. (2013)
7	user_timezone	✓	✗	✗	-
8	verification	✗	✗	✓	Castillo et al. (2013)
9	translation	✓	✗	✗	-
10	lists	✓	✗	✗	-
11	protection	✓	✗	✗	-
12	notifications	✓	✗	✗	-
13	default_profile	✓	✗	✗	-
14	face_in_picture	✓	✗	✗	-
15	name_sex	✓	✗	✗	-
16	opinion_shaper	✗	✗	✓	Andrews et al. (2016)
17	statuses_per_day	✓	✗	✗	-
18	lists_per_day	✓	✗	✗	-
19	male_name	✓	✗	✗	-
20	followers	✗	✗	✓	Castillo et al. (2013)
21	following	✗	✗	✓	Castillo et al. (2013)
22	followers_per_day	✓	✗	✗	-
23	following_per_day	✓	✗	✗	-
24	followers_following _ratio	✓	✗	✗	-

#	Feature	Original	Adjusted	Existing	Reference
25	characters	✗	✗	✓	Castillo et al. (2013)
26	exclamation	✗	✗	✓	Castillo et al. (2013)
27	question_mark	✗	✗	✓	Castillo et al. (2013)
28	commas	✓	✗	✗	-
29	periods	✓	✗	✗	-
30	multiple_marks	✗	✗	✓	Castillo et al. (2013)
31	hashtags	✗	✓	✗	Qazvinian et al. (2011)
32	urls	✗	✗	✓	Qazvinian et al. (2011)
33	words	✗	✗	✓	Castillo et al. (2013)
34	dictation_rate	✓	✗	✗	-
35	subjectivity	✓	✗	✗	-
36	polarity	✓	✗	✗	-
37	positive_words	✗	✗	✓	Castillo et al. (2013)
38	negative_words	✗	✗	✓	Castillo et al. (2013)
39	uppercase_letters	✗	✓	✗	Castillo et al. (2013)
40	places	✗	✗	✓	Hamidian et al. (2015)
41	first_person_pn	✗	✓	✗	Castillo et al. (2013)
42	second_person_pn	✗	✓	✗	Castillo et al. (2013)
43	third_person_pn	✗	✓	✗	Castillo et al. (2013)
44	witness	✗	✓	✗	X. Liu et al. (2016)
45	language_match	✓	✗	✗	-
46	mentions	✗	✓	✗	Hamidian et al. (2015)
47	favourite_count	✓	✗	✗	-
48	retweet_count	✗	✓	✗	Hamidian et al. (2015)
49	tweet_age	✗	✗	✓	Hamidian et al. (2015)
50	favourite_per_second	✓	✗	✗	-
51	retweet_per_second	✓	✗	✗	-
52	retweet_per_follower	✓	✗	✗	-
53	favourite_per_follower	✓	✗	✗	-

Table B.1: Rumour Detector Features

Appendix C

Kwon et al. (2013) created a dataset that included 110 events that were discussed in Twitter. The events and the tweets that were referring to these events were afterwards annotated as rumours or non-rumours. This dataset was mainly used in Master Thesis project to train and validate the Rumour Detector. Table C.1 presents the events that were annotated as non-rumours.

#	Description	# Tweets
1	Air France jet mission with 228 people over Atlantic after running into thunderstorms.	61
2	Pilot hailed for 'Hudson miracle': The pilot of an airliner that ditched in New York's Hudson River.	115
3	Amanda Knox to Take Stand in Murder Trial.	2593
4	Body of missing Yale student Annie Le found on campus in medical school lab.	145
5	Barnes and Noble Store Window Features Obama Alongside Monkey Book	39
6	There are protests in Korea as American beef is about to come back on the menu.	54
7	Ben and jerry's breast milk ice is on sale.	372
8	Couple who adopted 12 children shot to death: Byrd and Melanie Billings were found dead.	372
9	Reviews and status updates for comedy drama film, 'Charlie Wilson's War'.	3958
10	Reviews and viewers' emotion for a video meme, 'Christian the Lion'.	3650
11	Rockefeller poser gets up to 5 years for kidnapping.	1079
12	Cristiano Roaldo lucky escape from crash in his Ferrari.	162
13	Dell enters into smartphone market.	767
14	Reviews and status update about a movie, District 9.	9786
15	Information and reviews about a netbook, eee1101ha.	930
16	A famous video meme about an elephant painting.	477
17	Emma Watson posed for Crash Magazine.	98
18	Bank of England expected to cut interest rates to 1.5% or less. The lowest in the 315-year history.	152
19	Late term abortion Dr. George Tiller shot to death at his church.	14.495
20	Information and pictures of a giant coconut crab.	59

#	Description	# Tweets
21	Reviews and reader's emotion for a video meme 'Hamster on a piano'.	1154
22	Prominent black Harvard prof arrested for breaking into his own house (he was locked out.).	225
23	Heath ledger is dead.	3503
24	News and status update about Iran protest videos.	22653
25	David Cameron's special son Ivan died.	131
26	Abductee had two children with captor, authorities say: Jaycee Dugard, now 29, was kidnapped.	1838
27	Shocking news about Jennifer Hudson's mom and brother being shot and killed.	369
28	A celtic star john hartson has a brain cancer.	302
29	Update and emotions for Josef Fritzl's trial case. He's guilty for rape and incest.	4041
30	Reviews and viewers' emotion for a video meme, 'Lock bumping'.	13
31	Information and reviews about a camera, Nikon D300s	2511
32	Obama's Montana To-Do List: Discuss Health Care, Go Fly Fishing.	63
33	Obama Swats a Fly Like a Boss During an Interview and related video.	14992
34	A jury of 18 has been sworn in to hear the O.J. Simpson robbery case.	633
35	Information and reviews about a pocket PC, palm pre.	7345
36	Rio Pluma LLC is Reissuing Recall of Peanut Products. They may be Contaminated with Salmonella.	151
37	Tweets about one episodes of Plaxico Burress for gun shooting.	1329
38	Thomas Beatie, the transgender man who was born a woman, just welcomed his second child to the world.	3473
39	Prince chunk, the world's biggest cat.	53
40	Information and reviews about a mobile device, psp go.	13458
40	Sarah Jessica Parker expecting twins via surrogate.	1891
41	Reviews and status updates for comedy drama film 'Sicko'.	469
42	Tweets about an existing square shaped watermelon.	1872
43	Melbourne Gay couple have twins by Indian surrogate.	36
44	Information and reviews about a notebook, Toughbook 30.	131
45	Nine dead, 50 injured in Turkish Airlines passenger jet crash at Amsterdam.	875
46	Twitter bought Summize.	2849
47	Mother-of-two nursery worker arrested in child porn probe: Vanessa George.	159
48	News about a serious murderer.	10
49	Twitters about a video meme, Video: Western Spaghetti by team PES.	222
50	West Nile Virus Found in Knox County Mosquito Sample.	4961

Table C.1: Non-rumour Tweets in Kwon et al.(2013) dataset

Table C.2 presents the events that were annotated as rumours.

#	Description	# Tweets
1	Alligators live in sewer.	282
2	Asparagus is a cancer cure.	88
3	Barney Frank Snorts Cocaine	25
4	2 persons find the dead body of Bigfoot in Georgia.	505
5	Tweets about a video meme that shows larvae in the breast.	113
6	A rumor about chain mail to avoid curse of Carmen Winstead.	48
7	Use of cell phone at gas stations causes explosions.	20
8	A legendary animal, Chupacabra, is found.	517
9	Viral video and information that you can cook popcorn with mobile phones.	24
10	Dennis Kucinich saw UFO.	42
11	Deodorant can cause cancer.	258
12	There's a diet coke with bacon flavour.	81
13	A word 'dork' means whale's penis.	916
14	A duck's quack doesn't echo, and no one know why.	718
15	Be careful of earwigs, they can get in your ear and burrow through into your brain.	102
16	Emma Watson died in car accident.	103
17	Rumors about Sony Ericsson free laptop.	125
18	There is really big catfish eating human.	304
19	Google will buy Skype.	507
20	Information about the Postcard computer virus.	171
21	Harrison Ford is died.	414
22	Hercules, the world's biggest dog.	22
23	A man was caught putting his blood that HIV positive in bottles.	628
24	Hydrogen peroxide as a cancer treatment.	59
25	Iphone Nano will be launched.	2715
26	Iphone with OLED display will be launched.	1439
27	Ipod with 64gb capacity will be soon launched.	315
28	Jamie Lee Curtis is a hermaphrodite.	315
29	Jeff Golblum died.	77
30	Kayne West said that 'no one can match his sales' and he is the 'new King of Pop'.	116
31	Tweets mentioning Korean fan death rumors.	1006
32	Lady Gaga is a hermaphrodite.	4820
33	Listerine can deter mosquitoes.	240
34	Killer Catfish Goonch eats Human Flesh.	112
35	John McCain is the Manchurian Candidate	95
36	Megan Fox was originally a guy that had a sex change.	51
37	A viral video and information that combination of coke and mentos can be possibly fatal.	61
38	Information about IT product, midget PC.	17
39	Miley Cyrus is died.	142
40	Urban legend of a monster.	2771

#	Description	# Tweets
41	Mountain dew is not good for sperm	123
42	Remember cell phone's go public next month 888-382-1222 to avoid cell phone telemarketers.	846
43	Obama is muslim and anti christ.	3971
44	Obama cancelling National Day of Prayer.	56
45	Obama is not a natural born citizen.	2157
46	"The top 1% income earners had a pretty good run of it. They will pay higher taxes under Obama."	121
47	Onion can charge Ipod.	579
48	Pepsi can filled with AIDS blood.	41
49	Entering your PIN number backwards at an ATM summons the police.	129
50	Sarah Palin thinks dinosaurs existed 4000 years ago.	233
51	Sarah Palin called Obama and Hillary Sambo and the bitch.	14
52	Rat urine in soda can.	19
53	Sperm bank where they have hot nurses who pay you to give you a hand job.	14
54	Information about Steorn's free energy machine.	129
55	Picture of maggots in brain because of eating raw fish.	15
56	Swine flu can be propagated by pork.	25974
57	Zombie is becoming fact with the swine flue outbreak.	4807
58	The tooth fairy teaches children that they can sell body parts for money.	342
59	Xbox720 will be launched.	1012
60	Zunephone will be launched.	604

Table C.2: Rumour Tweets in Kwon et al.(2013) dataset

Bibliography

Aboulafia, Mitchell (1991). *Philosophy, social theory, and the thought of George Herbert Mead*. SUNY Press.

Albrecht, Allan J. and John E Gaffney (1983). "Software function, source lines of code, and development effort prediction: a software science validation". In: *IEEE transactions on software engineering* 6, pp. 639–648.

Alpaydin, Ethem (2014). *Introduction to machine learning*. MIT press.

Andrews, Cynthia et al. (2016). "Keeping up with the tweet-dashians: The impact of'official'accounts on online rumoring". In: *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. ACM, pp. 452–465.

Anson, Susan et al. (2017). "Analysing social media data for disaster preparedness: Understanding the opportunities and barriers faced by humanitarian actors". In: *International Journal of Disaster Risk Reduction* 21, pp. 131–139. URL: https://www.sciencedirect.com/science/article/pii/S221242091630156X.

Asiri, Sidath (2018). *Machine Learning Classifiers*. URL: https://towardsdatascience.com/machine-learning-classifiers-a5cc4e1b0623/.

Basu, Moumita, Somprakash Bandyopadhyay, and Saptarshi Ghosh (2016). "Post disaster situation awareness and decision support through interactive crowdsourcing". In: *Procedia Engineering*. Vol. 159. Elsevier, pp. 167–173. URL: https://www.sciencedirect.com/science/article/pii/S1877705816322998.

Bechhofer, Robert G (1995). *Design and analysis of experiment for statistical selection, screening, and multiple comparisons*. 04; QA279, B4.

Bergstra, James and Yoshua Bengio (2012). "Random search for hyper-parameter optimization". In: *Journal of Machine Learning Research* 13.Feb, pp. 281–305.

Bishop, Christopher M et al. (2006). "Pattern recognition and machine learning (information science and statistics)". In:

Braga-Neto, Ulisses M and Edward R Dougherty (2004). "Is cross-validation valid for small-sample microarray classification?" In: *Bioinformatics* 20.3, pp. 374–380.

Brownlee, Jason (2018a). *A Gentle Introduction to k-fold Cross-Validation*. URL: https://machinelearningmastery.com/k-fold-cross-validation/.

Brownlee, Jason (2018b). *Supervised and Unsupervised Machine Learning Algorithms*. URL: https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/.

Callaghan, Christian William (2016). "Disaster management, crowdsourced R&D and probabilistic innovation theory: Toward real time disaster response capability". In: *International Journal of Disaster Risk Reduction* 17, pp. 238–250. URL: https://www.sciencedirect.com/science/article/pii/S2212420915300698.

Caplow, Theodore (1947). "Rumors in war". In: *Social Forces*, pp. 298–302.

Carley, Kathleen M. et al. (2016). "Crowd sourcing disaster management: The complex nature of Twitter usage in Padang Indonesia". In: *Safety Science* 90, pp. 48–61.

Castillo, Carlos, Marcelo Mendoza, and Barbara Poblete (2011). "Information credibility on twitter". In: *Proceedings of the 20th international conference on World wide web*. ACM, pp. 675–684.

— (2013). "Predicting information credibility in time-sensitive social media". In: *Internet Research* 23.5. Ed. by Daniel Gayo-Avello, Panagiotis Takis Metax, pp. 560–588. URL: http://www.emeraldinsight.com/doi/10.1108/IntR-05-2012-0095.

Christopher, D Manning, Raghavan Prabhakar, and Schutza Hinrich (2008). "Introduction to information retrieval". In: *An Introduction To Information Retrieval* 151.177, p. 5.

Claesen, Marc and Bart De Moor (2015). "Hyperparameter search in machine learning". In: *arXiv preprint arXiv:1502.02127*.

Conrado, Silvia Planella et al. (2016). "Managing social media uncertainty to support the decision making process during Emergencies". In: *Journal of Decision Systems* 25.sup1, pp. 171–181. URL: https://doi.org/10.1080/12460125.2016.1187396.

Cortes, Corinna and Vladimir Vapnik (1995). "Support-vector networks". In: *Machine learning* 20.3, pp. 273–297.

Crowley, David N, Maciej Dabrowski, and John G Breslin (2013). "Decision support using linked, social, and sensor data". In:

Daume, Stefan, Matthias Albert, and Klaus von Gadow (2014). "Forest monitoring and social media – Complementary data sources for ecosystem surveillance?" In: *Forest Ecology and Management* 316. Forest Observational Studies: "Data Sources for Analysing Forest Structure and Dynamics", pp. 9–20. URL: http://www.sciencedirect.com/science/article/pii/S037811271300618X.

Diakopoulos, Nicholas, Munmun De Choudhury, and Mor Naaman (2012). "Finding and assessing social media information sources in the context of journalism". In: *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems - CHI '12*. CHI '12. New York, NY, USA: ACM, p. 2451. URL: http://doi.acm.org/10.1145/2207676.2208409%20http://dl.acm.org/citation.cfm?doid=2207676.2208409.

DiFonzo, Nicholas and Prashant Bordia (2007). *Rumor psychology: Social and organizational approaches*. Vol. 1. American Psychological Association Washington, DC.

Duan, Kaibo, S Sathiya Keerthi, and Aun Neow Poo (2003). "Evaluation of simple performance measures for tuning SVM hyperparameters". In: *Neurocomputing* 51, pp. 41–59.

Fleming, Candace C, Barbara Von Halle, et al. (1989). *Handbook of relational database design*. Vol. 989. Addison-Wesley New York.

Fottrell, E. and P. Byass (2009). "Identifying humanitarian crises in population surveillance field sites: Simple procedures and ethical imperatives". In: *Public Health* 123.2, pp. 151–155. URL: https://www.sciencedirect.com/science/article/pii/S0033350608002850?via%7B%5C%%7D3Dihub.

Fraser, Evan DG et al. (2006). "Bottom up and top down: Analysis of participatory processes for sustainability indicator identification as a pathway to community empowerment and sustainable environmental management". In: *Journal of environmental management* 78.2, pp. 114–127.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani (2001). *The elements of statistical learning*. Vol. 1. 10. Springer series in statistics New York, NY, USA:

Gao, Huiji, Geoffrey Barbier, and Rebecca Goolsby (2011). "Harnessing the crowdsourcing power of social media for disaster relief". In: *IEEE Intelligent Systems* 26.3, pp. 10–14. URL: http://dx.doi.org/10.1109/MIS.2011.52.

Geitgey, Adam (2018). *Face Recognition*. https://github.com/ageitgey/face_recognition.

Gen, Mitsuo and Runwei Cheng (2000). *Genetic algorithms and engineering optimization*. Vol. 7. John Wiley & Sons.

Goolsby, Rebecca (2010). "Social media as crisis platform". In: *ACM Transactions on Intelligent Systems and Technology* 1.1, pp. 1–11. URL: http://doi.acm.org/10.1145/1858948.1858955%20http://dl.acm.org/citation.cfm?doid=1858948.1858955.

Granell, Carlos and Frank O. Ostermann (2016). "Beyond data collection: Objectives and methods of research using VGI and geo-social media for disaster management". In: *Computers, Environment and Urban Systems* 59, pp. 231–243. URL: https://www.sciencedirect.com/science/article/pii/S0198971516300060.

Gross, Doug (2010). "Survey: More Americans get news from Internet than newspapers or radio". In: *CNN Tech*.

Guyon, Isabelle et al. (2002). "Gene selection for cancer classification using support vector machines". In: *Machine learning* 46.1-3, pp. 389–422.

Hamidian, Sardar and Mona Diab (2015). "Rumor detection and classification for twitter data". In: *Proceedings of the Fifth International Conference on Social Media Technologies, Communication, and Informatics (SOTICS)*, pp. 71–77.

Haworth, Billy (2016). "Emergency management perspectives on volunteered geographic information: Opportunities, challenges and change". In: *Com-*

puters, Environment and Urban Systems 57, pp. 189–198. URL: https://www.sciencedirect.com/science/article/pii/S0198971516300175.

Hermida, Alfred, Seth C. Lewis, and Rodrigo Zamith (2014). "Sourcing the Arab spring: A case study of Andy Carvin's sources on twitter during the Tunisian and Egyptian revolutions". In: *Journal of Computer-Mediated Communication* 19.3, pp. 479–499. arXiv: 0803973233. URL: https://academic.oup.com/jcmc/article/19/3/479-499/4067573.

Hevner, Alan R. et al. (2004). "Design Science in Information Systems Research". In: *MIS Q.* 28.1, pp. 75–105. URL: http://dl.acm.org/citation.cfm?id=2017212.2017217.

Hung, Kuo-Chih, Mohsen Kalantari, and Abbas Rajabifard (2016). "Methods for assessing the credibility of volunteered geographic information in flood response: A case study in Brisbane, Australia". In: *Applied Geography* 68, pp. 37–47. URL: https://www.sciencedirect.com/science/article/pii/S0143622816300054%20http://linkinghub.elsevier.com/retrieve/pii/S0143622816300054.

Imran, Muhammad, Carlos Castillo, et al. (2014). "AIDR: Artificial intelligence for disaster response". In: *Proceedings of the 23rd International Conference on World Wide Web*. ACM, pp. 159–162.

Imran, Muhammad, Shady Elbassuoni, et al. (2013). "Extracting information nuggets from disaster-related messages in social media." In: *Iscram*.

Imran, Muhammad, Prasenjit Mitra, and Carlos Castillo (2016). "Twitter as a Lifeline: Human-annotated Twitter Corpora for NLP of Crisis-related Messages". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Portoroz, Slovenia: European Language Resources Association (ELRA).

Ioffe, Sergey and Christian Szegedy (2015). "Batch normalization: Accelerating deep network training by reducing internal covariate shift". In: *arXiv preprint arXiv:1502.03167*.

James, Gareth et al. (2013). *An introduction to statistical learning*. Vol. 112. Springer.

Kaplan, Andreas M (2012). "If you love something, let it go mobile: Mobile marketing and mobile social media 4x4". In: *Business horizons* 55.2, pp. 129–139.

Kaplan, Andreas M and Michael Haenlein (2010). "Users of the world, unite! The challenges and opportunities of Social Media". In: *Business horizons* 53.1, pp. 59–68.

Kohavi, Ron et al. (1995). "A study of cross-validation and bootstrap for accuracy estimation and model selection". In: *Ijcai*. Vol. 14. 2. Montreal, Canada, pp. 1137–1145.

Kwon, Sejeong et al. (2013). "Prominent features of rumor propagation in online social media". In: *International Conference on Data Mining*. IEEE.

Lee, A (2000). "Systems Thinking, Design Science, and Paradigms: Heeding Three Lessons from the Past to Resolve Three Dilemmas in the Present to Direct a Trajectory for Future Research in the Information Systems

Field,"Keynote Address". In: *Eleventh International Conference on Information Management, Taiwan*.

Lee, Kathy et al. (2011). "Twitter trending topic classification". In: *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*. IEEE, pp. 251–258.

Leskovec, Jure, Anand Rajaraman, and Jeffrey David Ullman (2014). *Mining of massive datasets*. Cambridge university press.

Liu, Huan and Hiroshi Motoda (1998). *Feature extraction, construction and selection: A data mining perspective*. Vol. 453. Springer Science & Business Media.

Liu, Xiaomo, Quanzhi Li, et al. (2016). "Reuters Tracer: A Large Scale System of Detecting & Verifying Real-Time News Events from Twitter". In: *CIKM*, pp. 207–216.

Liu, Xiaomo, Armineh Nourbakhsh, et al. (2015). "Real-time rumor debunking on twitter". In: *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. ACM, pp. 1867–1870.

Loria, Steven et al. (2014). "Textblob: simplified text processing". In: *Secondary TextBlob: Simplified Text Processing*.

Ludwig, Thomas et al. (2017). "Situated crowdsourcing during disasters: Managing the tasks of spontaneous volunteers through public displays". In: *International Journal of Human Computer Studies* 102, pp. 103–121. URL: https://www.sciencedirect.com/science/article/pii/S1071581916301197.

Maresh-Fuehrer, Michelle M. and Richard Smith (2016). "Social media mapping innovations for crisis prevention, response, and evaluation". In: *Computers in Human Behavior* 54, pp. 620–629. arXiv: arXiv:1011.1669v3. URL: https://www.sciencedirect.com/science/article/pii/S0747563215301175.

Markus, M Lynne, Ann Majchrzak, and Les Gasser (2002). "A design theory for systems that support emergent knowledge processes". In: *MIS quarterly*, pp. 179–212.

Martin, Nora (2016). "Information Verification in the Digital Age:The News Library Perspective". In: 2, pp. i–51.

McCallum, Andrew, Kamal Nigam, et al. (1998). "A comparison of event models for naive bayes text classification". In: *AAAI-98 workshop on learning for text categorization*. Vol. 752. 1. Citeseer, pp. 41–48.

Meier, Patrick (2011). *New information technologies and their impact on the humanitarian sector*. URL: http://www.journals.cambridge.org/abstract%7B%5C_%7DS1816383112000318.

Michalski, Ryszard S, Jaime G Carbonell, and Tom M Mitchell (2013). *Machine learning: An artificial intelligence approach*. Springer Science, Business Media.

Middleton, Stuart E, Lee Middleton, and Stefano Modafferi (2014). "Real-time crisis mapping of natural disasters using social media". In: *IEEE Intelligent Systems* 29.2, pp. 9–17.

Norheim-Hagtun, Ida and Patrick Meier (2010). "Crowdsourcing for Crisis Mapping in Haiti". In: *Innovations: Technology, Governance, Globalization* 5,

pp. 81–89. URL: http://www.mitpressjournals.org/doi/pdf/10.1162/INOV%7B%5C_%7Da%7B%5C_%7D00046.

Norvig, Peter (2007). "How to write a spelling corrector". In: *De: http://norvig. com/spell-correct. html*.

O'Brien, John (2011). "MacChat: 2009–The Age of the Twitpocalypse". In: *Tech Blog*.

Özdamar, Linet and Mustafa Alp Ertem (2015). "Models, solutions and enabling technologies in humanitarian logistics". In: *European Journal of Operational Research*. Vol. 244. 1. North-Holland, pp. 55–65. URL: https://www.sciencedirect.com/science/article/pii/S0377221714009539.

Panagiotopoulos, Panos et al. (2016). "Social media in emergency management: Twitter as a tool for communicating risks to the public". In: *Technological Forecasting and Social Change* 111, pp. 86–96. URL: https://www.sciencedirect.com/science/article/pii/S0040162516301196.

Patterson, Scott (2010). "Letting the machines decide". In: *The Wall Street Journal* 13.

Pedregosa, Fabian et al. (2011). "Scikit-learn: Machine learning in Python". In: *Journal of machine learning research* 12.Oct, pp. 2825–2830.

Popoola, Abdulfatai and Dmytro Krasnoshtan (2013). "Information verification during natural disasters". In: *Proceedings of the 22nd international conference on World Wide Web*. WWW '13 Companion, pp. 1029–1032. URL: http://doi.acm.org/10.1145/2487788.2488111%20http://dl.acm.org/citation.cfm?id=2488111.

Pressr, Cambridge University (2018). *Overfitting*. URL: https://en.oxforddictionaries.com/definition/overfitting.

Qazvinian, Vahed et al. (2011). "Rumor has it: Identifying misinformation in microblogs". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 1589–1599.

Rakitin, Steven R (2001). *Software verification and validation for practitioners and managers*. Artech House, Inc.

Rennie, J et al. (2003). "Tackling the poor assumptions of Naive Bayes classifiers (PDF)". In: ICML.

Riccardi, Mark T. (2016). "The power of crowdsourcing in disaster response operations". In: *International Journal of Disaster Risk Reduction* 20, pp. 123–128. URL: https://www.sciencedirect.com/science/article/pii/S2212420916302199.

Rokach, Lior and Oded Z Maimon (2008). *Data mining with decision trees: theory and applications*. Vol. 69. World scientific.

Russell, Stuart J and Peter Norvig (2016). *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited,

Saleh, Babak et al. (2016). "Toward automated discovery of artistic influence". In: *Multimedia Tools and Applications* 75.7, pp. 3565–3591.

Schifferes, Steve et al. (2014). "Identifying and Verifying News through Social Media: Developing a user-centred tool for professional journalists". In: *Digi-*

tal Journalism 2.3, pp. 406–418. URL: http://www.tandfonline.com/doi/abs/10.1080/21670811.2014.892747.

Shibutani, Tamotsu (1966). *Improvised news*. Ardent Media.

Soden, Robert and Leysia Palen (2014). "From Crowdsourced Mapping to Community Mapping: The Post-earthquake Work of OpenStreetMap Haiti". In: *COOP 2014 - Proceedings of the 11th International Conference on the Design of Cooperative Systems, 27-30 May 2014, Nice (France)*. Ed. by Chiara Rossitto et al. Cham: Springer International Publishing, pp. 311–326. URL: http://link.springer.com/10.1007/978-3-319-06498-7%7B%5C_%7D19.

Starbird, Kate et al. (2014). "Rumors, false flags, and digital vigilantes: Misinformation on twitter after the 2013 boston marathon bombing". In: *IConference 2014 Proceedings*.

Statista, The Statistics Portal (2018a). *Most famous social network sites worldwide as of July 2018, ranked by number of active users (in millions)*. URL: https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/ (visited on 08/20/2018).

— (2018b). *Number of social media users worldwide from 2010 to 2021 (in billions)*. URL: https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/ (visited on 07/03/2018).

Stats, Internet Live (2018). *Twitter Usage Statistics*. URL: http://www.internetlivestats.com/twitter-statistics/ (visited on 08/20/2018).

Suh, Yirey et al. (2017). "A Comparison of Oversampling Methods on Imbalanced Topic Classification of Korean News Articles". In: *Journal of Cognitive Science* 18.4, pp. 391–437.

Sump-Crethar, A Nicole (2012). "Making the Most of Twitter". In: *Reference Librarian* 53.4, pp. 349–354. URL: https://doi.org/10.1080/02763877.2012.704566.

Takahashi, Bruno, Edson C. Tandoc, and Christine Carmichael (2015). "Communicating on Twitter during a disaster: An analysis of tweets during Typhoon Haiyan in the Philippines". In: *Computers in Human Behavior* 50, pp. 392–398. URL: https://www.sciencedirect.com/science/article/pii/S0747563215003076.

Tapia, Andrea H et al. (2011). "Seeking the trustworthy tweet: Can microblogged data fit the information needs of disaster response and humanitarian relief organizations". In: pp. 1–10.

Thornton, Chris et al. (2013). "Auto-WEKA: Combined selection and hyperparameter optimization of classification algorithms". In: *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pp. 847–855.

Walls, Joseph G, George R Widmeyer, and Omar A El Sawy (1992). "Building an information system design theory for vigilant EIS". In: *Information systems research* 3.1, pp. 36–59.

Watson, Hayley and Rowena Rodrigues (2017). "Bringing Privacy into the Fold: Considerations for the Use of Social Media in Crisis Management". In: *Jour-*

nal of Contingencies and Crisis Management 26.1, pp. 89–98. URL: http://doi.wiley.com/10.1111/1468-5973.12150.

Wendling, Cécile, Jack Radisch, and Stephane Jacobzone (2013). "The use of social media in risk and crisis communication". In:

Yuan, Faxi and Rui Liu (2018). *Feasibility study of using crowdsourcing to identify critical affected areas for rapid damage assessment: Hurricane Matthew case study.* URL: https://www.sciencedirect.com/science/article/pii/S221242091830150X.

Zimmer, J Christopher et al. (2010). "Investigating online information disclosure: Effects of information relevance, trust and risk". In: *Information & management* 47.2, pp. 115–123.

Zubiaga, Arkaitz et al. (2018). "Detection and resolution of rumours in social media: A survey". In: *ACM Computing Surveys (CSUR)* 51.2, p. 32.