JOHANNES KEPLER
UNIVERSITY LINZ

Submitted by
David Graf, BSc

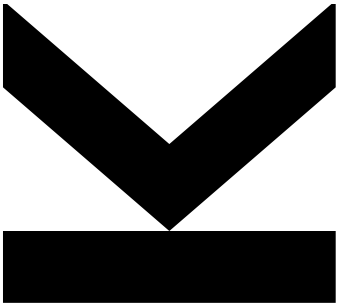
Submitted at
Department of Cooperative Information Systems, Institute of Telecooperation

Supervisor
Assoc. Prof. Mag. Dr. Wieland Schwinger, M.Sc.

Co-Supervisor
a.Univ.-Prof. Mag. Dr. Werner Retschitzegger

August 2018

Cross-domain Informativeness Classification for Disaster Situations



Master Thesis
to obtain the academic degree of
Diplom-Ingenieur
in the Master's Program
Computer Science

JOHANNES KEPLER
UNIVERSITY LINZ
Altenbergerstraße 69
4040 Linz, Österreich
www.jku.at
DVR 0093696

Statutory Declaration

I hereby declare that the thesis submitted is my own unaided work, that I have not used other than the sources indicated, and that all direct and indirect sources are acknowledged as references.
This printed thesis is identical with the electronic version submitted.

Linz, August 2018

David Graf, BSc

Acknowledgements

First of all, I would like to thank my thesis supervisors Prof. Werner Retschitzegger and Prof. Wieland Schwinger for supporting me throughout the process of working on this master thesis. I want to thank for always providing me with very detailed and constructive feedback, which helped me learning a lot about how to work and write in a scientific way. Thank you for getting the opportunity to work on this thesis at the Department of Cooperative Information Systems. As well I want to thank for cooperativeness and organizational support during the last months.

I furthermore want to thank Dr. Andrea Salfinger for the initial meetings at the beginning of my thesis and for providing me a starting point in this very interesting research field.

Finally, I would like to thank my family, my parents, my girlfriend Kathrin for providing me with unfailing support and encouragement throughout my years of study and through the process of working on this master thesis. Thank you!

Kurzfassung

Um im Rahmen des Krisenmanagements im Katastrophenfall möglichst rasch und effizient auf die vorliegende Situation reagieren zu können, gewinnen Social Media Platformen als Daten- und Informationsquelle immer mehr an Bedeutung. Das Filtern von Social Media Nachrichten hinsichtlich informativen und nicht informativen Inhalten kann dabei als ein wesentlicher Schritt verstanden werden. Dieses Filtern von Nachrichten wird häufig realisiert unter Anwendung von überwachtem Lernen (engl. "Supervised Machine Learning"). Aufgrund der Tatsache, dass kein Katastrophenfall mit einem anderen ident ist, beschränken sich in diesem Kontext viele Lösungsansätze auf einzelne Krisenarten, wie beispielsweise Erdbeben oder Flutkatastrophen. Zum Einen sind diese Ansätze nur bedingt geeignet für zukünftige Katastrophen anderer Arten, wie beispielsweise Waldbrände, zum Anderen bieten unterschiedliche Ausgangsdaten, verwendete Features, sowie Lernalgorithmen und deren Konfiguration zahlreiche Realisierungsmöglichkeiten, welche zu unterschiedlicher Qualität der Klassifizierung von Social Media Nachrichten führen. Diese Problemstellungen addressierend, umfasst die vorliegende Masterarbeit zwei Teile: Eine systematische und detaillierte Analyse eines existierenden krisenbezogenen Twitter Datensatzes trägt zu einem besseren Verständnis von Merkmalen dar, welche informative und nicht informative Nachrichten karakterisieren. Darauf aufbauend entwickelt diese Masterarbeit eine krisentyp unabhängige Klassifizierung von Twitter Nachrichten und zeigt abschließend in einer systematischen Evaluierung zahlreicher Experimente, dass diese krisentyp unabhängige Klassifizierung qualitativ bessere Ergebnisse erziehlt als eine Klassifikation bezogen auf spezifische Krisenarten.

Abstract

Social Media services gain increasing importance as a new data source for achieving Situation Awareness in disaster management. One crucial prerequisite is to automatically filter social media messages towards informativeness commonly realized by supervised machine learning. Since disaster situations are different, most classification approaches focus on informativeness classification of similar disasters. Thus their use is limited to particular disaster types, for instance earthquakes or floods, lacking general applicability. At the same time, how to get accurate informativeness classification for new disaster events is not yet totally understood due to variations in training data, features, classification algorithms and their settings. To address these issues, the contribution of this thesis twofold: First, a systematic and in-depth analysis of an existing twitter crisis data set is provided along four different dimensions in order to gain a comprehensive understanding of those characteristics indicating informative Tweets in disaster situations. Second, on basis of these insights, a cross domain classifier is engineered, which is applicable not only across different disaster events but also across disaster events of different types. Systematic classification experiments are conducted, demonstrating that the cross-domain classification approach presented in this thesis is more accurate than other disaster type specific ones.

Contents

List of Abbreviations

AI Artificial Intelligence

ML Machine Learning

SA Situation Awareness

POS Part-of-Speech

PCA Principal Component Analysis

List of Figures

List of Tables

Chapter 1

Introduction

Situation Awareness in disaster management. In crisis situations, such as natural disasters like earthquakes, floods, hurricanes or human-induced disasters like shootings or bombings, it is crucial for organizations and authorities to know the extent of the current situation to be able to react in an appropriate way. In many cases, however, detailed information about what happened exactly and what is going on in the affected area does not exist [Imran et al., 2015]. Counteracting this missing *Situation Awareness (SA)*, i.e., "understanding what is happening" [Vieweg et al., 2010], in a timely manner is crucial to reduce the impact on affected people [Cameron et al., 2012][Girtelschmid et al., 2016][Imran et al., 2015][Salfinger et al., 2016a] [Salfinger et al., 2016b].

Social Media for SA. Due to fast growing numbers of Social Media users in recent years all over the world, organizations and researchers are trying to use the paradigm of "humans as sensors" to get up to date Social Media messages, and "first-hand" information, from people, individuals or organizations who are affected by the actual crisis and further are trying to extract relevant information for efficient disaster management and decision making. "First-hand" [Imran et al., 2015] information by affected people or local organizations is one possible way to gather information about what is going on in the effected environment. Further on, in some situations "first-hand" information is even the only source of information, which can provide additional knowledge about the current situation [Vieweg, 2012]. Several studies showed, that Social Media are a very frequently used communication channel even during crisis situations [Acerbo and Rossi, 2017][Vieweg, 2012], thus providing a new source of data for gaining SA [Cameron et al., 2012]. In order to exploit the full potential of Social Media for enhancing SA in disaster management, first of all, Social media messages have to be automatically filtered with respect to *informativeness* [Salfinger, 2016], eliminating *non-related* messages like spam or advertisements and *non informative* ones like emotions or emphatic expressions. Overall, informativeness classification is the crucial basis for all further processing steps, like damage or impact assessment [Cresci et al., 2015].

Informativeness of tweets. The concept of *informativeness* is diverse in its use and discussed in various areas, including e.g. informativeness of web documents [Horn et al., 2013], term informativeness [Wong and Kit, 2011][Wu and Giles, 2013] or informativeness of Social Media messages in areas like news [Lloret and Palomar, 2016] as well as the crisis domain itself [Olteanu et al., 2015]. Yet, informativeness is a subjective concept, which heavily depends on the receiver of the information [Olteanu et al., 2015]. Since a variety of informativeness definitions exist [Derczynski et al., 2018][Horn et al., 2013][Lloret and Palomar, 2016][Longhini et al., 2017], the current work follows the informativeness definition of Olteanu et al. [Olteanu et al., 2015] where informativeness of crisis related tweets is captured by "checking whether the tweet contributes to a better understanding of the situation on the ground".

1.1 Social Media in disaster situation

Social Media platforms are an "easy-to-use way to communicate and network with each other" [Gundecha and Liu, 2012]. Therefore, millions of people use them every day to share messages for various reasons [Olteanu, 2016]. Studies showed that lots of people use Social Media platforms in crisis situations to communicate and share information about the "situation on the ground" [Imran et al., 2015], for example about areas with power outages, people without food or water [Imran et al., 2015], [Vieweg, 2012], [Olteanu, 2016]. The amount of information shared on Social Media after a disaster event is even more than usually shared by users [Cresci et al., 2015]. However, there are many limitations of using social data [Olteanu, 2016], which have to take into considerations when working with data from Social Media. Social Media data are "vast, noisy, distributed, unstructured, and dynamic" [Gundecha and Liu, 2012]. In addition, information shared online by users depend on their goal [Imran et al., 2015]. In other words, not every piece of information, shared online by users, can contribute to situation awareness or is informative in some way at all. Even worse, information on Social Media platforms can not be verified easily and may be a rumor, i.e. may be not true or consists of wrong facts [Zubiaga et al., 2015]. Yet, if these issues are taken into considerations while using Social Media services, they can contribute to situation awareness in early phases of disaster situations.

1.2 Twitter

A very popular social media platform is Twitter[1], a microblogging service which limits one messages to a maximum of 140 characters. Last year, 2017, Twitter doubles the amount of

[1] www.twitter.com

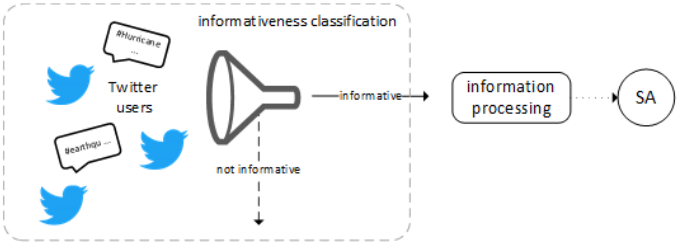


Figure 1.1: Informativeness filtering

allowed characters to 280 per message. A message is called tweet. Twitter is used all over the world to communicate, share and discuss any topics. The network uses the concept of followers and followees. Unlike other social media platfroms like Facebook[2], this concept leads to a directed user network. A user X who follows another user Y does not imply that user Y necessarily follows back user X. A user X who follows user Y receives all tweets posted by user Y. Retweets allow to answer on a previous tweet immediately. Over time Twitter users evolved common practices for communication. For instance, "RT" stands for Retweet, "@" followed by a certain user refers to this particualr user and "#" followed by a word is used for Hashtags [Kwak et al., 2010]. hashtag information is used in tweets to refer tweets to certain topics which are currently discussed. For instance, the hashtag "#hurricanesandy" was used in Twitter communication to state this tweet is related to this particular Hurricane. The reason why Twitter is frequently used for scientific research is because Twitter provides an API to access real tweet data. In general, Twitter provides three types of APIs, namely the Search API, the Streaming API and the Twitter Firehose. While the Search API works like searching based on keywords, the Streaming API and the Firehose are rather a push of data by Twitter, i.e. Twitter delivers tweets in near real-time directly to the API's user. Twitter's Search API and the Streaming API are accessible for free, yet there exist some limitations by Twitter, for instance the amount of accessible data or number of Retweets. In contrast, Twitter Firehose provides full access and less restrictions, however usage is costly.

1.3 Motivation

Past disaster events showed that Social Media users post a huge amount of messages during and after crisis situations [Olteanu et al., 2015]. However, not all messages are useful in the sense of informativeness and can contribute to situation awareness. Some messages contain a lot of emotion and are hardly informative, for instance ad and spam messages or even not related messages which do not provide any crisis related information at all. Therefore, filtering potentially crisis related Social Media messages regarding their informativeness, i.e., a particular message contributes to crisis related situation

[2] www.facebook.com

awareness or in other words is informative or is not informative, is one major task, visualized in Figure 1.1. In an ideal case immediately after a user shares a message, an automated task can decide, whether this particular message is informative or not, i.e., filters out only informative messages to support disaster management and disaster decision making. This is desirable since due to the huge number of Social Media messages in a crisis event, human processing and filtering of those messages is almost impossible. For instance, when analysing the Hurricane "Sandy" event in 2012 and 2013, Stowe et al. [Stowe et al., 2016] got a collection of 22.2 million messages after performing keyword selection on Twitter. Manual processing of such a big amount of data is not feasible. Hence, the state-of-the-art approach for this task is using machine learning to classify Social Media posts in informative and not informative messages automatically [Acerbo and Rossi, 2017][Imran et al., 2016][Khare et al., 2017][Verma et al., 2011]. Recent scientific literature deal with this issue and propose a variety of different approaches to address this information filtering task.

1.4 Machine Learning from Social Media data

Machine Learning types. Machine Learning is a huge field in the domain of artificial intelligence. In general, scientific literature distinguish three main types of machine learning [Russell and Norvig, 2016]. First, *supervised learning*, where an algorithm learns a function for input-output pairs based on previous examples, which is used to predict the output class of new input data. Classification is an example of supervised learning. The second type, *unsupervised learning* tries to uncover patterns in data without explicitly defining output classes. Clustering is the most common unsupervised learning task [Russell and Norvig, 2016]. The third type, *reinforcement learning* is based on "series of reinforcements-rewards or punishments" [Russell and Norvig, 2016] which tell the algorithm whether previous decisions were right or wrong. This master thesis uses concepts of supervised learning or classification respectively, to determine informativeness of potentially crisis related tweets. Furthermore, clustering is used for tasks of the data analysis part.

Classification. Extracting information from Social Media can be considered as a classification problem. A certain message can be informative or not, can contain information about affected people, about the state of infrastructure or affected areas. Due to the huge amount of data generated by users in crisis situation, manually categorization by humans of these messages is not feasible. Hence, categorization of Social Media messages is a classification task done by computer algorithms, so-called classifiers. The classifier is trained on labeled messages of past disaster events. Each message is described by a set of features, which are properties of this particular message. The classifier learns, based on features, how to classify messages, i.e., which properties of a message leads to which result. The

trained classification model then can be used to classify new messages in new crisis events, for instance, whether a message is informative or not, contain information about affected people, about the state of infrastructure or affected areas.

1.5 Research scope

Informativeness classification. Current approaches for informativeness classification mainly employ supervised machine learning [Acerbo and Rossi, 2017][Imran et al., 2016] [Khare et al., 2017][Verma et al., 2011]. Learning from past events and classification on new events, is, however quite challenging not least since disaster situations are different in many ways [Pekar et al., 2016]. For instance, a Hurricane differs from a plan crash regarding affected people, affected geographic location, duration of the event and further leads to different decisions and reactions in disaster management. Thus, most classification approaches focus on informativeness classification for specific types of events, for instance earthquakes or floods, only, lacking general applicability. At the same time, how to get accurate informativeness classification for new crisis events is not yet totally understood due to variations in training data, features, classification algorithms and their settings.

Crisis related data. Some studies already investigated the behaviour of people using Social Media in different crisis situations [Vieweg, 2012]. Furthermore, previous research showed that events also differs significantly regarding the information shared on Social Media [Olteanu et al., 2015]. Since every crisis is unique in general, the question is, are there still commonalities between events which allow learning from past events? What are common characteristics of different crisis events? Where are the differences? How are differences and similarities related to informativeness in messages?

Problem definition. While many research deal with various classification approaches to extract information out of Social Media messages [Imran et al., 2013][Pekar et al., 2016][Stowe et al., 2016][Verma et al., 2011], very few research address the actual characteristics of past events and their contained implicit temporal, spatial, linguistic and source information, in particular related to informativeness. Detailed insights in potentially crisis related data might contribute to a more efficient and intelligent way of classification. Moreover, a majority of classification approaches uses a relatively small corpus, regarding number of crisis events, disaster types as well as number of samples, for training, because manually labeling is a very time consuming and cost intensive task. Hence, only little existing research has addressed cross-domain training, i.e., to train the classifier with samples from more than one event domain [Pekar et al., 2016][Verma et al., 2011][Acerbo and Rossi, 2017]. Due to different event characteristics, training the classifier on one crisis event and further using it for classification on a current event might not be effective. Differences in event dimensions

like type, whether it is an earthquake or a flood, or like temporal aspects, whether it is an instantaneous or progressive event, might lead to insufficient classification results. In other words, learning from an event in the past might not be suitable for classification of future events. The questions is whether this assumption is true for informativness classification also? Since a corpus for each type of event is practically not feasible [Cresci et al., 2015], a classifier, cross-domain trained on a heterogeneous dataset, which achieves sufficient accuracy over various events, is beneficial.

Contributions. This master thesis consists of two main parts. First, a systematic and in-depth analysis of an existing twitter crisis data set is provided along four different dimensions, covering temporal, spatial, linguistic and source information with respect to informativeness in order to gain a comprehensive understanding of those characteristics indicating informative tweets in disaster situations. Additionally, insight in crisis related data uncovers differences and similarities cross crisis events of the same type and cross crisis events of different types to create an appropriate training set for informativeness classification. On basis of these insights, in the second part of this thesis, a cross domain classifier is engineered, which is applicable not only across different disaster events but also across disaster events of different types. Features and training data used for the classification approach are based on data analysis results. Since the time aspect plays an important role in crisis, this thesis proposes a novel time dependent classification approach, which partitions the entire classification in sub- classification tasks with respect of the time when the tweet was shared. For the previous stated reasons, this thesis focuses on cross-domain training, i.e., use more than one event as training corpus, to show how well this informativeness classification approach performs on a heterogeneous training set, tested on various different events of different event types. However, being able to compare cross-domain training results with other approaches used in recent literature this thesis runs a variety of classification experiments using different train-testset combinations, including in-domain and out-domain classification.

Thesis scope. The overall scope of this thesis is informativeness classification. Thus, all statements, data analysis, results and experiments in this thesis are considered regarding informativeness classification, which might be different from other classification tasks in the context of mining crisis related data. Not the primary scope of this thesis is improving informativeness classification by optimizing and developing more complex features like using lexicons and ontologies to represent messages in a more abstract way, since a lot of other research already did and improvements are slight at some point [Khare et al., 2017]. Instead, this work's focus is more on the data itself and how data and knowledge about it can be used to improve informativeness classification of tweets. However, using data insights to develop additionally more complex features might lead to further classification improvements and thus is suited for future work.

1.6 Thesis outline

This master thesis is structured as follows. Chapter 2 gives some background information to get a better understanding of the current research field. The following Chapter 3 discusses and compares recent literature and tries to point out main differences with respect to this thesis. Further, Chapter 4 focuses on the dataset and provides a systematic manual data analysis regarding temporal, spatial, linguistic and source characteristics. Chapter 5 and Chapter 6 propose an informativeness classification approach based on data insights and explain details of the prototype implementation. Chapter 7 discusses and compares various classification experiments and their results. The final Chapter 8 contains a lessons learned sections as well as an outlook for future work.

Chapter 2

Background

The current chapter gives some background information related to the research field of this master thesis. Since a majority of research focusing on Twitter as information source, the following section discusses briefly the usage of Twitter for research. Furthermore, relevant aspects regarding the data set are explained in more detail. In addition, state-of-the-art Machine Learning algorithms used for research are discussed. Finally, this chapter covers some definitions used in the Machine Learning domain to explain meanings of used terms in this thesis.

2.1 Twitter for research

A majority of researchers use Twitter, respectively tweets posted by Social Media users, as a source of information [Vieweg, 2012][Cresci et al., 2015][Olteanu et al., 2015], because Twitter provides a public accessible Application Programming Interface (API)[1], where it is possible to access and stream tweets in near real-time. Hence, research in this domain is rather focused and adapted on characteristics of Twitter, for instance the limited amount of characters used in a message or the concept of Retweets or the directed linking concept of followers and followees. As a consequence, research and their conclusions based on Twitter might not be applicable to other Social Media services directly. At least, differences between Social Media services have to be considered when working with different platforms. In general, information extraction from Social Media for scientific reasearch is not restricted to Twitter, Pogrebnyakov et al. [Pogrebnyakov and Maldonado, 2017] uses Facebook as information source. Nevertheless, getting data from other Social Media platforms than Twitter might be challenging since only a minority of Social Media platforms provide their data to public.

[1] https://developer.twitter.com/en/docs.html

2.2 Available datasets

For this thesis the most relevant work with respect to crisis related datasets is the work of Olteanu et al. [Olteanu et al., 2015]. They accessed the Twitter API by using Hashtag and keyword filtering to collect tweets of 26 crisis events, which take place in the years 2012 and 2013. Furthermore, they employed crowed-source worker to judge each tweet and label them manually regarding informativeness, i.e., whether a certain tweet is informative or not, regarding information type, i.e., categories like for example infrastructure or donations, and regarding information source, i.e., whether a tweet was posted by media organizations, by government or directly by affected people. The outcome of their work is a labeled corpus with around 28K tweets public available under the name CrisisLexT26 dataset[1]. The CrisisLexT26 dataset is described in more detail in section 4, because this thesis build up on the CrisisLexT26 dataset.

Stowe at al. [Stowe et al., 2016] created for their work a crisis related corpus using tweets posted by Twitter users before, during and after Hurricane "Sandy" in the year 2012. Then "domain experts, social scientists, and linguists" [Stowe et al., 2016] labeled them manually which resulted in a corpus of 7.5K tweets. The corpus of [Imran et al., 2013] was manually label by an external crowd-sourcing platform and contains approximately 4.5K tweets from the Tornado in Joplin, Missouri 2011.

Due to the reason that creating a corpus for training and testing is very costly, many researches use already existing and labeled data sets or subsets of them. Pekar et al. [Pekar et al., 2016] uses the entire CrisisLexT26 dataset, Khare et al. [Khare et al., 2017] a subset of the it. In total 3.2K tweets out of 9 events, where the majority of tweets are in English. Imran et al. [Imran et al., 2016] uses 11 events, five earthquakes and six flood events, of the CrisisLexT26 combined with tweets from a platform called AIDR, which results in total in 11.7K tweets. A subset of the CrisisLexT26, 12 events and only natural hazards, are used by Acerbo et al. [Acerbo and Rossi, 2017] for their classification approach. Adding additional related tweets from the Twitter API they ended up with a dataset of 36K tweets. Yet, in their approach they removed tweets which are labeled as "not applicable" or "not related".

Cresci et al. [Cresci et al., 2015] used a manually annotated dataset of four crisis events taken place in Italy. 5.6K tweets in total consists of two floods and two earthquakes. Stowe et al. [Stowe et al., 2016] created their own corpus using tweets from Hurricane Sandy. They ended up with 7.5K tweets posted by 93 different users. A realatively small dataset was used by Verma et al. [Verma et al., 2011]. Their dataset contains 2K tweets, approximately 500 tweets per event, collected in the years 2009 and 2010. Imran et al. [Imran et al., 2013] used 4.4K tweets from to the Tornado in Joplin, Missouri in 2011.

[1] http://www.crisislex.org/data-collections.html

To conclude, for research purposes, not many labeled datasets exist due to the time consuming and costly creation and labeling process of datasets. Many research work set up on already existing datasets. Since the CrisisLexT26 is one of the largest datasets available, which consists of various events, it is chosen as a basis for this thesis.

2.3 Data classification

Beside the dataset used for classification, choosing the "right" features is not less important to get accurate classification results [Pekar et al., 2016]. Various types of supervised Machine Learning algorithms use training examples from the dataset to train a classification model. The outcome is the trained model or classifier, respectively, which than can be used to predict unseen tweets of future events. Many different types of supervised Machine Learning algorithms exist to classify crisis related tweets and many of them are also used in related literature. Depending on the particular task, different algorithms lead to more or less suitable results. However, using which classification algorithm with which configurations for which dataset is almost impossible to generalize, although there exist some preferred approaches in related literature. The current section describes algorithms and settings used for classification. Furthermore, this section describes the data used for training and testing the classifier and points out the differences in in-domain, out-domain and cross-domain classification.

2.3.1 Algorithms

Various implementations of supervised Machine Learning algorithms exist. However, for classifying crisis related information some of them are more frequently used than others, namely Support Vector Machines (SVM), Maximum Entropy Models, K-Nearest Neighbor (KNN) and Random Forest classifications. Linear SVM are very frequently used Machine Learning algorithms when it comes to higher dimensional data [Pedregosa et al., 2011]. In the work of [Cresci et al., 2015][Stowe et al., 2016][Khare et al., 2017][Mohammad et al., 2013] and [Pekar et al., 2016] SVMs are used for their classification task. Moreover, Naive Bayes classifier and Maximum Entropy Models, used in the work of [Stowe et al., 2016][Verma et al., 2011] and [Pekar et al., 2016], lead to suitable classification results. Pekar et al. [Pekar et al., 2016] additionally experimented and compared in their work K-Nearest Neighbor (KNN) and Random Forest classifier. Nevertheless, best classification results achieved SVM and Maximum Entropy Models. While the Random Forest classifier, used by Pekar et al. [Pekar et al., 2016], was not the best choice regarding classification accuracy, in the work of Acerbo et al. [Acerbo and Rossi, 2017] the Random

Forest classifier lead to the best results. This aspect shows that classification accuracy does not only depend on the algorithm, rather different algorithm configurations or even more relevant, different features used can lead to different classification results while using the same classification algorithm. Which classifier to choose depends on various aspects and thus training the classifier is more an experimental setup and evaluation task. Nevertheless, testing the classifier on unseen data or applying Cross-Validation estimates the performance of the classification model on real data and ensures that future classification will lead to reliable and suitable classification results. The underlying assumption thereby is that training data represents the unseen future data, which again emphasizes the importance of using the "right" data for training the classifier.

2.3.2 Data for training and testing

As already argued at the end of the previous section, using the "right" training data is crucial to get reliable and suitable classification results. In general, literature categorizes classification experiments in three main types with respect to crisis events. First, training on data or tweets, respectively of a particular type of event, for example tweets of Hurricane Sandy, and testing on another event, but from the same type, for instance tweets of Hurricane Irma. Researchers [Cresci et al., 2015] call this *in-domain*, i.e., training and testing within the same type of event. The second category is called *out-domain* classification. Thereby, the classifier is trained on one type of event, e.g. Hurricane Sandy, and tested on tweets of a different type of event, e.g. an earthquake. In the third category training data consists of more than one event, for instance using tweets of Hurricane Sandy, tweets of an earthquake in Chile as well as tweets from a helicopter crash in the UK to train the classifier. In the current thesis this type of classification is called *cross-domain* (training). In addition to these three main types and not comparable to cross-domain there exist *Cross-Validation* [Pedregosa et al., 2011], which is a concept frequently used in the Machine Learning domain and also in crisis related literature. Cross-Validation splits the training data, e.g. from one particular event and uses parts of the same event for training as well as for testing. This is typically used for estimating the classification results during the experimental phase.

In general, the accuracy of classification results tend to decrease from Cross-Validation over in-domain, out-domain to cross-domain, due to the reason that in an event are more similarities within the same event than with another event or even with an event of another type. Based on the used dataset, literature uses a variety of training and testing scenarios. Since it is practically not feasible or very time consuming and costly, respectively to create a corpus for each type of event, e.g. earthquakes, floods, shootings, researchers pay out-domain and cross-domain approaches more attention than in-domain

approaches, although out-domain and cross-domain classification are reasonable harder to realize [Cresci et al., 2015].

2.3.3 Machine Learning terms

For clarification, the following list gives a set of definitions used in the Machine Learning domain. All explanations and examples are related to informativeness classification, i.e. supervised learning, which is used in the context of this master thesis. Moreover, definitions are based on the books of [Russell and Norvig, 2016] and [Bishop, 2006].

- A **Dataset or Corpus** contain data or input-output samples, which are used for training and testing the classifier. Usually the entire dataset is split up into training data and test data. The ratio for splitting depends on the requirements of the classification task.

- **Training data** contains input-output samples used for training the classifier (e.g. past disaster events). Thereby, input are features or their values, respectively, and the output is the corresponding class, i.e. informative or not informative. The learning algorithm tries to find a function which match given input data, i.e. features, to a given output class.

- **Test data** is used for testing the classifier and simulates a new disaster event. Thus, based on input data, i.e. features, the classifier tries to predict the corresponding class.

- **Classes** are the categories, which the classifier tries to classify certain input data. For informativeness classification there are two classes, informative and non informative.

- **Features** are properties which describes a single object, for instance a certain tweet. Feature are extracted for instance from tweet text, e.g. the number of words in the text. Each object for classification is described by a set of n features, which results in a n-dimensional representation of the object.

- **Cross-Validation** is a technique used for evaluation and selecting the best performing classifier and is used during development. In S-fold Cross-Validation, the entire dataset is split into S parts where S-1 parts are used for training the classifier and the remaining part is used for testing the classifier. The final performance score is the average over all S runs.

	predicted informative	predicted non informative
true informative	**TI**	FN
true non informative	FI	**TN**

Table 2.1: Confusion matrix

- **Performance evaluation** and performance measures are used to compare different classification approaches or results of different classifiers. Since informativeness classification is a binary classification, the confusion matrix of table 2.1 show prediction results.

  Frequently used performance metrics in literature are:

  – Accuracy: The number of correct classified objects.

  $$Acc = \frac{TI + TN}{TI + FN + FI + TN} \tag{2.1}$$

  – Precision: Proportion of correct informative predictions

  $$Prec = \frac{TI}{TI + FI} \tag{2.2}$$

  – Recall: Proportion of correct classified informative tweets

  $$Rec = \frac{TI}{TI + FN} \tag{2.3}$$

  – F1-value:

  $$F1 = 2 * \frac{Prec * Rec}{Prec + Rec} \tag{2.4}$$

Chapter 3

Related Work

With increasing popularity of Social Media services more researchers are trying to use information from Social Media also in disaster situations [Olteanu et al., 2015].

This chapter discusses related literature with respect to the current thesis. The first section deals with informativeness classification in general, in other areas as well as in the crisis domain. The second section focuses on related work with respect to a systematic manual analysis of crisis related data and in the following section, main differences to other closely related classification approaches are factored out. Table 3.1 shows a summary of closely related work, sorted firstly with respect to the classification task, secondly concerning the dataset used and thirdly regarding other related aspects, such as used features or cross-domain training. Related classification approaches are discussed along this table.

3.1 Informativeness classification

Informativeness is a broadly discussed concept in literature and applied in various areas like news articles [Lloret and Palomar, 2016], web documents [Horn et al., 2013] and linguistic sciences [Kireyev, 2009][Wu and Giles, 2013], to mention just a few. In the crisis domain, informativeness classification is one early step in processing information from Social Media. While few research thereby focuses on the informativeness classification task [Acerbo and Rossi, 2017][Imran et al., 2016][Longhini et al., 2017], other work take that for granted and focus on more specific classification tasks like damage assessment [Cresci et al., 2015] or develop platforms [Cameron et al., 2012] and frameworks [Avvenuti et al., 2016][Ren et al., 2017], which support crisis management as a whole. While many approaches for informativeness classification focus on particular disaster types [Acerbo and Rossi, 2017][Cresci et al., 2015][Longhini et al., 2017], only few works, however, address classification cross a variety of disaster types

[Khare et al., 2018][Khare et al., 2017] and also in other application areas like e.g. news [Dai et al., 2007] cross-domain approaches are sparse.

3.2 Manual crisis data analysis

Important for any Machine Learning task is an appropriate corpus for training and testing the classifier. Based on training examples or past tweets respectively, the classifier learns a classification model, which then is used to predict new unseen tweets. Recent studies found out that information posted online during disaster varies a lot between different types of events [Olteanu et al., 2014][Munro and Manning, 2012][Kanhabua and Nejdl, 2013]. While many research focus on feature-engineering, i.e., inventing and evaluating new features, [Khare et al., 2018][Khare et al., 2017][Longhini et al., 2017] or classification itself [Pekar et al., 2016][Stowe et al., 2016][Verma et al., 2011], only few research deal with crisis data analysis *with respect to informativeness*. Acerbo & Rossi investigate "common patterns" inside informative and non informative tweets, which is similar to the data analysis of the current thesis, yet, they focus on similarities and differences in words. A statistical analysis of data with respect to informativeness is done by Lloret & Palomar [Lloret and Palomar, 2016] where they present linguistic features which "an informative tweet should have in order to be informative", focusing, however, on the news domain. Ning et al. [Ning et al., 2017] present an analysis of six disasters, to identify linguistic, sentimental and emotional features. In contrast to this thesis, they address "relatedness" rather than "informativeness" of tweets. The manual crisis data analysis of the current thesis is based on the results of Olteanu et al. [Olteanu et al., 2015], additionally, on top of that this thesis focuses on a detailed analysis with respect to informativeness.

3.3 Classification approaches

Most closely related works deal with informativeness classification of Social Media messages. This section wants to discuss closely related work along Table 3.1 and tries to point out main differences with respect to this thesis.

Considering closely related informativeness classification approaches, Acerbo & Rossi [Acerbo and Rossi, 2017] base their work on a subset of the CrisisLexT26 and solely used natural hazards with more than one event per event type. Hence, their dataset contains three disaster types: floods, earthquakes and fires. In contrast to this thesis, their goal was not to learn a cross-domain classifier, rather they focus on a novel text metric to use as additional feature for classification. In addition, they exclude all tweets, which are labeled as "not related" and "not applicable" from the original dataset. Since the classification

approach proposed in this thesis is designed for online classification, i.e. classifying tweets in real time on the message stream, removing tweets from the dataset does not correspond with the final use case and thus the current thesis uses all tweets, shared by users, for classification. Used features by Acerbo & Rossi like hashtags, URLs, "@" character and source information are also used in the current thesis. Additionally, they accessed the Twitter API to get user information as features, for instance, number of followers and followees, registration date, total number of tweets by user and also Retweets, which are, however, not used in this thesis, but might be part of future works.

Closely related with respect to the number of cross-domain experiments is the work of Imran et al. [Imran et al., 2016], which is also based on a subset of the CrisisLexT26, yet, they only use two disaster types, namely floods and earthquakes. In contrast, this thesis uses the entire CrisisLexT26 dataset including 26 events from 13 different event types. Beside tweets in English, this dataset contains also tweets of other languages, mainly Spanish, Portuguese and Tagalog. It includes human-induced events as well as natural hazards. Important for data analysis is using a variety of events and disaster types to uncover differences and similarities among them. In order to uncover characteristics within the same disaster type it is important to analyze several events of the same type, i.e. in-domain. As well, it is important to analyze a variety of different event types to investigate out-domain differences and similarities. Thus, by using all 26 events, this thesis covers both, a variety of event types as well as several events from the same type. Additionally, Imran et al. experiments with adding portions of the test set to the training set, i.e. test on the same data already used for training, which led to reasonable classification improvements. However, this is likewise not applicable for online classification and therefore not in the scope of the current thesis.

Longhini et al. [Longhini et al., 2017] present a "language-agnostic model" for informativeness classification and show the impact of a new feature "source", indicating hardware for communication, e.g. mobile or not. Cross-domain classification, however, is not part of their work.

Considerable more disaster types as all previously mentioned approaches are used by Khare et al. [Khare et al., 2017]. As in this thesis, they run cross-domain experiments, yet their classification task is "relatedness" (whether a tweet is related to a crisis event or not) instead of "informativeness". Statistical features used by Khare et al. like Part-of-Speech information, and tweet lenght are part of this thesis. To address the problem of "lexical sparsity of tweets" [Cresci et al., 2015], i.e., tweets can be per definition only 140 characters long, Khare et al., as well as [Cresci et al., 2015] and [Imran et al., 2013], use a similarity, synonym and hypenym lexicon based approach. Words in tweets were replaced with synonyms from a lexicon to eliminate the issue of same meanings by different

words. In addition, by using "Babelfy"[2] they extracted "semantic entities" out of the tweet text, which they further used to look up for hypernyms or synonyms on a more abstract level in a knowledge base called Bebel.net[3]. Using a more abstract level, e.g. the words "firemen" or "policemen" are translated in the more general term "defender", allows to encapsulate the specific event type. The current thesis focuses mainly on simple linguistic featues, however, more complex features, like semantic concepts, considering hypernyms and synonyms might be part of future work. Most closely related to this thesis is another very similar work of Khare et al. [Khare et al., 2018] where they use the entire CrisisLexT26 dataset to address cross-domain classification, focusing, however, on "relatedness", not on "informativeness".

The work of [Cresci et al., 2015] is closely related with respect to features used for classification like number of tokens, "@"-symbols, hashtags, punctuation, Emoticons, sentiment and Part-of-Speech (POS) information, like number of adjectives, number of verbs or number of nouns. Additionally, they address cross-domain classification by using four disaster events of two disaster types in Italy. They do not focus, however, on informativeness classification but rather on "damage assessments" of Italian tweets. Another type of feature was used in the work of Cresci et al., and also used by Mohammad et al. [Mohammad et al., 2013]. They tried to extract the sentiment, which is implicit in the tweet content, and used it as features. Many research is done to determine the sentiment of tweets in an automated way. Cresci et al. used in their work already existing sentiment polarity lexicons as well as an own developed classification approach which learns and classifies the sentiment based on positive and negative tweets in a labeled corpus. Mohammed at al. [Mohammad et al., 2013] created an own "Hashtag sentiment lexicon", where Hashtags used in tweets are directly related to positive or negative sentiment. For example, Hashtags like #good, #angry express the emotion of the user who posted a certain tweet. In order to determine the sentiment of a tweet and further to use as a feature, the current thesis applied an existing sentiment library, namely TextBlob[1].

Closely related with respect to cross-domain classification are the works of Li et al. [Li et al., 2015] and Imran et al. [Imran et al., 2013], yet they consider two different disaster types, only.

3.4 Conclusion

Finally, this section summarizes the main aspects which distinguish this thesis from other related work.

[2]http://babelfy.org/
[3]http://babelnet.org/
[1]http://textblob.readthedocs.io/en/dev/index.html

Approach	Goal	Dataset					Features			Training & Evaluation			
		Dataset	Language	No. tweets	No. disasters	No. types	Linguistic	Temporal	Spatial	Algorithm	In-domain	Out-domain	Cross-domain
[Acerbo and Rossi, 2017]	Informativeness	CrisisLexT26	EN,ES,TL	36.0K	12	3	✓	✓		Random Forest			✓
[Imran et al., 2016]	Informativeness	CrisisLexT26	EN,ES,TL	11.7K	11	2	✓			Random Forest	✓	✓	✓
[Longhini et al., 2017]	Informativeness	CrisisLexT26	EN,ES,TL	12.9K	12	3	✓	✓		Random Forest	✓		
[Khare et al., 2017]	Relatedness	CrisisLexT26	EN	3.2K	9	7	✓			SVM			✓
[Khare et al., 2018]	Relatedness	CrisisLexT26	EN,ES,TL,PT	32.0K	26	11	✓			SVM		✓	✓
[Cresci et al., 2015]	Damage assessment	own	IT	5.6K	4	2	✓			SVM	✓	✓	✓
[Ning et al., 2017]	Relatedness	CrisisLexT6	EN	32.5K	6	5	✓		✓	CNN			✓
[Li et al., 2015]	Relatedness and others	own	EN	2.7K	2	2	✓			Naive Bayes			✓
[Imran et al., 2013]	Informativeness and others	own	EN	6.4K	2	2	✓			Naive Bayes	✓		
[Graf et al., 2018]	*Informativeness*	*CrisisLexT26*	*EN,ES,TL,PT*	*28.0K*	*26*	*13*	✓	✓		*SVM*	✓	✓	✓

Table 3.1: Overview on informativeness classification approaches.

First, the current thesis [Graf et al., 2018] uses a larger dataset, regarding amount of samples as well as regarding variety of events and disaster types. Imran et al. [Imran et al., 2016] showed in their work that accuracy improves significantly when including more data for training, even when this data is from other events. As table 3.1 shows, many experiments use significantly smaller datasets, some events for training contain only a few hundred tweets. Hence, not only the size of the entire dataset matters, also the number of tweets per event, which are used for training, is highly relevant. The number of training examples decreases even more when applying some filter or using a under-sampled balanced dataset, i.e., create an equal distribution of classes for training. The current thesis uses a dataset including 12K more tweets than the average dataset used in closely related work of Table 3.1. According the number of events and the variety of event types, the dataset used in this thesis is even the largest.

Moreover, this thesis provides a systematic manual data analysis, which is based on the results of Olteanu et al. [Olteanu et al., 2015], and additionally, on top of that, focuses particularly on informativeness. Such a detailed analysis of the CrisisLexT26 dataset with respect to informativeness does not exist so far. First, this data analysis allows suitable feature selection, i.e., detailed insight in data show whether certain data characteristics are suitable to use as a feature for classification. In addition, data analysis provides information for appropriate composition of train and test set to achieve reliable and accurate classification results on unseen events.

The classification approach proposed in this thesis is designed for real time, i.e., online, classification. Therefore, certain features used in related work, like number of Retweets or other tweet reply information are not used for classification due to the reason that some information is not available when a certain tweet is classified.

By running in-domain, out-domain and cross-domain experiments, the current thesis is able to show the impact of the crisis domain on informativeness classification results, i.e., to answer the question whether a cross-domain trained classifier is more accurate in classifying crisis related tweets with respect to informativeness than a crisis specific one. In

general, less research focus on cross-domain training, since cross-domain training requires an appropriate dataset including a higher number of events as well as events from different types.

Chapter 4

Data Analysis

Analyzing crisis-related data and furthermore using those insights in data to improve informativeness classification is the overall scope of the current master thesis. The CrisisLexT26, created by [Olteanu et al., 2015], seems to be best suitable for this task for mainly two reasons. Firstly, it is one of the largest dataset available containing approximately 28K tweets, which are manually labeled with respect to informativeness. Secondly, the CrisisLexT26 dataset contains a variaty of different event types, natural hazards as well as human induced events. Furthermore, these events happened all around the globe. Some events took place in North and South America, moreover, there are events from the Asian area, for instance Philippines or Bangladesh, disaster events took place in Europe, Spain, Italy or UK, as well as in Australia. Thus, also in the sense of geographic location, the CrisisLexT26 can be considered as a heterogeneous collection of potentially crisis related tweets. The first section of the current chapter explains the CrisisLexT26 in detail, mainly based on the work of Olteanu et al. [Olteanu et al., 2015], to get a better understanding of how the CrisisLexT26 dataset was developed and manually annotated.

Based on the work of Olteanu et al. [Olteanu et al., 2015], the systematic analysis of the CrisisLexT26 presented in the second section of the current chapter addresses two main

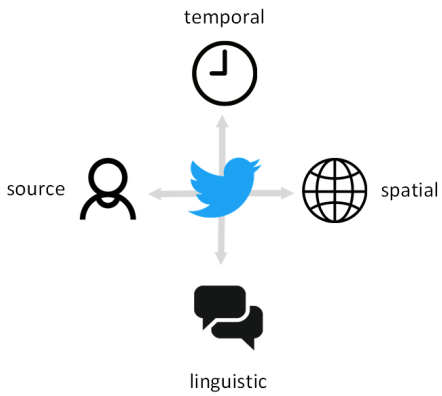


Figure 4.1: Analysis dimensions

goals: Firstly, analysis wants to uncover similarities and differences of different crisis events and in particular of crisis events of different disasters types. This provides the basis for creating an appropriate training set, which addresses the hypothesis accurate informativeness classification over various disaster types, i.e., addressing the domain adaptation problem. Secondly, analysis wants to uncover the impact of tweet characteristics with respect to informativeness, to use specifically those having high impact on informativeness as features for classification, which addresses the hypothesis of a more accurate informativeness classification compared to other approaches. The systematic analysis is based on 4 different dimensions (cf. Figure 4.1), comprising i) WHEN a tweet was shared, i.e., *temporal* dimension, ii) WHERE a tweet geographically belongs to, i.e., *spatial* dimension, iii) HOW a tweet is written, i.e., *linguistic* dimension and iv) WHO posts a tweet, i.e., *source* dimension. In order to uncover tweet characteristics correlating with informativeness and their differences with respect to disaster types, analysis of each dimension follows a systematic hierarchical process. To be more specific, the analysis considers, firstly, each dimension on an aggregated level, secondly, on a crisis event level, and, thirdly, in combination with other dimensions. Overall, for each dimension, the analysis is detailed as far as significant differences with respect to informativeness are encountered.

The final section of the current chapter focuses on event similarity by applying hierarchical clustering to show similarities of events based on event dimensions given by Olteanu et al. and tweet characteristics discussed in the current chapter. Clustering results are able to show the impact of tweet characteristics on event similarities. In other words, clustering visualizes similarity of events based on the tweets. More details are explained in Section 4.3.

4.1 CrisisLexT26 dataset

The CrisisLexT26 dataset contains potentially crisis-related tweets from 26 past disaster events, which happened in the years 2012 and 2013 all over the world. Basically, events can be distinguished between natural hazards, for instance earthquakes or floods, and human induced crisis events, for instance shootings or plane crashes. The dataset contains events from both categories, 15 natural events, 10 human induced events and one event which can be seen in both categories. Olteanu et al. [Olteanu et al., 2015] categorize those 26 events furthermore according their subcategory, type, development and spread. All the dimension and characteristics of each events are listed in table 4.1. The column "Days" refers to the duration of the event, i.e., only tweets from this time frame are included in the dataset. The column "Tweets" does not show the final number of tweets in the dataset, but rather the number of tweets collected to a certain crisis event.

Year	Country	Crisis Name	Days	Tweets	Category	Subcategory	Type	Development[a]	Spread[b]
2012	Italy	Italy earthquakes	32	7.4K	Natural	Geophysical	Earthquake	Instantaneous	Diffused
2012	US	Colorado wildfires	31	4.2K	Natural	Climatological	Wildfire	Progressive	Diffused
2012	Philippines	Philippines floods	13	3.0K	Natural	Hydrological	Floods	Progressive	Diffused
2012	Venezuela	Venezuela refinery explosion	12	2.7K	Human-induced	Accidental	Explosion	Instantaneous	Focalized
2012	Costa Rica	Costa Rica earthquake	13	2.2K	Natural	Geophysical	Earthquake	Instantaneous	Diffused
2012	Guatemala	Guatemala earthquake	20	3.3K	Natural	Geophysical	Earthquake	Instantaneous	Diffused
2012	Philippines	Typhoon Pablo	21	1.9K	Natural	Meteorological	Typhoon	Progressive	Diffused
2013	Brazil	Brazil nightclub fire	16	4.8K	Human-induced	Accidental	Fire	Instantaneous	Focalized
2013	Australia	Queensland floods	19	1.2K	Natural	Hydrological	Floods	Progressive	Diffused
2013	Russia	Russian meteor	19	8.4K	Natural	Others	Meteorite	Instantaneous	Focalized
2013	US	Boston bombings	60	157.5K	Human-induced	Intentional	Bombings	Instantaneous	Focalized
2013	Bangladesh	Savar building collapse	36	4.1K	Human-induced	Accidental	Collapse	Instantaneous	Focalized
2013	US	West Texas explosion	29	14.5K	Human-induced	Accidental	Explosion	Instantaneous	Focalized
2013	Canada	Alberta floods	25	5.9K	Natural	Hydrological	Floods	Progressive	Diffused
2013	Singapore	Singapore haze	19	3.6K	Mixed	Others	Haze	Progressive	Diffused
2013	Canada	Lac-Megantic train crash	14	2.3K	Human-induced	Accidental	Derailment	Instantaneous	Focalized
2013	Spain	Spain train crash	15	3.7K	Human-induced	Accidental	Derailment	Instantaneous	Focalized
2013	Philippines	Manila floods	11	2.0K	Natural	Hydrological	Floods	Progressive	Diffused
2013	US	Colorado floods	21	1.8K	Natural	Hydrological	Floods	Progressive	Diffused
2013	Australia	Australia wildfires	21	2.0K	Natural	Climatological	Wildfire	Progressive	Diffused
2013	Philippines	Bohol earthquake	12	2.2K	Natural	Geophysical	Earthquake	Instantaneous	Diffused
2013	UK	Glasgow helicopter crash	30	2.6K	Human-induced	Accidental	Crash	Instantaneous	Focalized
2013	US	LA Airport shootings	12	2.7K	Human-induced	Intentional	Shootings	Instantaneous	Focalized
2013	US	NYC train crash	8	1.1K	Human-induced	Accidental	Derailment	Instantaneous	Focalized
2013	Italy	Sardinia floods	13	1.1K	Natural	Hydrological	Floods	Progressive	Diffused
2013	Philippines	Typhoon Yolanda	58	39.0K	Natural	Meteorological	Typhoon	Progressive	Diffused

[a] temporal development of crises
[b] geographic spread of a crisis

Table 4.1: Crisis events of the CrisisLexT26 dataset [Olteanu et al., 2015]

4.1.1 Data source

Source of the CrisisLexT26 dataset is Twitter's Sample API, which is available for public use. Yet, Twitter's Sample API only provides 1% of the entire tweet stream of Twitter, i.e., all tweets posted by users. Previous studies have showed that although only 1% of all tweets are accessible, this 1% of all tweets are up to some point representative for all posted tweets and do not bias the results [Olteanu et al., 2015]. Disaster events, which had associated more than 100,000 tweets are considered for this dataset of 26 events, since 1% of tweets from the Sample API result in at least 1K tweets. In a first step, all public available tweets from the years 2012 and 2013 were accessed using Twitter's Sample API. In a second step tweets were filtered and associated to certain disaster events based on Hashtags and Keywords. Very short tweets, equal or less than three tokens long, were removed from the dataset, because usually they contain no useful information at all. They did not remove Retweets or duplicates from the dataset, since they want to keep the information of people repeating already existing messages.

4.1.2 Labeling process

Manual labeling of tweets in the work of Olteanu et al. was done by crowdsource workers employed for this particular task. Each tweet was manually annotated towards three dimensions, namely informativeness, information type, information source. To reduce labeling bias by subjective interpretations, the authors provided detailed definitions and

instructions for the task. Informativeness was defined as "checking whether the tweet contributes to better understanding the situation on the ground" [Olteanu et al., 2015]. Still, this definition allows some subjective judgment. To address this issue, first, at least three worker labeled each tweet, and second, also the authors contributed in the labeling process and compared their annotations with the annotations of the crowed workers to reduce the labeling bias. The following list shows examples of informative and non informative messages:

1. Informative
   informative (direct or indirect); curating or producing content; contribute to situational awareness; situational information; contextual information to better understand the situation

2. Not Informative
   trolling; humor; off-topic; rumor; humor or irrelevant/spam;

Finally, the informativeness labeling process ended up with four classes, namely "informative and related", "related but not informative", "not related" and "not applicable".

Information type and information source were annotated in the same way. Labeled classes for information source are "Eyewitness", "Government", "NGOs", "Business", "Media" and "Outsiders" [Olteanu et al., 2015].

1. Eyewitness
   citizen reporters, members of the community; eyewitnesses; local, peripheral, personally connected; local individuals ; local perspective, on the ground reports; direct experience (personal narrative and eyewitness reports); direct observation, direct impact, relayed observation;

2. Government
   (news organizations and) authorities; government/administration; police and fire services; police; government; public institutions; public service agencies, flood specific agencies;

3. NGOs
   non-profit organizations; non-governmental organization; faith-based organizations; Business commercial organizations; enterprises; for-profit corporation;

4. Media
   news organizations (and authorities), blogs; journalists, media, and bloggers; news organization; professional news reports; media; traditional media (print, television,

radio), alternative media, freelance, journalist; blogs, news-crawler bots, local, national and alternative media; media sharing (news media updates, multimedia);

5. Outsiders
   sympathizers; distant witness; remote crowd; non-locals

Class labels of the dimension information type are "Affected individuals", "Infrastructure and Utilities", "Donations and volunteer", "Caution and advice", "Sympathy and emotion", "Other useful information" [Olteanu et al., 2015]. Since this dimension is mainly related to semantic content of a tweet and hardly related to informativeness, information type is not discussed in detail in the current thesis.

Finally, Olteanu et al. ended up with approximately 28K labeled tweets and published their collection of 26 events for research purposes as CrisisLexT26[1] dataset.

4.2 Systematic crisis data analysis

In the following a systematic and in-depth analysis of crisis related tweets with respect to informativeness is provided along four different dimensions, covering temporal, spatial, linguistic and source characteristics (cf. Figure 4.1) on basis of the CrisisLexT26 [Olteanu et al., 2015] dataset.

Since this thesis deals with informativeness classification, the distribution of informativeness in the dataset is an important aspect. To get a complexity reduced view on the data, this thesis only considers two informativeness classes, namely "informative" and "not informative". Thus, the class labels "related but not informative", "not related" and "not applicable" are joined together and all belong to one class "non informative". Second reason for joining these classes to one is that for a practical use on the Twitter message stream, this binary classification can be considered as an early stage filtering task, where only informative tweets are further processed. In total, the dataset contains 60.11% informative and 39.89% non informative tweets. Figure 4.2 visualizes the overall class distribution, i.e., the distribution of informative and non informative tweets, of all events. The majority of events included in the CrisisLexT26 tend to have more tweets in the "informative" class. The most unbalanced event is the NYC train crash where over 90% of tweets belong to the "informative" class and thus less than 10% to the "non informative" class. Class distribution, i.e., whether classes are balanced or not, is important for the classification approach. An unbalanced training set may bias classification accuracy and thus have to be considered in the classification approach.

[1] http://www.crisislex.org/data-collections.html

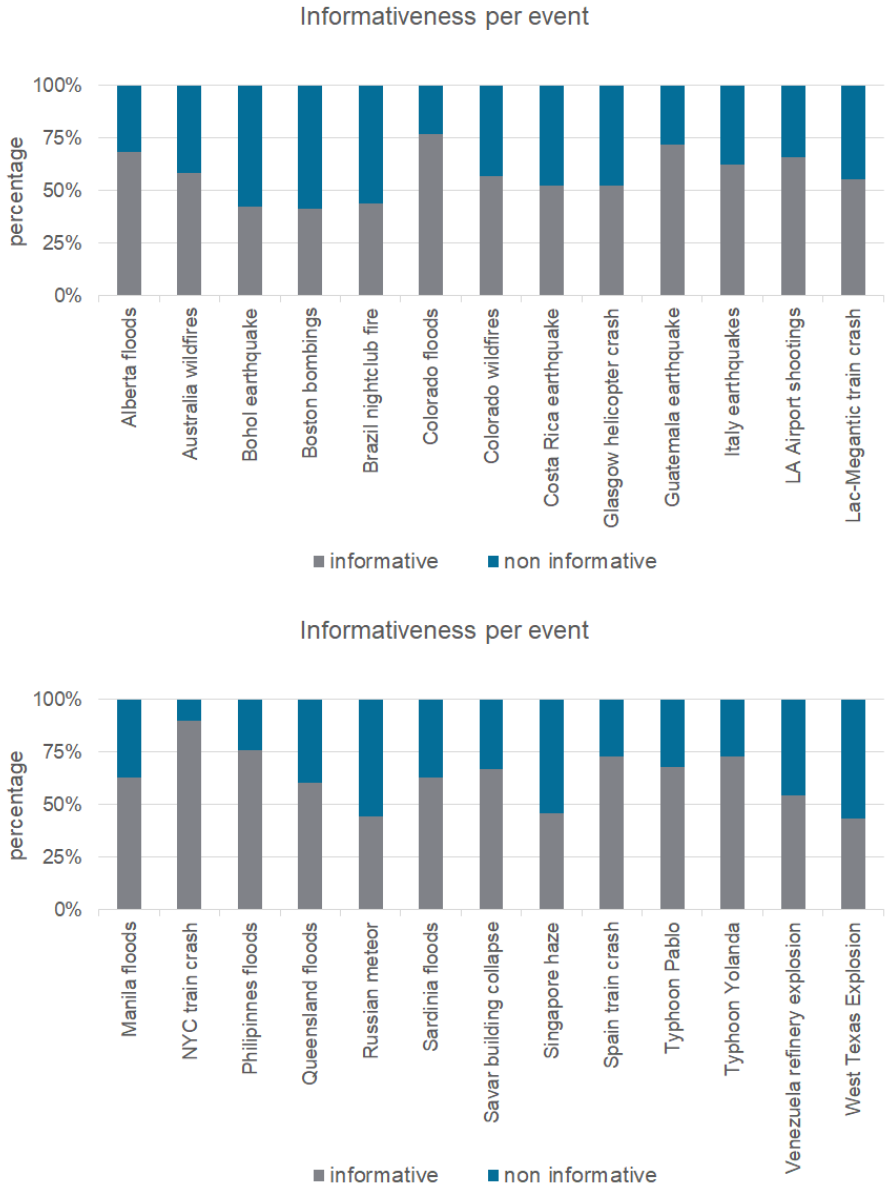


Figure 4.2: Informativeness per event

4.2.1 Temporal dimension

By considering the temporal dimension of tweets this thesis wants to analyze the evolution of informativeness in time over an entire crisis. For this, it is investigated how far tweet characteristics determined by spatial, source and linguistic dimensions change from the beginning of a crisis to its end. Particular emphasis is put on whether there are differences between crisis events or between disaster types, apart expectable peculiarities induced by the instantaneous or progressive character [Olteanu et al., 2015] of certain disaster types like bombings or floods. In the following, most important aspects are discussed in more detail.

4.2.1.1 Informativeness per time

The evolution of communication in time regarding informativeness during crisis events may give interesting insights in people's Social Media behavior. If the relation of informative to non informative tweets differ in time, for instance in early stages there are more non informative tweets and barely informative tweets or the other way round, this knowledge can be used for adapting the classification approach, for instance using different classifiers for different time phases during the entire disaster.

Figure 4.3 shows the evolution of informativeness over all events during the first 31 days after the event took place. 31 days represents one month after the event occurred and include 99.45% of all tweets from the entire dataset. Especially progressive events demand to consider over a longer period of time and yet 31 days are still possible to visualize in form of a time line. The chart shows that in the first few days of an event the majority of tweets about that event were posted. Furthermore, Figure 4.4 visualizes the difference in communication regarding the time development dimension of events, which is clearly related to time. Instantaneous events, like earthquakes, shootings or train crashes trigger immediately lot of online communication in the very early stages after the event starts. In contrast, progressive event types, which might build up in severeness gradually, like floods or bush fires lead to more continuous online communication which might build up over time. While there are differences regarding the time development dimension of an event, whether it is an instantaneous or progressive event, in both categories there are still informative and non informative tweets over the entire time period. However, the amount of tweets decreases significantly over time.

Regarding in-domain, out-domain and cross-domain classification, similarity of events with respect to informativeness over time is essential to look at. The charts in Figure 4.7 show the events in time grouped by their event type. Hence, similarities or even not existing similarities in the same domain or cross another domain are visible. Out of this charts some

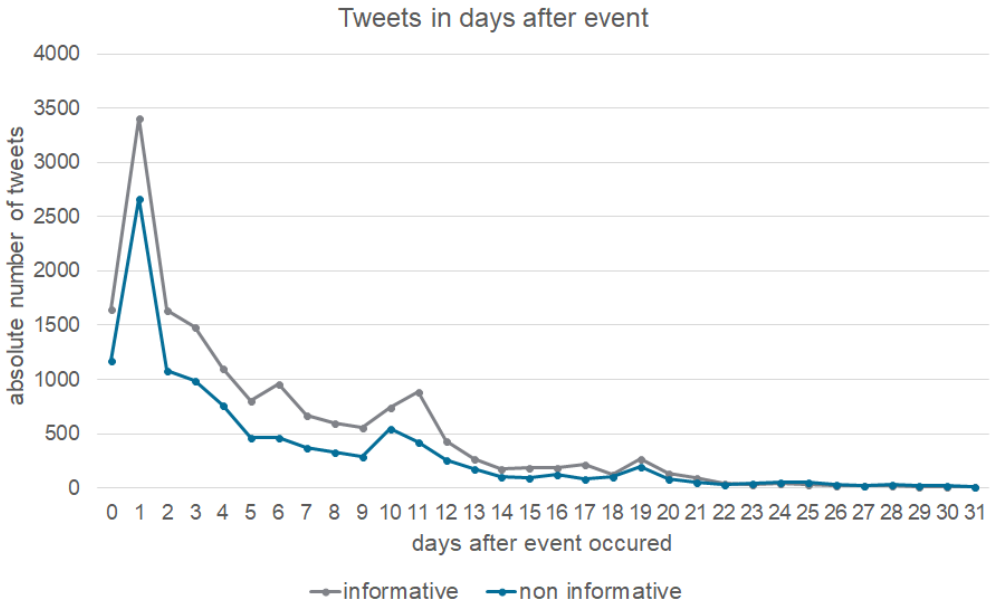


Figure 4.3: Informativeness per time

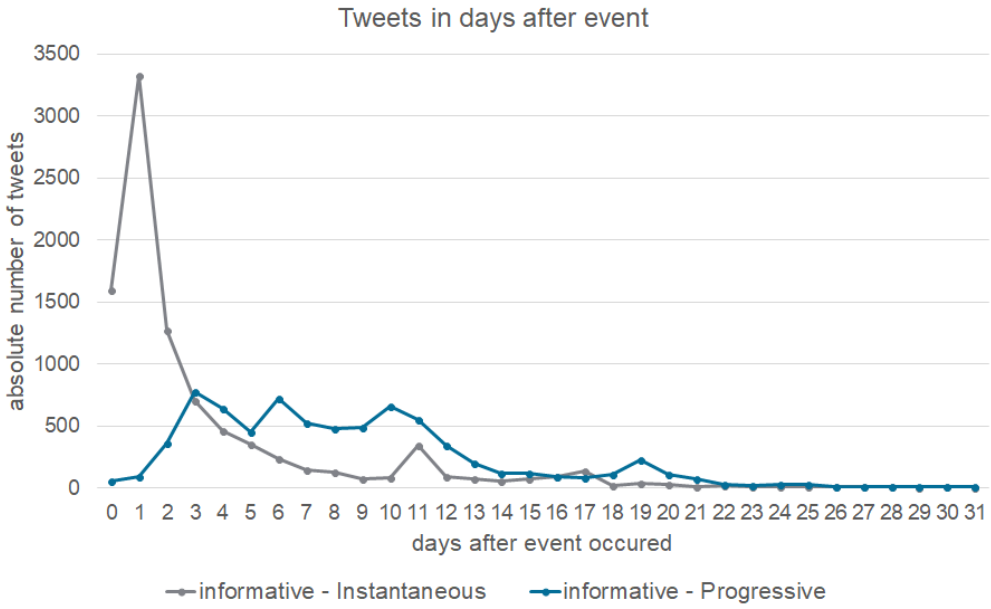


Figure 4.4: Informative tweets in respect to time development dimension

observations can be drawn. Firstly, the temporal progress of progressive events, like floods or wildfires, look different even within the same event type. For instance, comparing both wildfires, Colorado and Australia, show that tweets were posted on different days. The point in time when most of the tweet communication take place depends more on particular occurrences within the event itself. In the case of floods, this may be a rising water level. As a consequence, more people are directly affected by the higher water level and so more people share information online. Secondly, instantaneous events, whether they are natural or human-induces does not make significant differences, tend to look similar even over other domains. Comparing the charts of earthquakes, derailments, explosions and other instantaneous events, in all of these events a majority of the entire communication take place in early stages of the event, independent from the actual type of event. Hence, informativeness over time tend to be the same in instantaneous types of events. Typhoons are a progressive type of event and both of them look similar in time, this is not necessarily representative for all Typhoons. Like in floods, the number of affected people in Typhoons during the entire crisis can vary and consequently has impact on online communication.

Possible existing geographic and cultural differences in informativeness evolution over time show the charts in Figure 4.9. Disaster events are clustered based on their country and each continent is represented in one chart. In general, the relation informative to non informative tweets seems to be similar over all five charts. However, there exist some fine grained distinctions. In Asian events predominate non informative tweets in the first two days. Whereas, this is not the case for all other events, rather it is the other way round. Especially in European events, South- and North America, in early stages of the events tweets tend to be mainly informative. Yet, informative tweets drop down below non informative tweets in later phases of the event in Europe or in South America. For instance, in South American events, six days after the event occurred, there are more non informative tweets than informative tweets.

To conclude, a majority of tweet communication take place in early stages of the event. 50.06% online communication is within the first three days after the event took place. This is even more the case in instantaneous events, where 71.87% of all tweets are posted within the first three days. In general, the timeline of informative tweets of instantaneous events look more similar even over different event types. This is not the case in progressive events where tweet behavior over time depends on particular occurrences within the event itself.

4.2.1.2 Findings

Main findings regarding the temporal dimension can be summarized as follows:

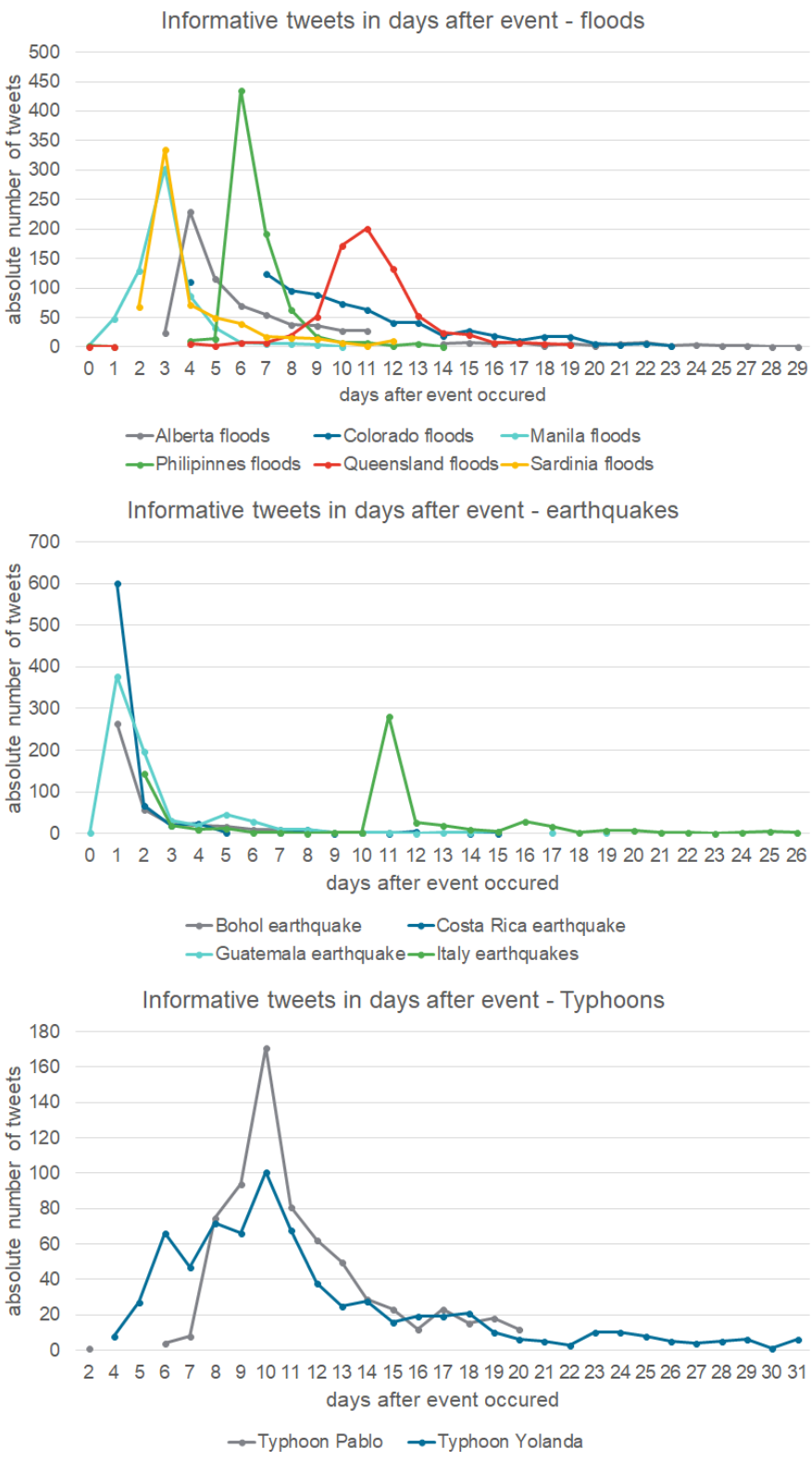


Figure 4.5: Informative tweets per event type (1)

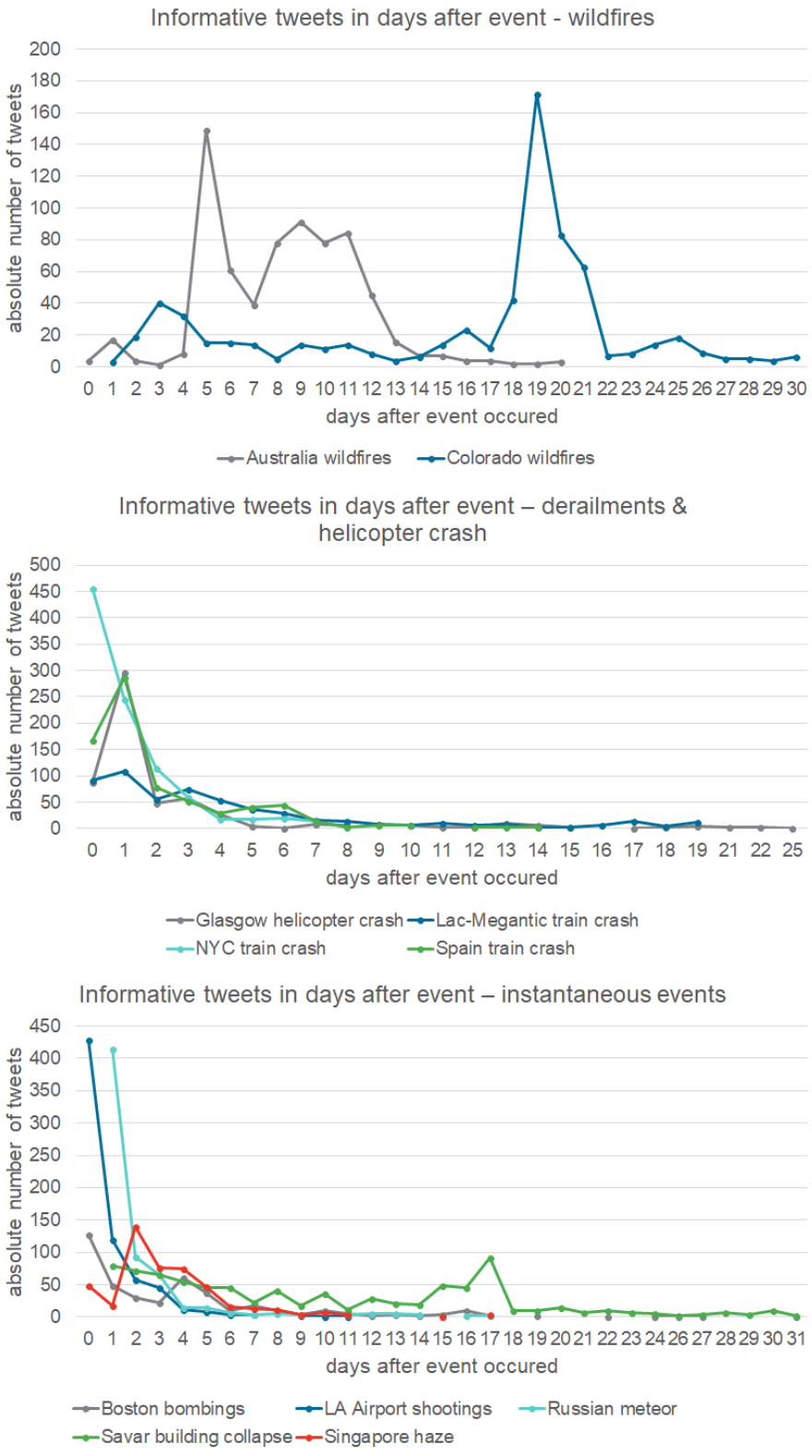


Figure 4.6: Informative tweets per event type (2)

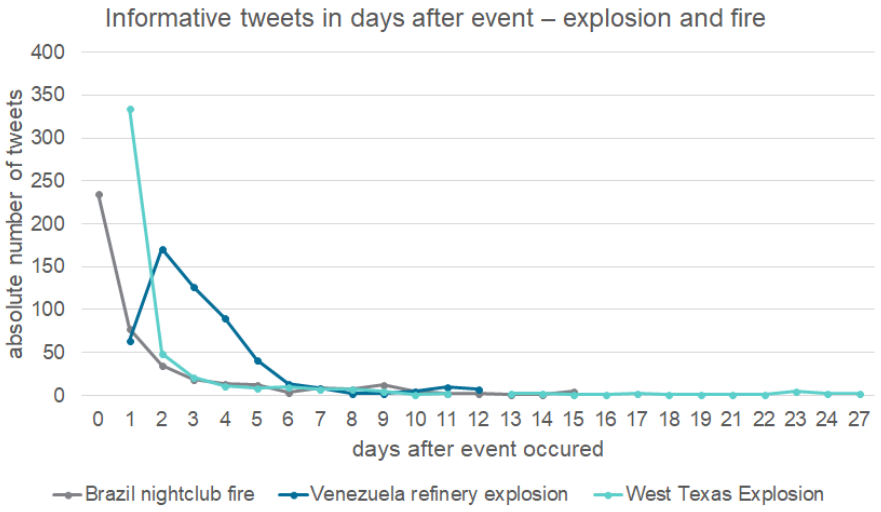


Figure 4.7: Informative tweets per event type (3)

1. Figure 4.10 visualizes the relation of informative tweets to non informative tweets over time. Overall, this relation stays constant indicated by the dashed gray line. Thus, response time itself, i.e., the period elapsed since the event started and the tweet was sent, is not highly informative for classification.

2. Considering differences in informativeness over time with respect to the *other three analysis dimensions*, only the source dimension shows some peculiarities. While for all events ”media”, ”government” and ”NGOs” tend to be much more informative over the entire time period independent of the disaster type, this is not true for sources, ”business”, ”eyewitness” and ”outsiders”. ”Eyewitnesses” over all events, e.g., tend to share, interestingly, in early stages of a crisis more non informative than informative tweets, which turns around after a few days (cf. dashed yellow line in Figure 4.10). Since *response time* in combination with other dimensions, like source, shows differences with respect to informativeness, response time is therefore suited to be used as feature for informativeness classification.

3. Analysis shows *no considerable differences* with respect to informativeness between *different disasters types*, except expectable differences with respect to the amount of shared tweets over time between instantaneous disaster types where a majority of tweet communication takes place in the first days and progressive disaster types, where communication is more constant over the entire crisis or correlates with particular occurrences within the event itself, e.g., a rising water level in case of a flood.

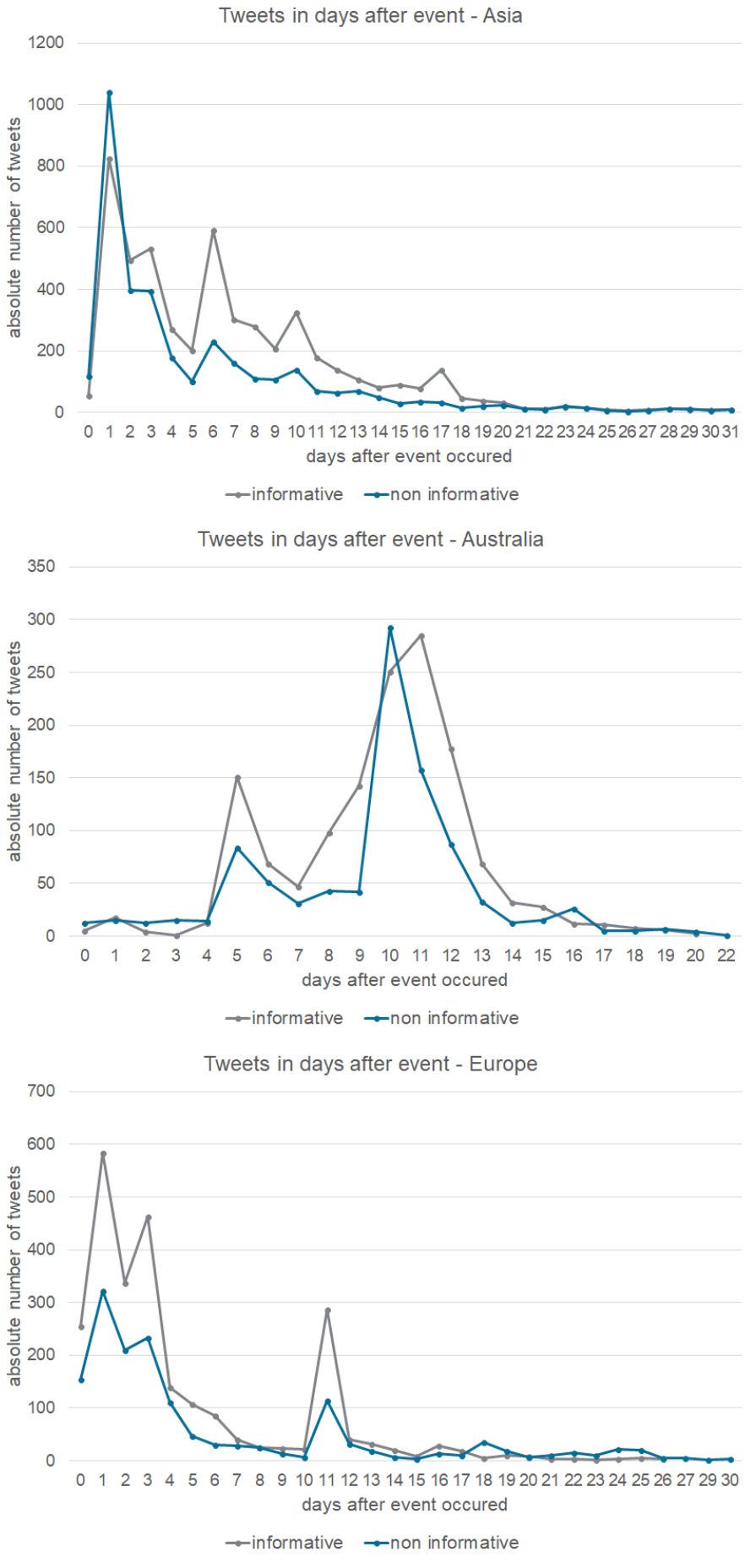


Figure 4.8: Informativeness per geographic location (1)

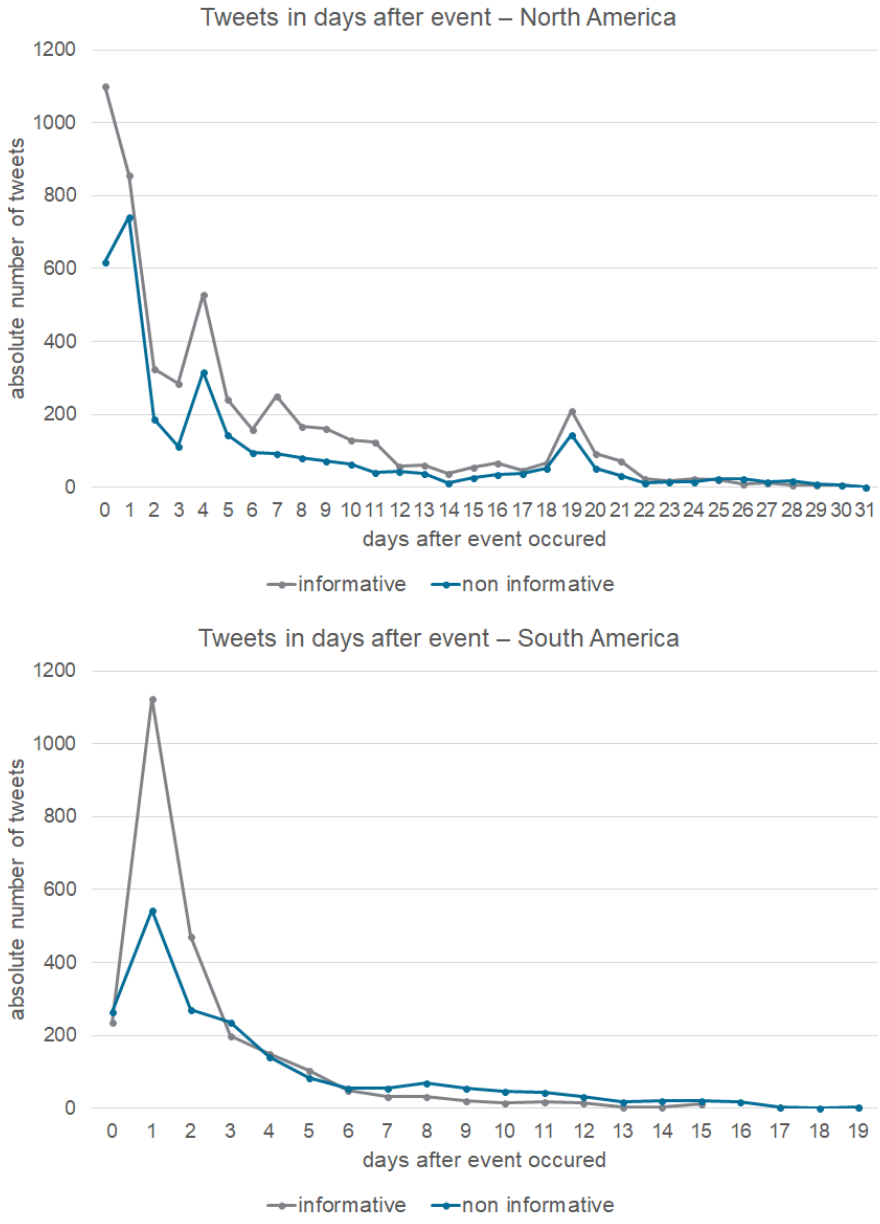


Figure 4.9: Informativeness per geographic location (2)

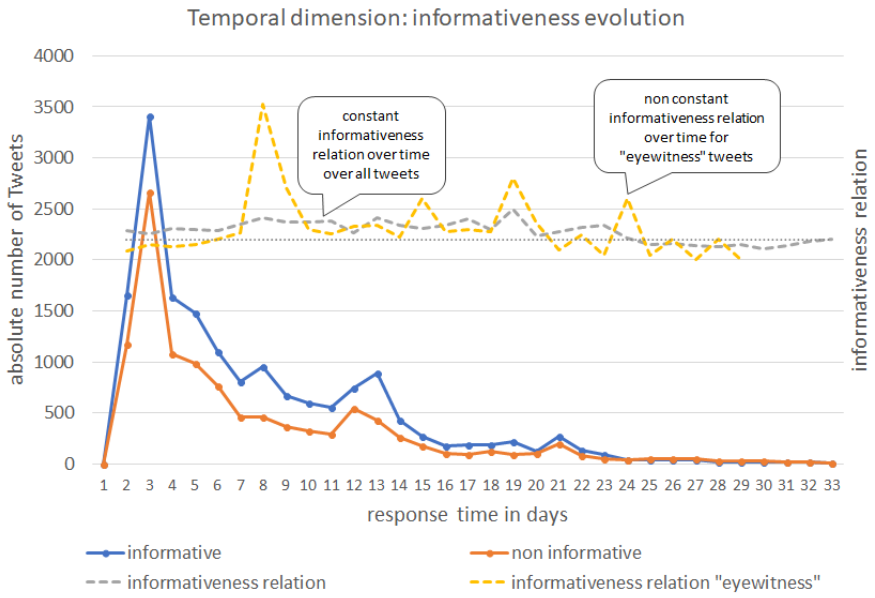


Figure 4.10: Temporal - informativeness evolution

4.2.2 Spatial dimension

By considering the spatial dimension of tweets this thesis wants to analyze whether there exist differences in informativeness of tweets with respect to the geographic location where the crisis event happened. Due to the geographic distribution of crisis events in the dataset, events were grouped together at a continent level (Asia, Europe, Australia, North- and South-America) based on their country. In the following, most important aspects are discussed in more detail.

4.2.2.1 Informativeness per geographic locations

Grouping events based on the country show whether there are differences in informativeness towards geographic location of the event. Figure 4.11 shows the informativeness in events grouped per continent. However there are no significant differences in informativeness between geographic location. Western countries, i.e., events in Europe and North America have a slightly higher proportion of informative tweets.

To conclude, informativeness with respect to geographic locations does not show any obvious patterns.

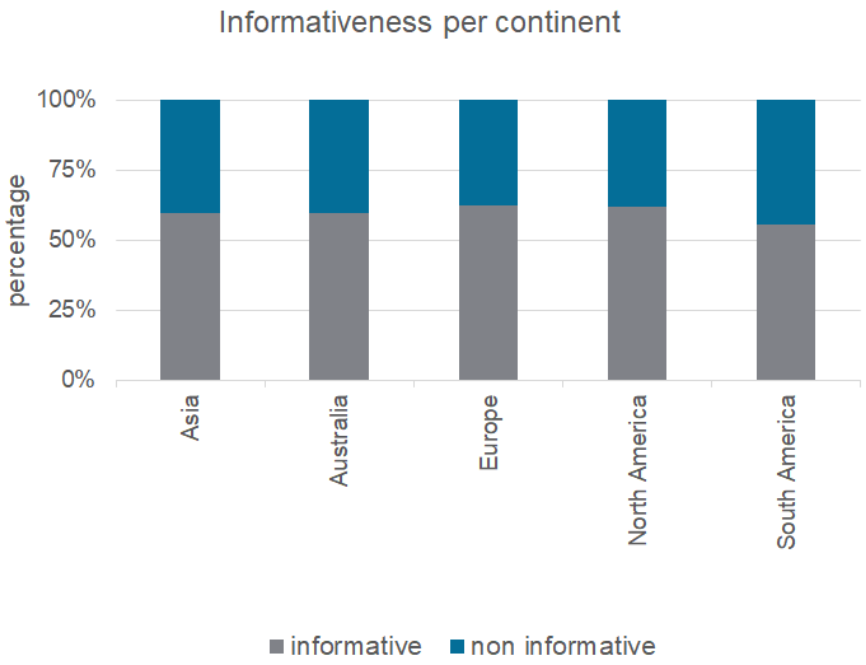


Figure 4.11: Informativeness per geographic locations

4.2.2.2 Findings

Main main findings towards the spatial dimension can be summarized as follows:

1. The overall relation between informative and non informative tweets in disaster events is *similar* across all continents although there is little variation between single events (cf. Figure 4.12). However, the spatial dimension does not provide additional information with respect to informativeness and thus is not used for classification.

2. There are *no considerable differences* between disaster events and between disasters types within one continent, which as a consequence support cross-domain classification.

4.2.3 Linguistic dimension

By considering the linguistic dimension of tweets this thesis wants to analyze which linguistic characteristics differentiate informative tweets from non informative ones. Since naturally a variety of linguistic characteristics exist, we focus on those ones which have been already used for classification in other domains such as news [Lloret and Palomar, 2016]. In particular, we analyze i) language, length and sentiment of a tweet, ii) Part-of-Speech (POS) information covering nouns, verbs, adverbs and adjectives, iii) frequency of spe-

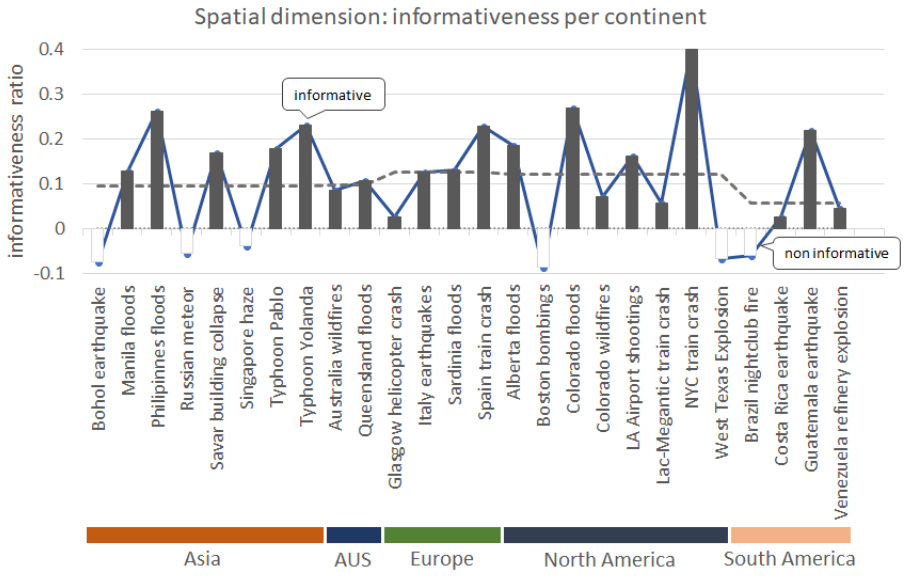


Figure 4.12: Spatial - informativeness per continent

cial characters, Emoticons and crisis-related hashtags and finally iv) punctuation. Since a majority of tweets, 71%, are in English, POS annotations and sentiment analysis are based on the English language. In order to analyze linguistic characteristics some of them require pre-processing steps, like Tokenization, removing stop-words, and Stemming. In the following, most important aspects are discussed in more detail.

4.2.3.1 Language

Language may play an important role for extracting, creating and selecting linguistic features, mainly in a pre-processing step. Using different words, a different grammatical structure in the tweet text can possibly bias classification results. Therefore, this section takes a look on used languages in the dataset. A majority of tweets from the entire dataset are written in English (EN), however there are some proportions of Tagalog (TL), which is a common language used in the Philippines, Spanish (ES) and Portuguese (PT) as well as very small portions of various other languages, included. Since the top four languages automatically determined in the dataset, EN, ES, TL, PT, cover over 90% of tweets, the focus lies on these four languages. Figure 4.13 shows the languages included in the dataset. To assess the language of the tweet text the python package "langdetect" [4] was used. Since the language is based on the tweet text, the remaining less than 15% of tweets are those tweets of other languages, 9.50%, as well as tweets which can not be surely recognized as one of the top 4 languages but with a probability of less than 0.75% be still from one of these top 4 languages. Table 4.2 shows the proportion of used languages in detail.

[4]https://pypi.python.org/pypi/langdetect

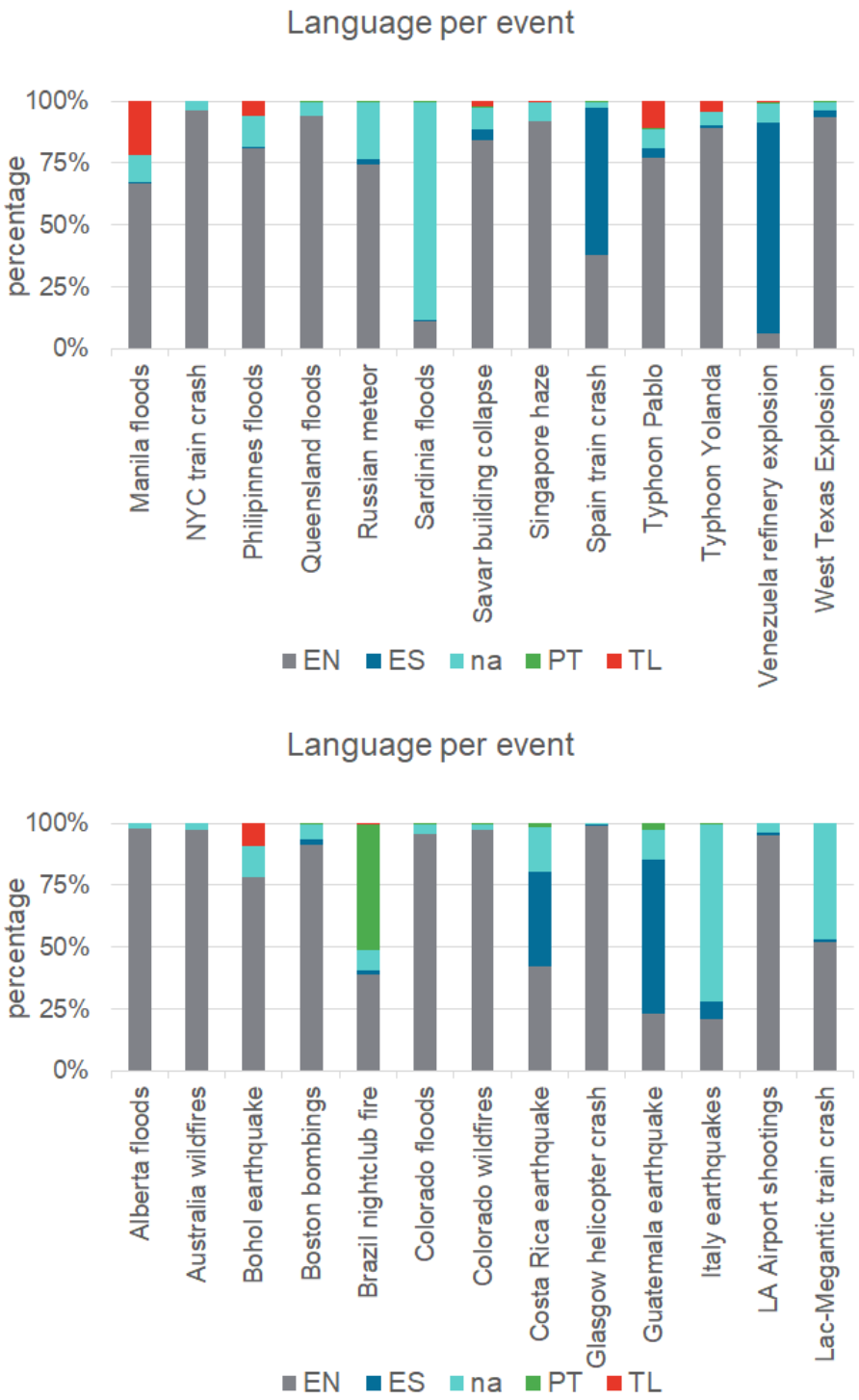


Figure 4.13: Language per event

Language	%
EN	71.08%
ES	10.42%
PT	2.16%
TL	2.01%
na	14.33%

Table 4.2: Dataset languages

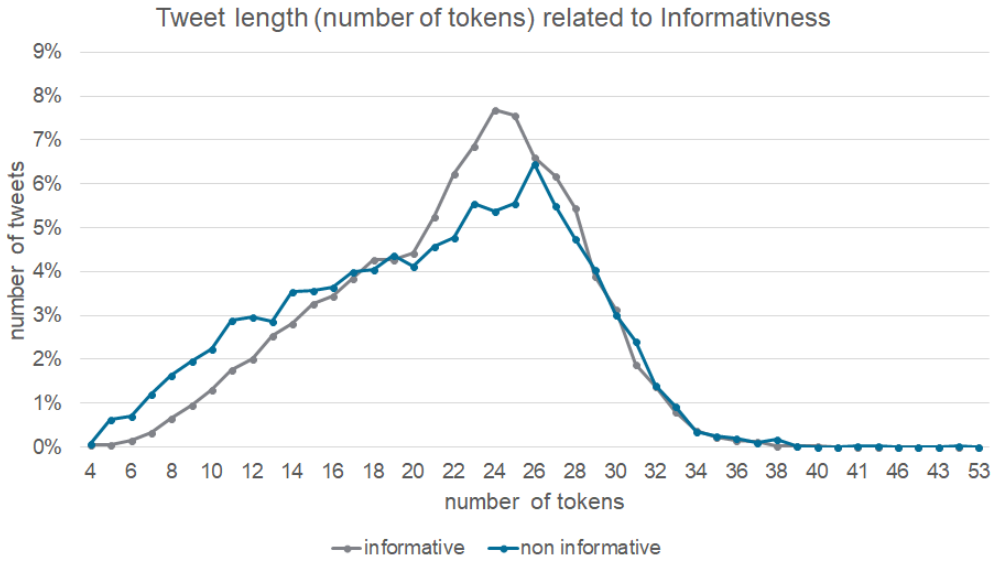


Figure 4.14: Tweet text length

To conclude, top four used languages are English, Spanish, Portuguese and Tagalog. More than 70% of all tweets are written in English. Since a majority of tweets are written in English and all 26 events contain at least portions of English tweets, pre-processing and extracting of features for classification can primary focus on the English language.

4.2.3.2 Tweet text length

Figure 4.14 shows the tweet length in number of tokens to visualize potential differences in tweet length regarding the informativeness classes. Yet, Chart 4.14 does not show significant differences in tweet length between informative and non informative tweets over all events. A more detailed view on certain events gives Table 4.3. Overall, non informative tweets tend to be slightly shorter than informative tweets. The average informative tweet is 21.13 tokens long, the average non informative tweet 19.62 tokens. Non informative tweets in other languages than English, like ES, TL and PT, are even shorter.

	Informative		**Not informative**	
	mean	**st.dev.**	**mean**	**st.dev.**
Total	**21.13**	**7.22**	**19.62**	**8.31**
EN	21.41	7.04	20.42	7.73
ES	20.31	7.53	18.78	9.47
TL	21.59	7.24	17.88	7.66
PT	20.21	8.16	16.03	10.18
Business	23.05	8.17	23.97	6.25
Eyewitness	22.50	7.37	20.32	7.98
Government	21.88	7.26	22.80	6.96
Media	20.88	6.67	20.08	7.99
NGO	23.56	7.84	21.63	8.47
Outsiders	20.45	7.93	19.39	8.56
Instantaneous	**20.59**	**7.08**	**19.17**	**8.59**
Human-induced	**20.78**	**7.02**	**19.92**	**8.76**
Boston bombings	21.71	6.26	19.80	7.95
Brazil nightclub fire	21.00	7.38	18.51	9.79
Glasgow helicopter crash	21.44	6.36	21.08	7.19
LA Airport shootings	20.98	6.93	21.01	8.96
Lac-Megantic train crash	19.37	8.04	18.00	9.81
NYC train crash	21.13	6.63	21.01	7.96
Savar building collapse	20.24	6.82	21.65	7.65
Spain train crash	20.26	7.50	17.35	10.36
Venezuela refinery explosion	20.19	7.18	20.87	9.06
West Texas Explosion	22.06	6.68	20.14	8.03
Natural	**20.20**	**7.19**	**17.97**	**8.16**
Bohol earthquake	18.81	7.83	17.67	7.77
Costa Rica earthquake	20.81	7.44	18.67	7.53
Guatemala earthquake	20.55	7.23	20.05	8.19
Italy earthquakes	20.78	6.27	19.75	7.81
Russian meteor	19.46	7.05	15.99	8.64
Progressive	**21.81**	**7.33**	**20.38**	**7.76**
Singapore haze	21.23	6.65	21.34	7.90
Natural	**21.85**	**7.37**	**20.23**	**7.73**
Alberta floods	22.69	7.14	22.13	7.44
Australia wildfires	22.31	7.25	21.96	7.74
Colorado floods	20.35	6.99	21.53	7.70
Colorado wildfires	20.70	6.64	20.67	7.17
Manila floods	22.81	8.35	19.00	7.95
Philipinnes floods	24.89	6.31	20.49	7.80
Queensland floods	21.85	6.48	18.82	7.08
Sardinia floods	20.37	7.66	19.22	7.91
Typhoon Pablo	22.05	6.51	19.11	8.02
Typhoon Yolanda	20.42	8.73	19.60	7.97

Table 4.3: Mean and standard deviation of informativeness classes

Figures 4.15 and 4.16 provide the results of Table 4.3 in graphical form. Wider gaps in the chart between the informative and non informative line mean there is a difference in tweet length between informative and non informative tweets.

To conclude, informative tweets tend to be longer, on average 1.51 tokens, than non informative tweets.

4.2.3.3 Part of Speech information of tweets

In Natural Language Processing (NLP), Part of Speech (POS) information is commonly used for information processing tasks in general or pre-processing of features in Machine Learning, respectively. Part of Speech taggers annotate tokens as part of a sentence with additional lexical information. Nouns, verbs, adjectives or adverbs are examples for POS annotations. POS information of sentences used in tweets might be different between informative tweets or not informative tweets. For instance, more nouns may correlate with more facts and furthermore, tend to be more informative, whereas more adjectives or adverbs are an indicator for emotion, which tend to not informative information. POS information is usually used as features in classification approaches.

Part of Speech taggers are based on language rules, thus, this thesis only uses events containing more than 50% of tweets in English for analyzing POS information. In addition, out of those events all tweets other than English are filtered out. Otherwise this may lead to biased results. POS tagging was implemented by using the "Natural Language Toolkit"[4] to identify nouns, verbs, adjectives and adverbs in tweet sentences.

Table 4.4 shows the resulting POS information based on the mean and separated by informativeness classes. Overall, some observations can be made from these results. On average, informative tweets contain slightly more nouns and adjectives. In contrast, non informative tweets have a tendency to contain more verbs and adverbs on average. This is even more the case for eyewitness tweets where an informative tweet contains on average 2.69 more nouns than an average non informative tweet. This fact is clearly visible in Figure 4.18. Differences visualized in grey stand for a higher mean in the informative class, difference bars visualized in white stand for a higher mean in the informative class.

In Table 4.4, events with less than 50% tweets in English are replaced with "n.a" instead of values.

To conclude, informative tweets tend to contain more nouns and adjectives, whereas non informative tweets tend to contain more verbs and adverbs. As a consequence, POS

[4] https://www.nltk.org/index.html

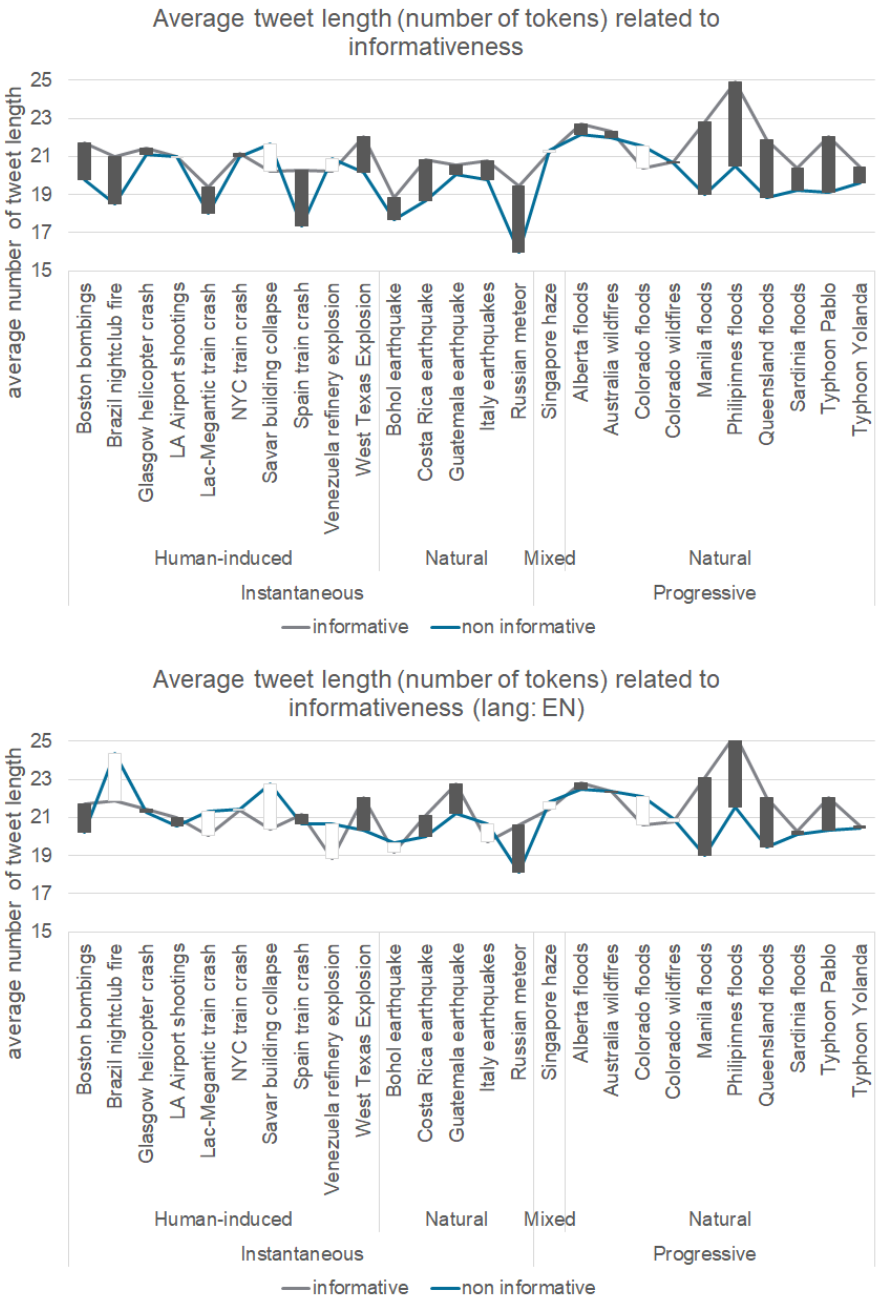


Figure 4.15: Average tweet length per event

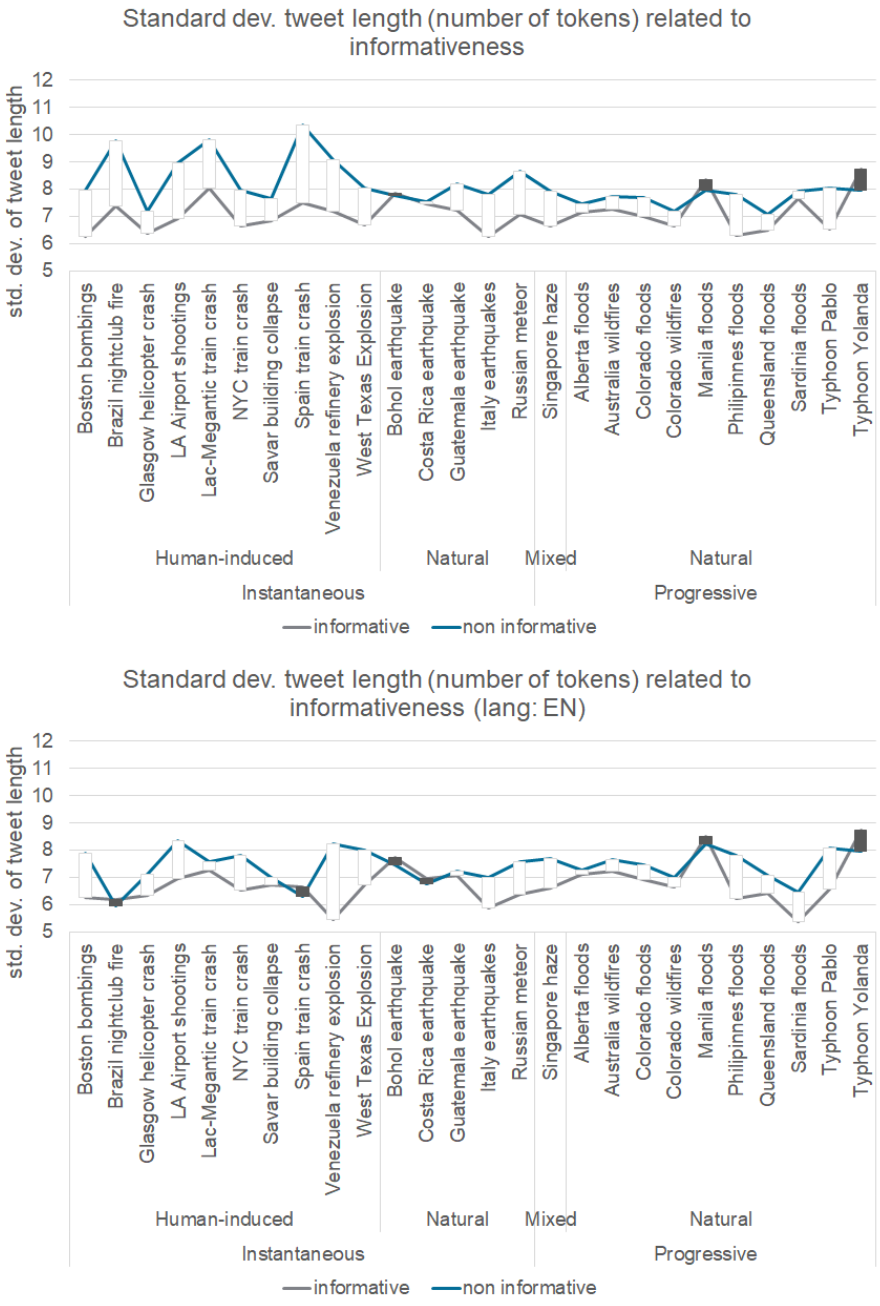


Figure 4.16: Standard deviation tweet length per event

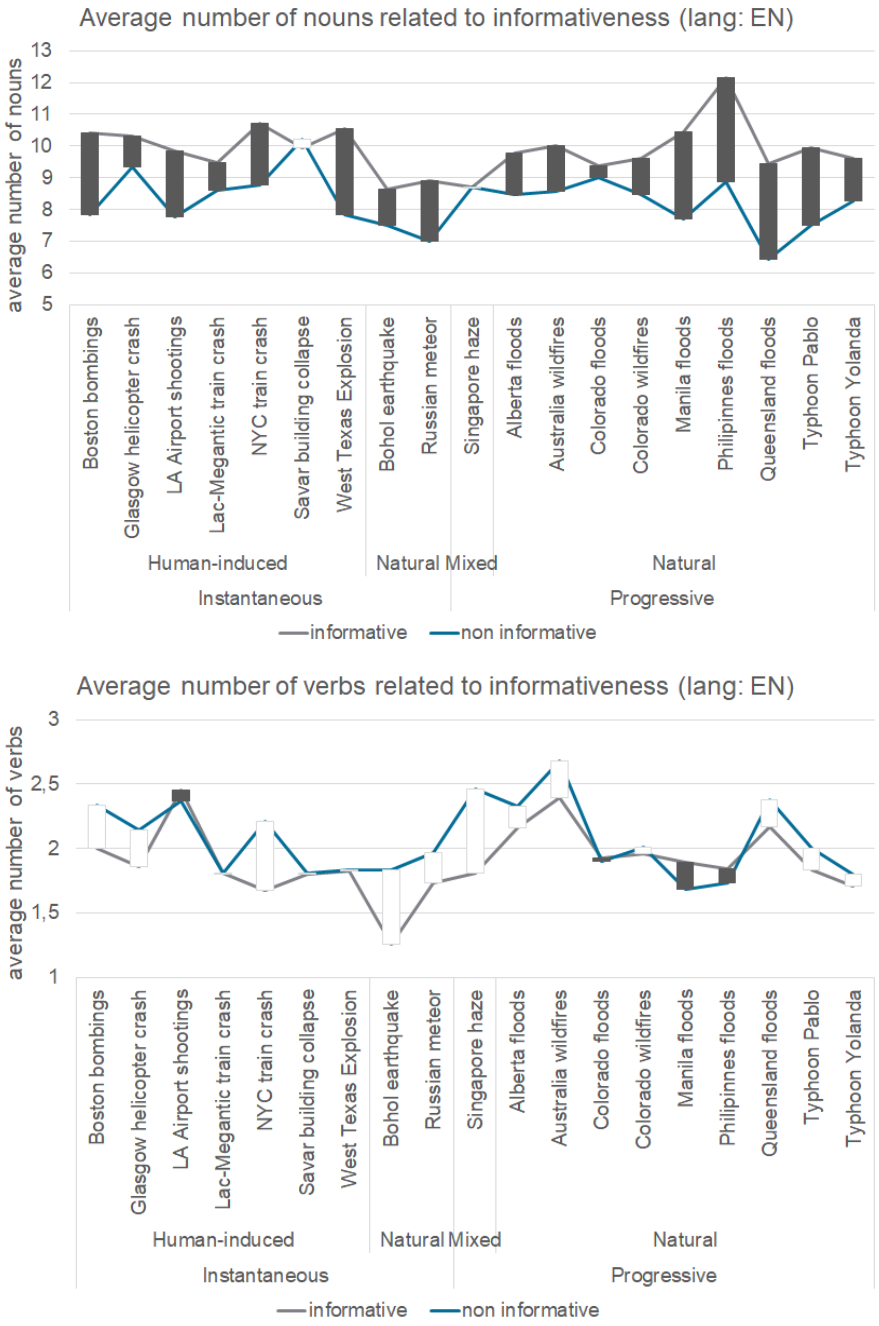


Figure 4.17: Mean of Part-of-Speech (1)

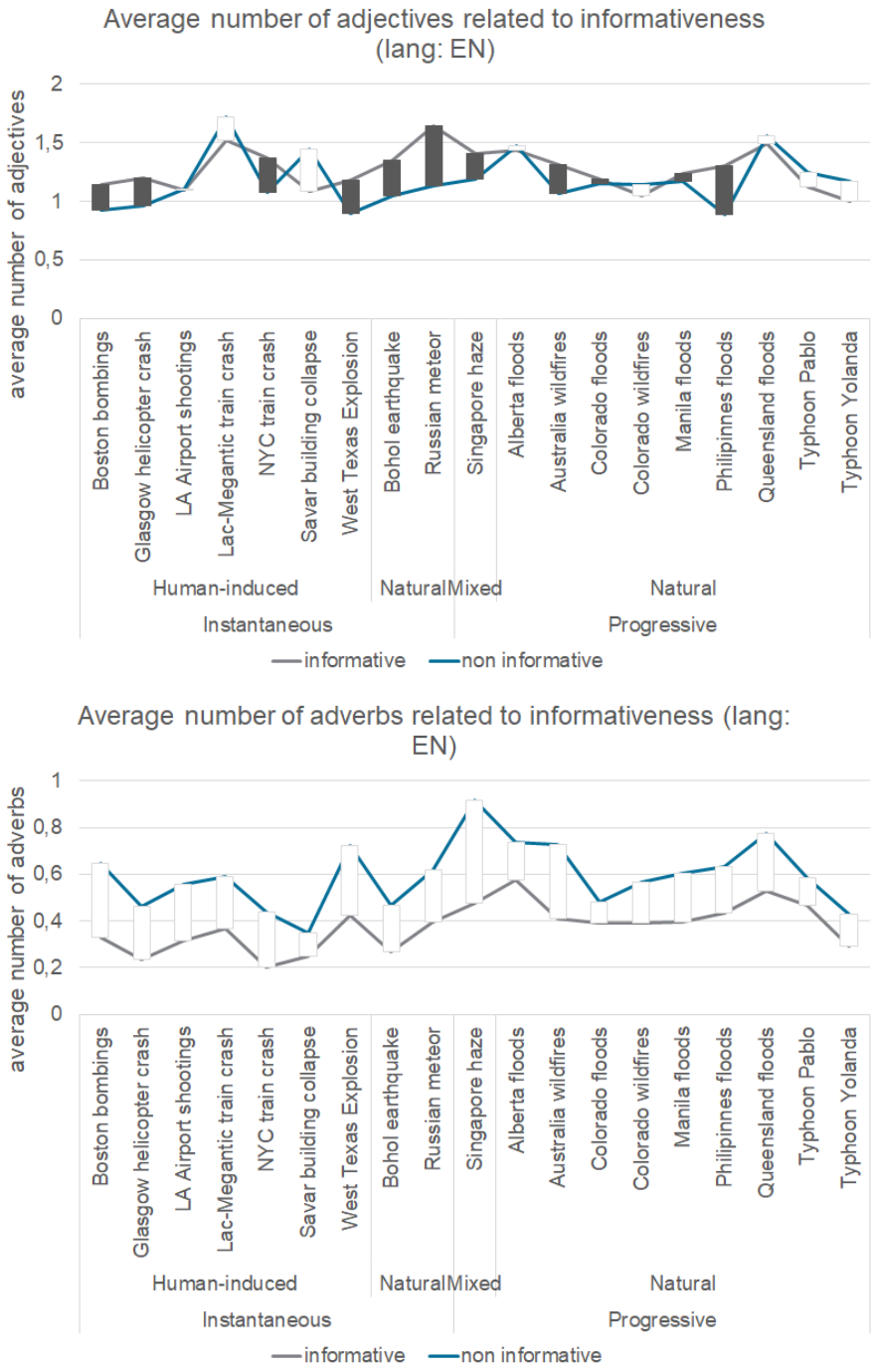


Figure 4.18: Mean of Part-of-Speech (2)

	Informative				Not informative			
	Noun	**Verb**	**Adj**	**Adv**	**Noun**	**Verb**	**Adj**	**Adv**
Total	**9.94**	**1.89**	**1.26**	**0.37**	**8.10**	**2.03**	**1.14**	**0.60**
Business	10.34	2.18	1.24	0.37	10.72	2.12	1.05	0.49
Eyewitness	10.37	1.97	1.24	0.66	7.68	2.48	1.32	0.93
Government	10.35	1.88	1.22	0.37	9.87	2.30	1.29	0.52
Media	10.12	1.75	1.29	0.29	9.54	1.88	1.20	0.42
NGO	10.64	2.13	1.17	0.40	9.37	2.35	1.11	0.64
Outsiders	9.00	2.09	1.20	0.43	7.57	2.17	1.02	0.62
Instantaneous	**9.94**	**1.82**	**1.27**	**0.29**	**8.07**	**1.95**	**1.06**	**0.54**
Human-induced	**10.20**	**1.90**	**1.22**	**0.28**	**8.50**	**2.03**	**1.07**	**0.55**
Boston bombings	10.41	1.99	1.13	0.33	7.78	2.29	0.91	0.63
Brazil nightclub fire	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Glasgow helicopter crash	10.32	1.86	1.20	0.24	9.29	2.12	0.96	0.46
LA Airport shootings	9.83	2.45	1.10	0.31	7.83	2.33	1.09	0.54
Lac-Megantic train crash	9.41	1.75	1.52	0.35	8.43	1.73	1.77	0.56
NYC train crash	10.68	1.66	1.37	0.20	8.78	2.20	1.07	0.44
Savar building collapse	9.94	1.79	1.08	0.25	10.23	1.74	1.42	0.34
Spain train crash	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Venezuela refinery explosion	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
West Texas Explosion	10.56	1.81	1.18	0.42	7.85	1.80	0.90	0.71
Natural	**8.80**	**1.48**	**1.49**	**0.33**	**7.03**	**1.73**	**1.04**	**0.51**
Bohol earthquake	8.57	1.23	1.33	0.27	7.24	1.62	0.98	0.42
Costa Rica earthquake	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Guatemala earthquake	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Italy earthquakes	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Russian meteor	8.98	1.67	1.60	0.38	6.87	1.82	1.09	0.57
Progressive	**9.94**	**1.94**	**1.25**	**0.42**	**8.14**	**2.12**	**1.22**	**0.66**
Singapore haze	8.67	1.79	1.41	0.48	8.69	2.40	1.18	0.88
Natural	**10.03**	**1.95**	**1.24**	**0.42**	**8.04**	**2.07**	**1.22**	**0.62**
Alberta floods	9.77	2.14	1.43	0.57	8.39	2.29	1.46	0.73
Australia wildfires	10.00	2.39	1.31	0.41	8.54	2.62	1.06	0.71
Colorado floods	9.31	1.90	1.18	0.38	8.91	1.85	1.15	0.46
Colorado wildfires	9.58	1.94	1.05	0.39	8.41	1.98	1.13	0.56
Manila floods	10.47	1.82	1.27	0.39	7.92	1.59	1.21	0.59
Philipinnes floods	12.26	1.74	1.32	0.40	8.77	1.58	0.89	0.57
Queensland floods	9.46	2.14	1.47	0.52	6.29	2.31	1.55	0.75
Sardinia floods	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Typhoon Pablo	10.01	1.79	1.11	0.46	7.55	1.82	1.26	0.56
Typhoon Yolanda	9.63	1.69	1.00	0.29	8.23	1.69	1.17	0.42

Table 4.4: Mean of POS values of informativeness classes

information is suitable to use as feature for informativeness classification since differences in informative and non informative tweets exist.

4.2.3.4 Characters used in tweets

The usage of special characters in tweets might be an additional indicator for informativeness of tweets. Table 4.5 lists the average frequency of characters in tweets and 4.20 visualizes the results of the special characters "#", "?", "!", URLs and "@". Over all events, there is a trend that the characters "?" and "!" are more used in non informative tweets, which may correlate with emotion of users. The "#" character is usually used as hashtag in tweets. There is a tendency towards the character "#" is more frequently used in non informative tweets. In contrast, there is a tendency that URLs and the "@" character are more used in informative tweets. This is maybe related to evidence, which is provided by an additional link in the tweet as well as tweet answers directly referred to users using the "@" character.

To conclude, there is a tendency that the characters "#", "!" and "?" are more frequently used in non informative tweets, whereas URLs and "@" are more frequently used in informative tweets. Again, this kind of information is suitable to use as feature since there exist differences with respect to informativeness of tweets.

4.2.3.5 Emoticons

Emoticons, e.g. ":)" or ":(", used in tweets may be an indicator for emotion of users. Studies showed that Emotions are more frequent in non informative tweets [Acerbo and Rossi, 2017]. The same conclusion can be drawn from the current data analysis. Figure 4.21 show the frequency of positive and negative Emoticons used in tweet text. Over all events, there is a tendency that in general more negative Emotions are used. However, there are differences regarding the event. In general, when considering differences between the informativeness classes, Emoticons are used more frequently in non informative tweets.

To conclude, Emoticons are more used in non informative tweets. Thus, existing Emoticons in tweets can be used as feature for classifying tweets.

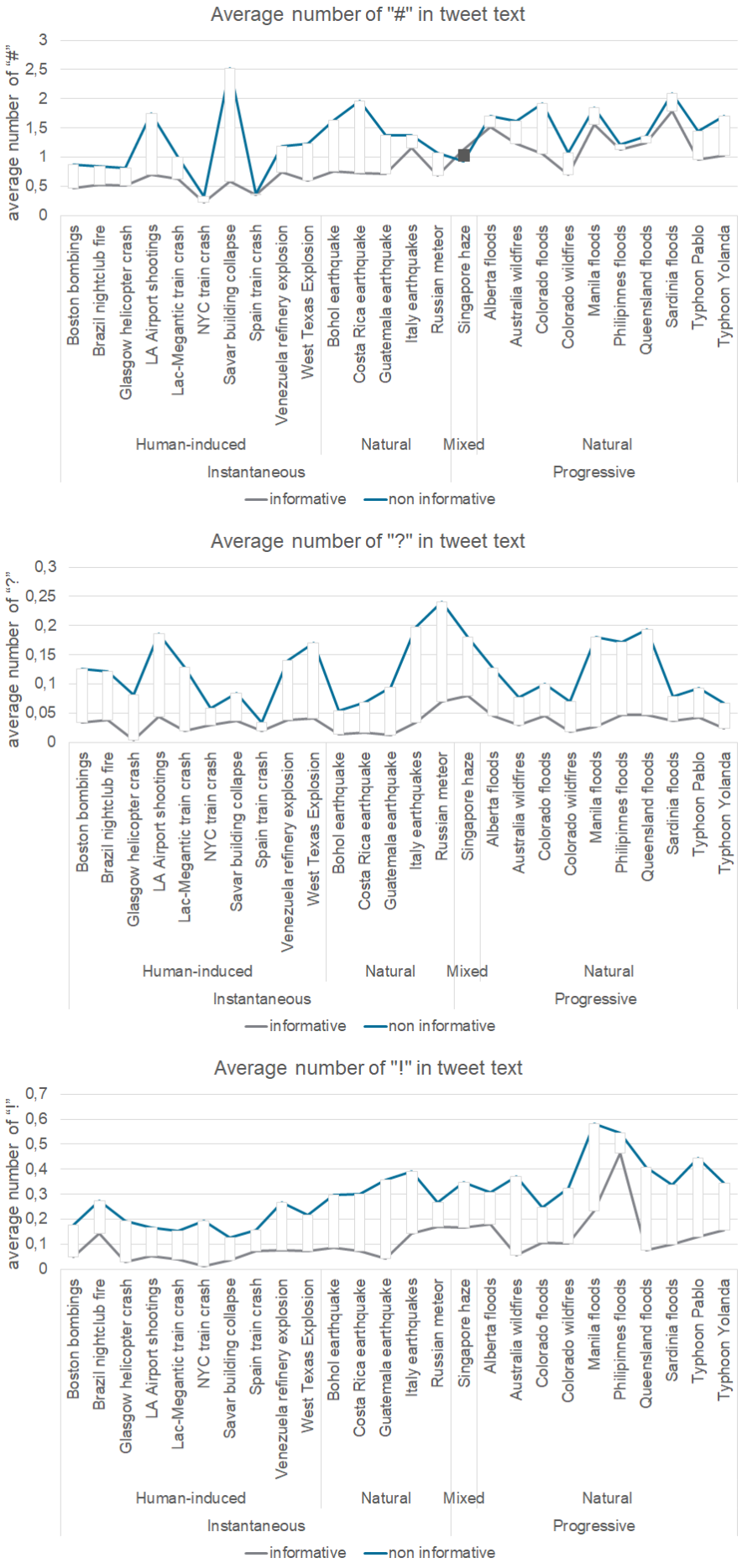


Figure 4.19: Average characters in tweet text (1)

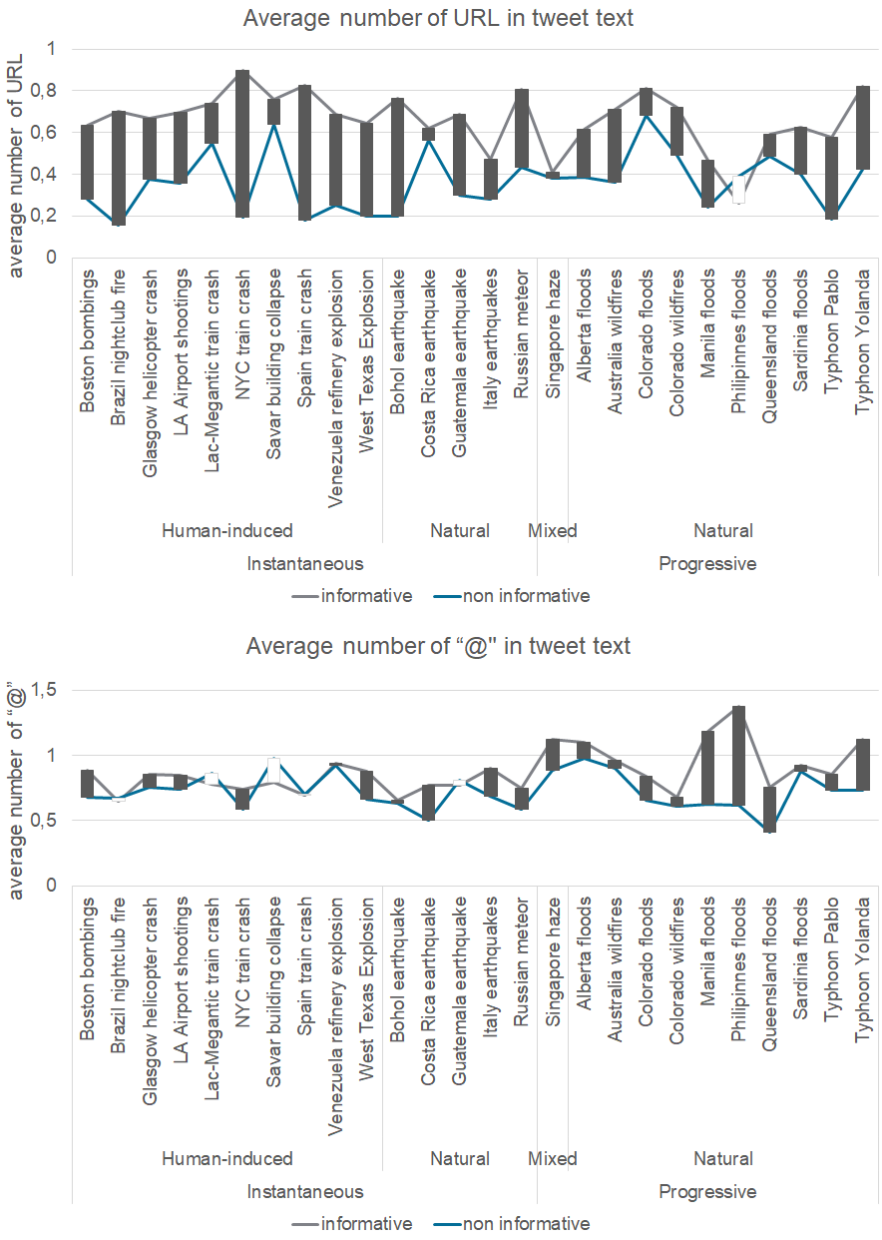


Figure 4.20: Average characters in tweet text (2)

	#		?		!		URL		@	
	Inf.	**Not**	**Inf.**	**Not**	**Inf.**	**Not**	**Inf.**	**Not**	**Inf.**	**Not**
Total	**0.87**	**1.08**	**0.03**	**0.13**	**0.11**	**0.29**	**0.67**	**0.26**	**0.88**	**0.72**
Business	1.16	0.87	0.02	0.08	0.17	0.18	0.52	0.38	1.07	1.24
Eyewitness	1.28	1.15	0.05	0.12	0.40	0.43	0.42	0.27	1.04	0.64
Government	1.21	1.44	0.02	0.07	0.05	0.09	0.54	0.39	0.92	1.03
Media	0.71	1.11	0.02	0.10	0.05	0.15	0.80	0.68	0.85	0.98
NGO	1.44	1.46	0.04	0.11	0.21	0.46	0.55	0.48	1.14	1.03
Outsiders	0.89	1.05	0.07	0.13	0.16	0.28	0.52	0.18	0.81	0.68
Instantaneous	**0.61**	**1.27**	**0.03**	**0.13**	**0.07**	**0.24**	**0.72**	**0.34**	**0.79**	**0.70**
Human-induced	**0.51**	**1.14**	**0.03**	**0.12**	**0.05**	**0.20**	**0.74**	**0.32**	**0.80**	**0.76**
Boston bombings	0.47	0.87	0.03	0.13	0.05	0.18	0.63	0.28	0.89	0.68
Brazil nightclub fire	0.52	0.84	0.04	0.12	0.14	0.27	0.70	0.16	0.65	0.67
Glasgow helicopter crash	0.51	0.81	0.00	0.08	0.03	0.19	0.67	0.38	0.86	0.75
LA Airport shootings	0.70	1.75	0.04	0.19	0.05	0.17	0.70	0.36	0.85	0.74
Lac-Megantic train crash	0.63	1.00	0.02	0.13	0.04	0.15	0.74	0.55	0.77	0.86
NYC train crash	0.22	0.33	0.03	0.06	0.01	0.19	0.90	0.19	0.74	0.59
Savar building collapse	0.58	2.52	0.04	0.08	0.03	0.13	0.76	0.64	0.79	0.98
Spain train crash	0.36	0.37	0.02	0.03	0.07	0.15	0.83	0.18	0.70	0.70
Venezuela refinery explosion	0.74	1.19	0.04	0.14	0.08	0.27	0.69	0.25	0.94	0.92
West Texas Explosion	0.59	1.23	0.04	0.17	0.07	0.22	0.64	0.20	0.88	0.66
Natural	**0.80**	**1.48**	**0.03**	**0.14**	**0.10**	**0.31**	**0.66**	**0.38**	**0.78**	**0.61**
Bohol earthquake	0.76	1.63	0.01	0.05	0.09	0.30	0.76	0.20	0.66	0.63
Costa Rica earthquake	0.73	1.96	0.02	0.07	0.07	0.30	0.62	0.56	0.77	0.50
Guatemala earthquake	0.71	1.37	0.01	0.09	0.04	0.36	0.69	0.30	0.77	0.81
Italy earthquakes	1.15	1.37	0.03	0.20	0.14	0.39	0.47	0.28	0.90	0.68
Russian meteor	0.68	1.07	0.07	0.24	0.17	0.27	0.81	0.43	0.75	0.59
Progressive	**1.20**	**1.49**	**0.04**	**0.12**	**0.16**	**0.38**	**0.61**	**0.40**	**0.99**	**0.73**
Singapore haze	1.12	0.93	0.08	0.18	0.17	0.35	0.41	0.38	1.12	0.89
Natural	**1.21**	**1.57**	**0.04**	**0.11**	**0.16**	**0.39**	**0.62**	**0.40**	**0.98**	**0.71**
Alberta floods	1.51	1.70	0.05	0.13	0.18	0.31	0.62	0.39	1.10	0.97
Australia wildfires	1.22	1.61	0.03	0.08	0.06	0.37	0.71	0.36	0.96	0.90
Colorado floods	1.06	1.92	0.05	0.10	0.11	0.25	0.81	0.68	0.84	0.66
Colorado wildfires	0.70	1.07	0.02	0.07	0.10	0.32	0.72	0.49	0.68	0.61
Manila floods	1.56	1.85	0.03	0.18	0.23	0.58	0.47	0.24	1.18	0.62
Philipinnes floods	1.12	1.21	0.05	0.17	0.47	0.54	0.26	0.39	1.37	0.62
Queensland floods	1.24	1.36	0.05	0.19	0.08	0.41	0.59	0.49	0.75	0.41
Sardinia floods	1.79	2.09	0.04	0.08	0.10	0.34	0.63	0.40	0.92	0.88
Typhoon Pablo	0.96	1.45	0.04	0.09	0.13	0.44	0.58	0.18	0.86	0.73
Typhoon Yolanda	1.03	1.71	0.02	0.07	0.16	0.34	0.82	0.42	1.12	0.73

Table 4.5: Average characters in tweet text

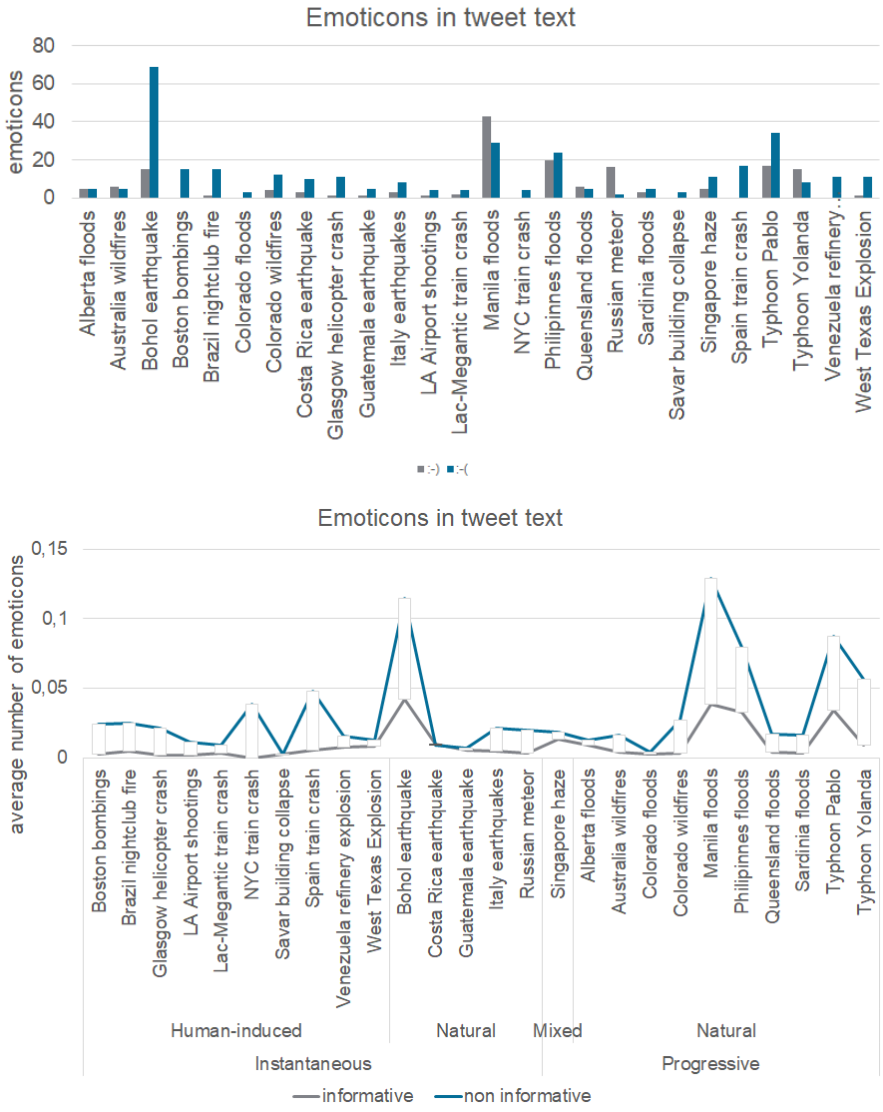


Figure 4.21: Emoticons in tweet text

4.2.3.6 Sentiment

More precise than only looking for Emoticons in tweet text is determining the sentiment of the tweet based on the language and terms used. Sentiment information is commonly used for text classification. To determine the sentiment of a tweet, the Python library Textblob[5] was used. Based on the language used in the tweet text, TextBlob provides a sentiment polarity and a sentiment subjectivity score. Positive values, in an interval from zero to one, express positive sentiment and negative values, in an interval from minus one to zero, express negative sentiment. The sentiment subjectivity score expresses the objectivity, i.e., an estimation for reliability, of the sentiment polarity score where values close to zero are very objective and values close to one are very subjective. To visualize the sentiment in an comparable way in Figure 4.22, a positive as well as a negative sentiment score was calculated:

$$postiveSentimentScore = Polarity * (1 - Subjectivity) \tag{4.1}$$

$$negativeSentimentScore = |Polarity| * (1 - Subjectivity) \tag{4.2}$$

Since sentiment analysis is based on language, again only tweets where more than 50% of tweets written in English are considered.

Figure 4.22 shows, all events contain tweets with portions of sentiment, yet, there are no distinguishable patterns between events. However, considering informativeness, there is a tendency that informative tweets contain less positive sentiment than not informative tweets, whereas considering negative sentiment there is a trend in the other way round, i.e., informative tweets contain more negative sentiment than not informative tweets. Over all, tweets contain almost twice as much positive sentiment than negative sentiment.

To conclude, there is a trend that informative tweets contain less positive and more negative sentiment. As a consequence, positive and negative sentiment are suitable feature candidates for informativeness classification.

4.2.3.7 Tweet hashtags

Figure 4.23 shows the average number of hashtags, e.g "#HurricaneSandy", which are related to the particular crisis event, used within the tweet text. These hashtags are compared against a defined set of hashtags, which were commonly used to address the particular crisis event. These set of hashtags was defined by the CrisisLexT26 dataset. In contrast to the character analysis of "#" in the section before, the current hashtag analysis

[5] http://textblob.readthedocs.io/en/dev/index.html

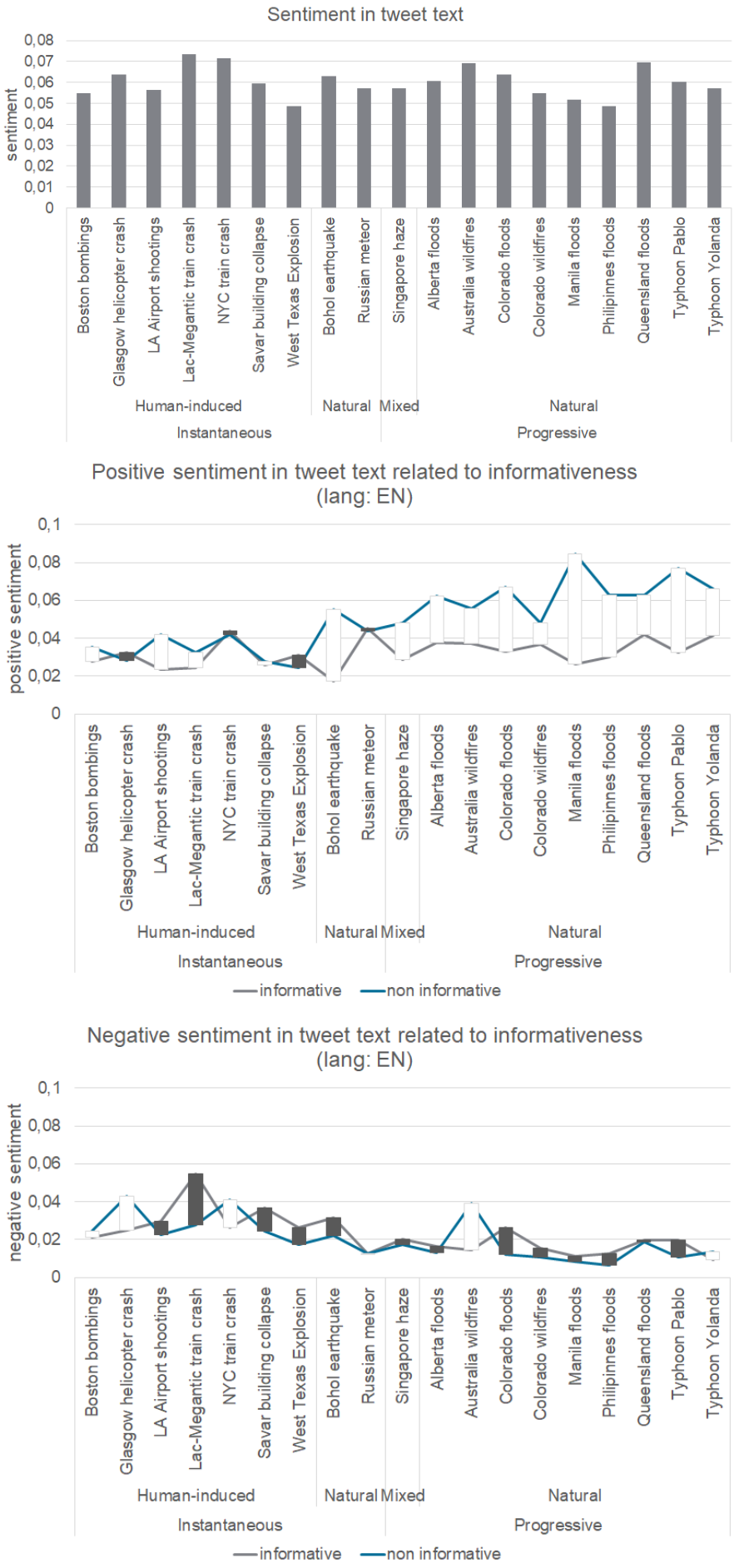


Figure 4.22: Sentiment in tweet text

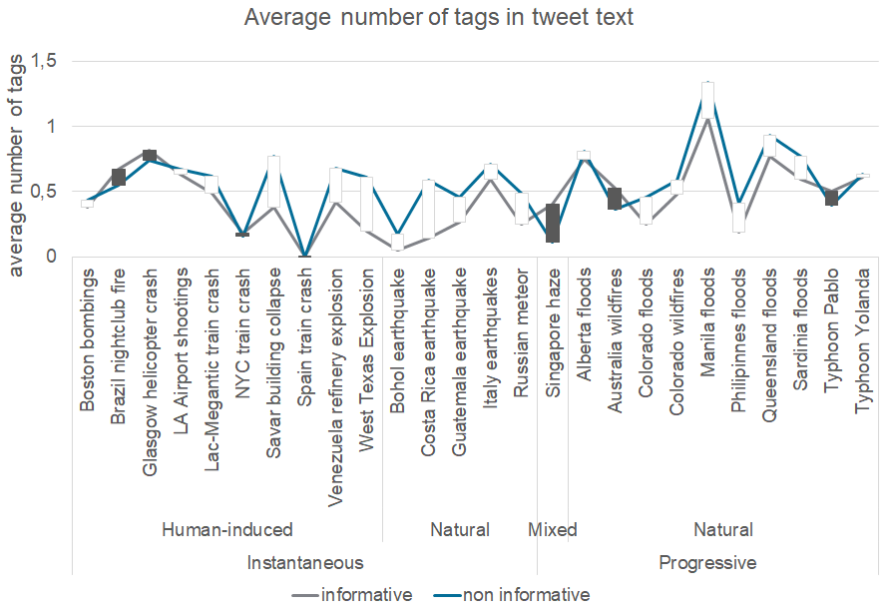


Figure 4.23: Tags used in tweets

considers the entire hashtag-string, which is also used by users to search for particular topics. There is a trend towards using less crisis related hashtags in informative tweets. Yet, this is not the case for all events.

To conclude, there is a tendency that less crisis related hashtags are used in informative tweets. Again, since there exist differences towards informativeness classes, hashtag information is suitable to use as feature for classification.

4.2.3.8 Negation in tweets

Terms which are associated with negation, for instance ”not”, ”none”, ”neither” or ”never”, might provide additional information regarding informativeness. A full list of negation words used is listed in Chapter 5. Figure 4.24 shows existing negation terms in both classes. Except in the Colorado wildfire event, more negation terms are used in non informative tweets.

To conclude, more negation terms are used in non informative tweets. As a consequence, negation on tweets is suitable to use as feature for informativeness classification.

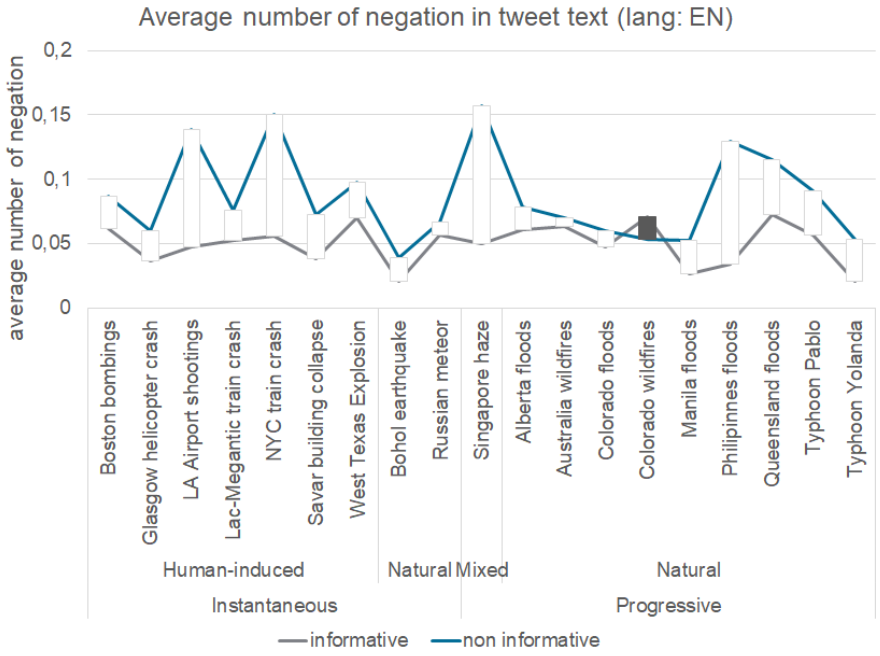


Figure 4.24: Negations used in tweet text

4.2.3.9 Findings

Our main findings towards the linguistic dimension are visualized in Figure 4.25 and can be summarized as follows:

Informative tweets tend to:

1. be longer, on average 1.51 tokens longer than non informative ones, thus indicating that tweet length could be a suitable classification feature.

2. contain more nouns and adjectives.

3. contain URLs and the character "@" more frequently.

4. contain less positive and more negative sentiment.

5. contain less crisis related hashtags.

Non informative tweets tend to:

1. be shorter.

2. contain more verbs and adverbs.

3. contain the characters "#", "!" and "?" more frequently.

4. contain more Emoticons.

5. contain more negation terms in tweet text, such as "no", "not" or "never".

6. finish with punctuation more likely.

However, overall *no considerable differences* between different events and different disaster types over all analyzed tweet characteristics exist.

4.2.4 Source dimension

The source of a tweet was labeled by Olteanu et al. [Olteanu et al., 2015] and the dataset contains tweets originating from source i) business, ii) eyewitness, iii) government, iv) media, v) NGOs, iv) outsiders, and tweets not applicable, which are grouped in our work as "others". Figure 4.26 shows the distribution of the information source classes in detail. A majority of tweets from the entire dataset, 36%, belong to the "Media" class. In contrast, only around 8% of tweets were shared by eyewitnesses, i.e. those who are directly affected by the crisis. In the following, most important aspects are discussed in more detail.

4.2.4.1 Informativeness per source

As Figure 4.28 shows there exist an unequal distribution of informativeness between different information sources, which allows some conclusions. First, tweets from the information source "media" strongly tend to be "informative". In contrast, tweets from the information source "outsiders" and "others" strongly tend to be "non informative". Second, eyewitness reports, which intuitively may be more informative, due to the reason that eyewitness reports from affected people are first-hand information, geographically nearby information, are almost balanced in class. 44.01% tweets of information type "eyewitness" belong to the "informative" class and 55.99% of eyewitness tweets belong to "not informative" class.

To conclude, informativeness of different sources vary significantly. Tweets shared by source media are significantly more informative than tweets shared by source outsiders. Hence, information source is important for informativeness classification and is suitable to use as feature for classification.

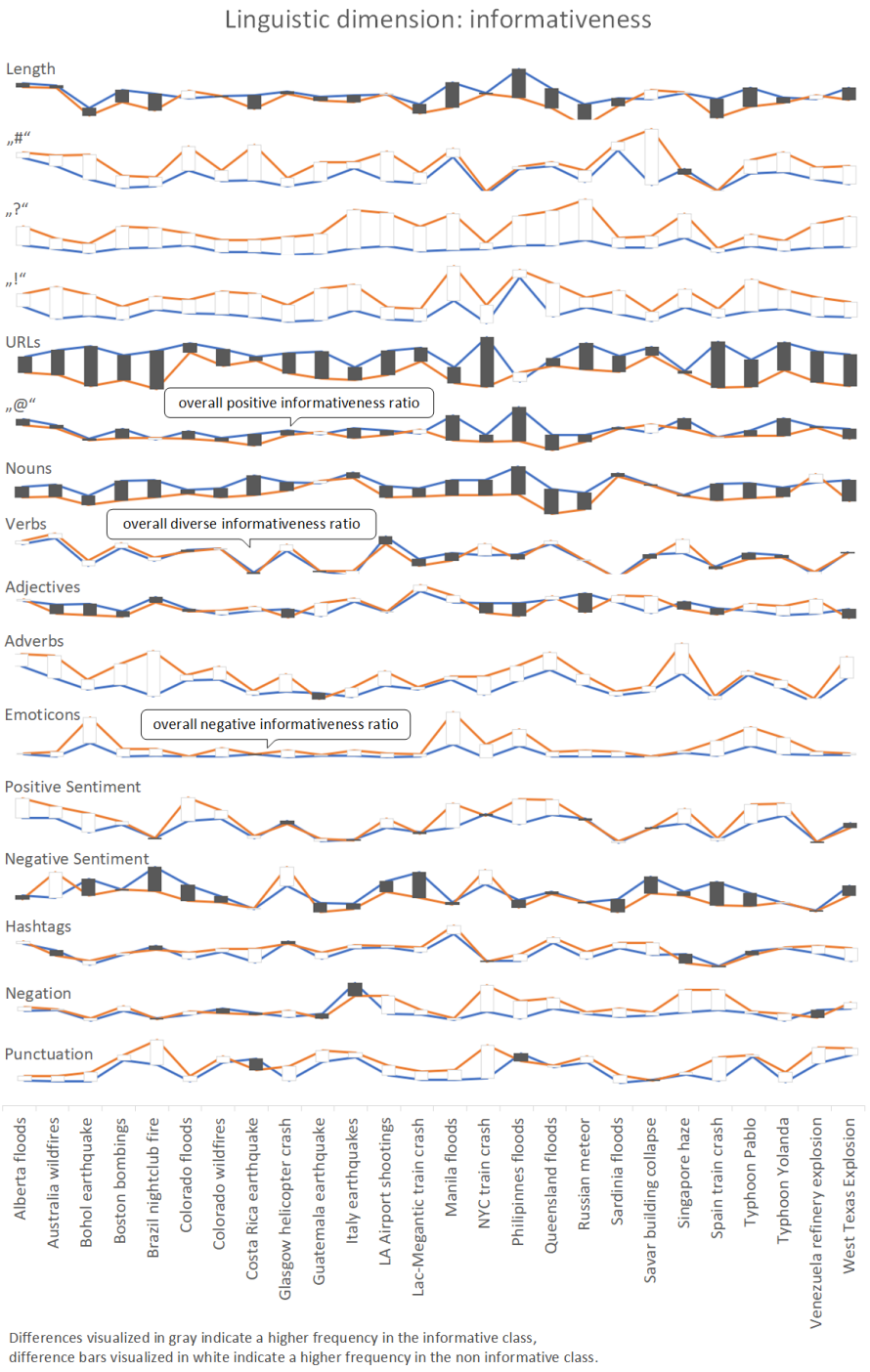


Figure 4.25: Linguistic - informativeness of characteristics

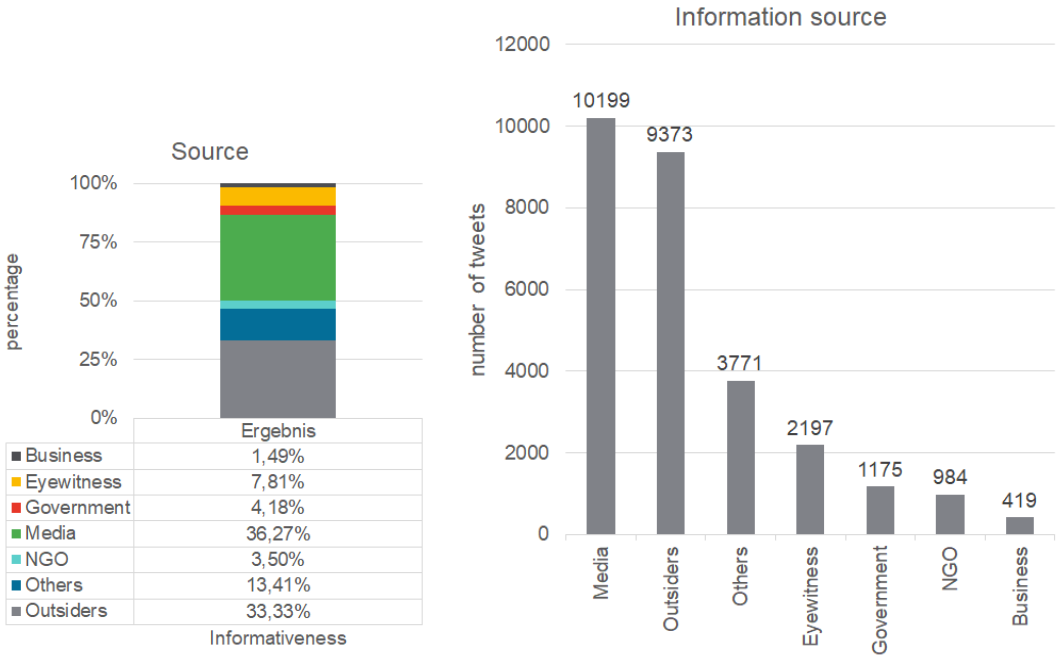


Figure 4.26: Information source distribution

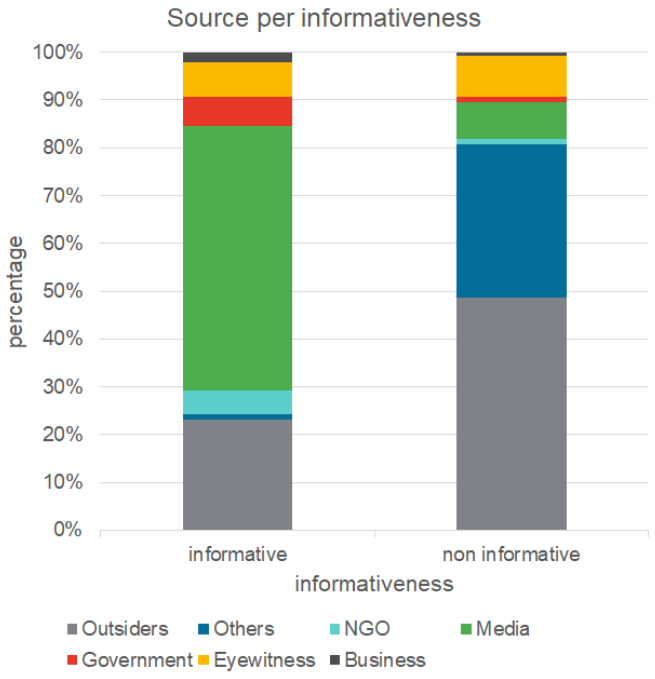


Figure 4.27: Informativeness per information source

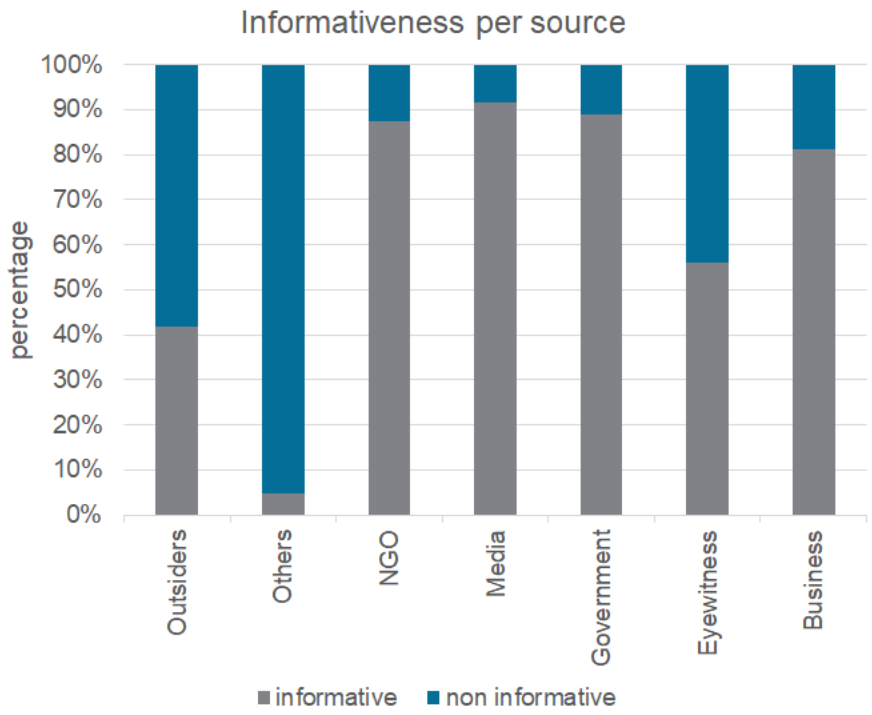


Figure 4.28: Informativeness per information source

4.2.4.2 Information source per event

Furthermore, Figure 4.29 shows the information source for each event individually. Some interesting observations can be drawn. Except the Alberta floods, events in the Asian area predominately tend to contain more eyewitness reports and less tweets from source media. For instance, Philippines floods and Singapore haze have significantly more eyewitness proportion. This fact may lead to worse classification results for these disaster events since tweets from source eyewitness are informative as well as not informative almost equally. However, differences regarding information source between events do not follow obvious visible patterns.

To conclude, there are no obvious patterns regarding information source over different events. However, events from the Asian area contain more eyewitness reports and less media information than others.

4.2.4.3 Information source per geographic location

Figures 4.32 and 4.32 show informativeness per geographic location for the particular information sources media and eyewitness. Correlations with the previous Figure set 4.30 are obvious and visible in the charts. One conclusion is that tweets from source labeled

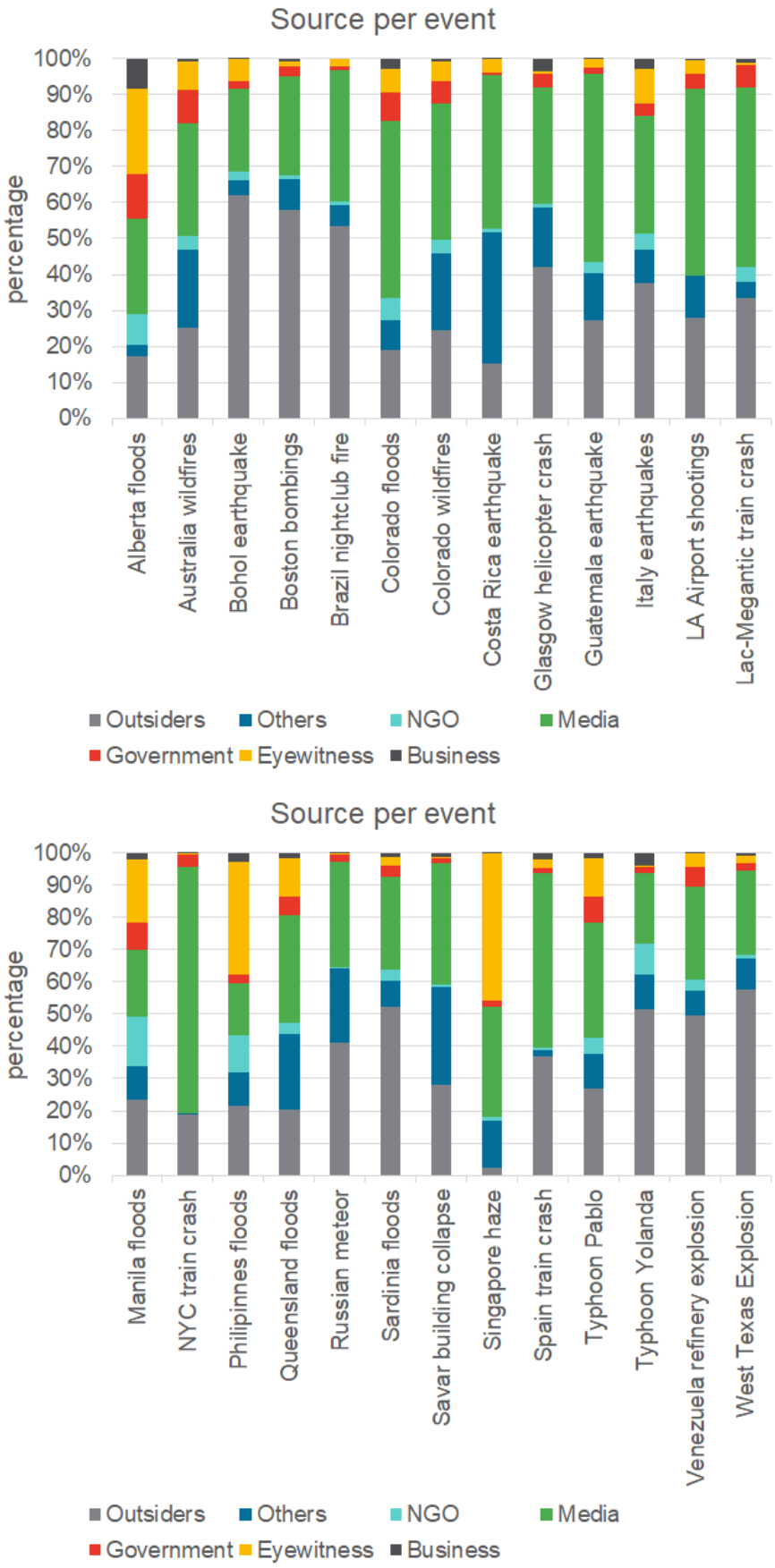


Figure 4.29: Information source per event

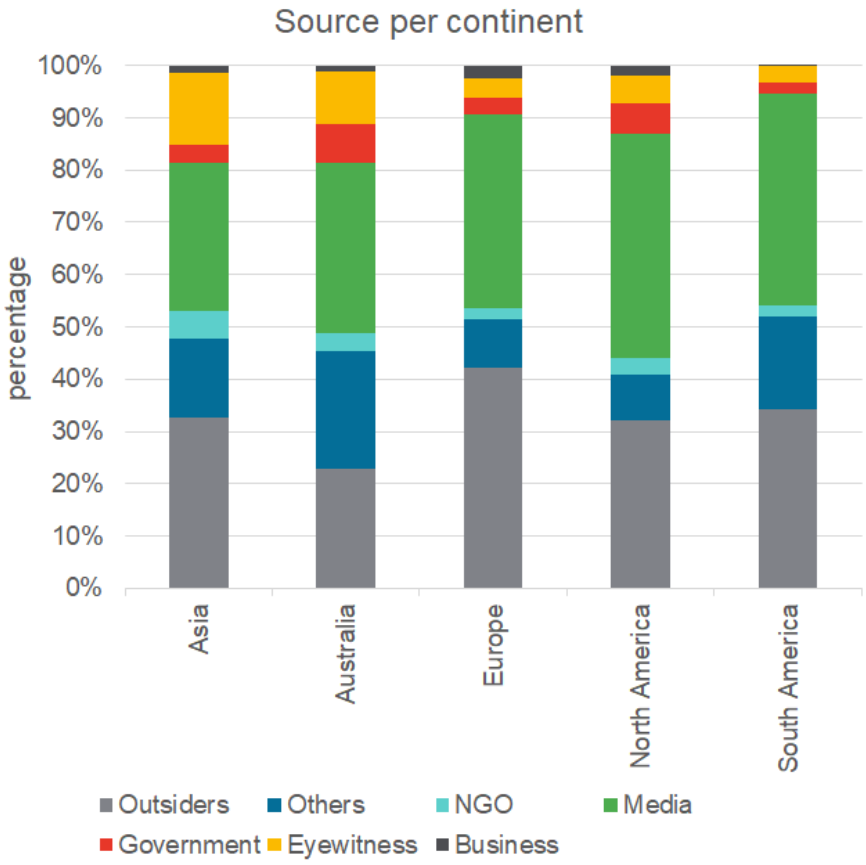


Figure 4.30: Information source per geographic location

media can be considered es highly informative. Over the entire duration of the events, in all geographic locations, there are more informative tweets than non informative tweets. Furthermore, except in Australia, in early phases media tweets are significantly more informative than non informative. Eyewitness reports, i.e., tweets from source eyewitness, expose a different perspective, which do not lead to such obvious conclusions. In the case of Asian events, there is a correlation with Figure 4.30, which indicates that eyewitness reports in early stages are the reason why there are more non informative tweets also in total. Hence, the informativeness of eyewitness reports is at least in that case not very high. Other interpretations are hard to generalize, since the entire dataset contains only 7.81% tweets with source eyewitness, which results in only very few samples per event.

To conclude, events from the Asian are tend to contain more tweets from source eyewitness and less tweets from source media than others. Especially, events took place in America have a higher proportion of tweets shared by source media. Since source media tend to be mainly informative, American events may lead to better informativeness classification than Asian events, because tweets from information source eyewitness can be informative or non informative equally.

4.2.4.4 Information source per time

Separating the evolution of informativeness in time regarding the information source, see Figure 4.37, shows some interesting insights. While information sources, media, government and NGOs tend to be much more informative also over the entire time period of the event, this is not true for information sources, business, eyewitness and outsiders. Considering the source Business in more detail, the chart shows more non informative communication in early stages of the event and tend to be more informative later on. Considering the information source eyewitness, the chart shows in early stages of the event, there are more non informative tweets than informative tweets. This changes slightly over time, in later phases of the event informative and not informative tweets tend to be equal. The peak, visible in the chart on day 6 might result from a particular progressive event where something unusual happens, which can be considered as an outlier. Moreover, tweets from the information source outsiders tend to be non informative, in early stages even worse.

To conclude, tweets from information sources media, business, government and NGOs are informative also over time, whereas tweets from information sources outsiders, eyewitness and others are not. As a consequence, considering the point in time when a tweet was posted may give additional information regarding informativeness in the case of information source eyewitness.

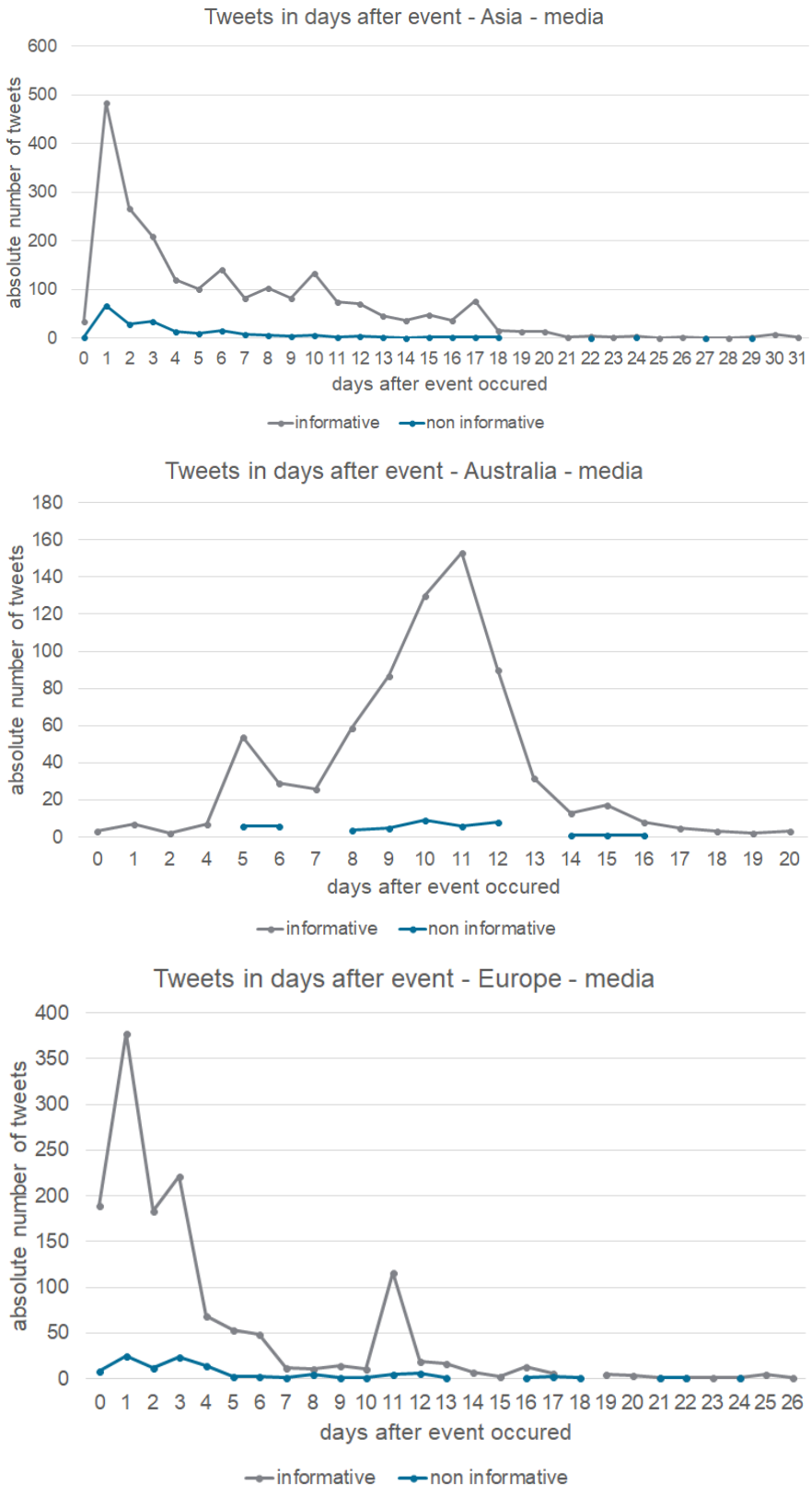


Figure 4.31: Informativeness per geographic location and information source media (1)

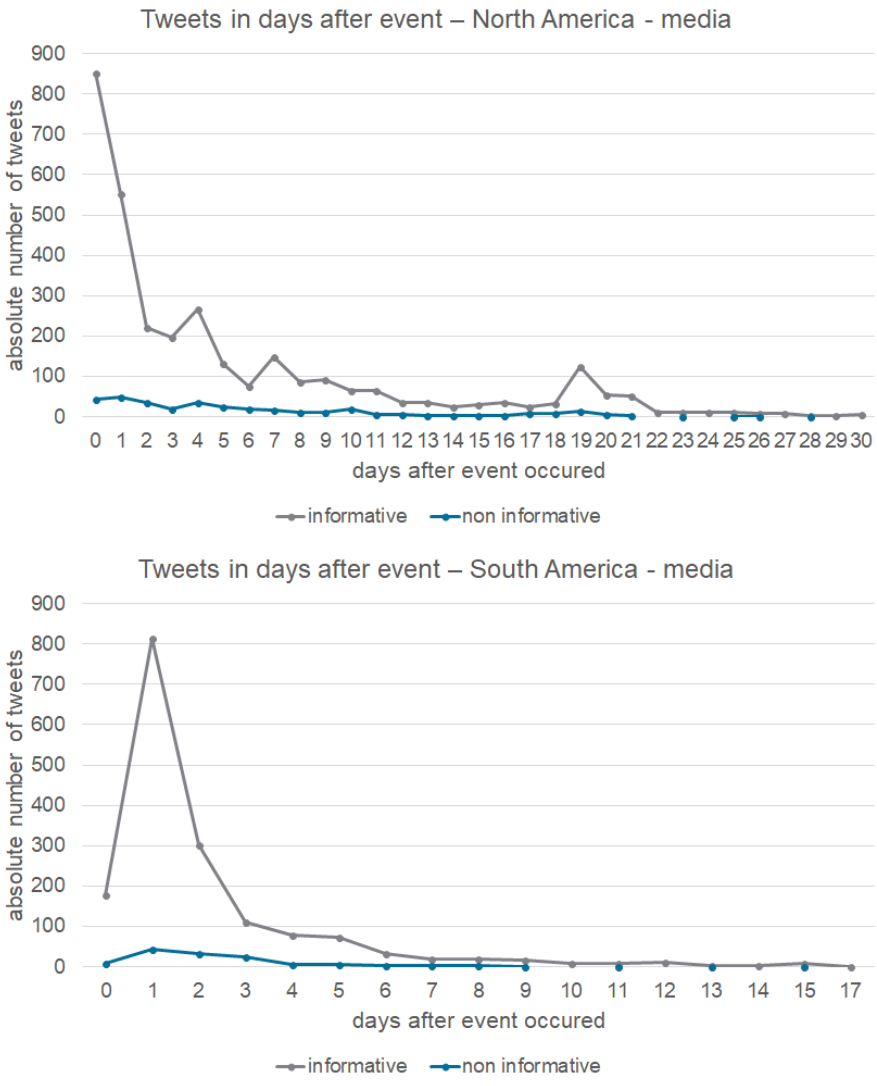


Figure 4.32: Informativeness per geographic location and information source media (2)

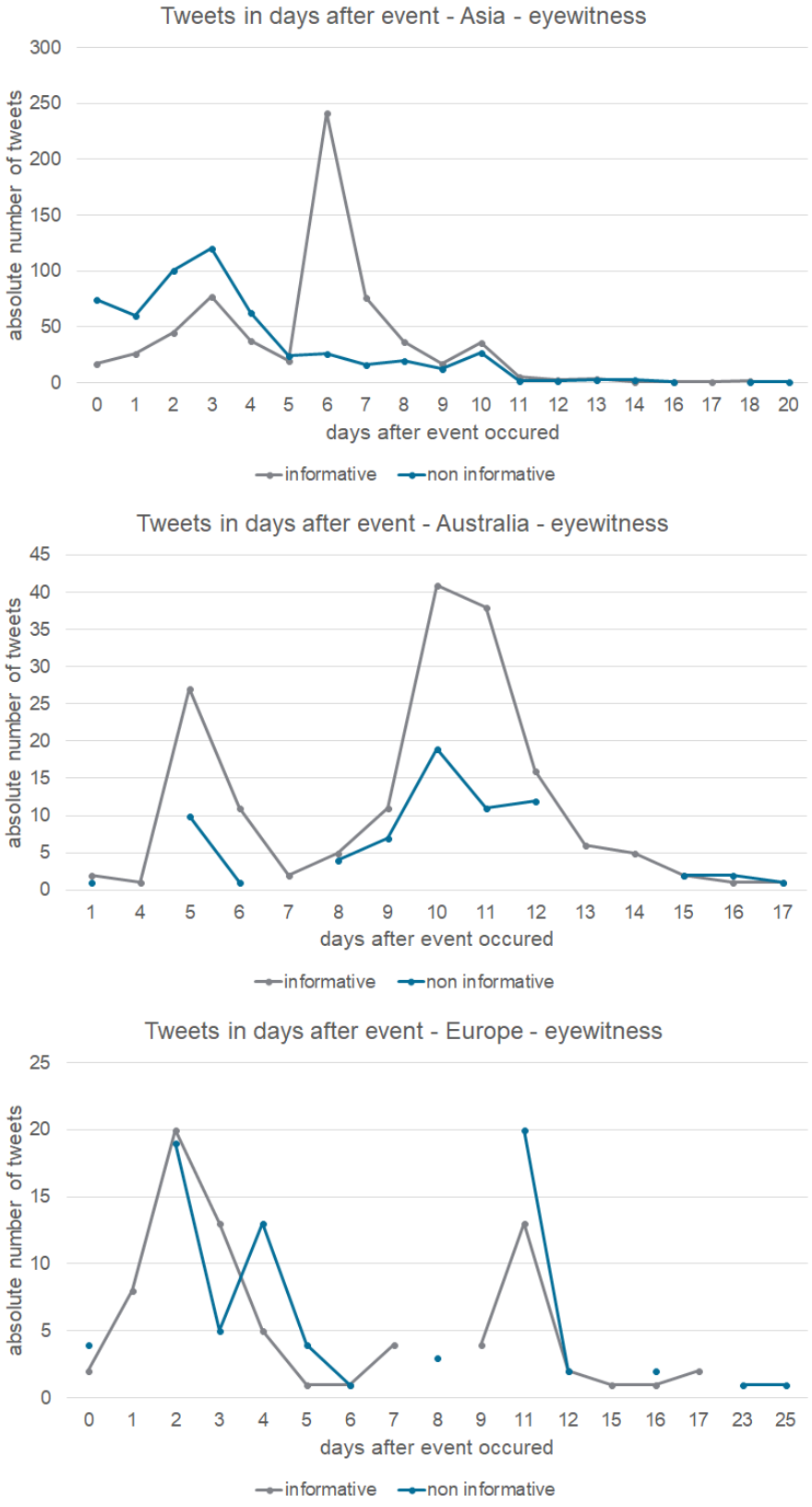


Figure 4.33: Informativeness per geographic location and information source eyewitness (1)

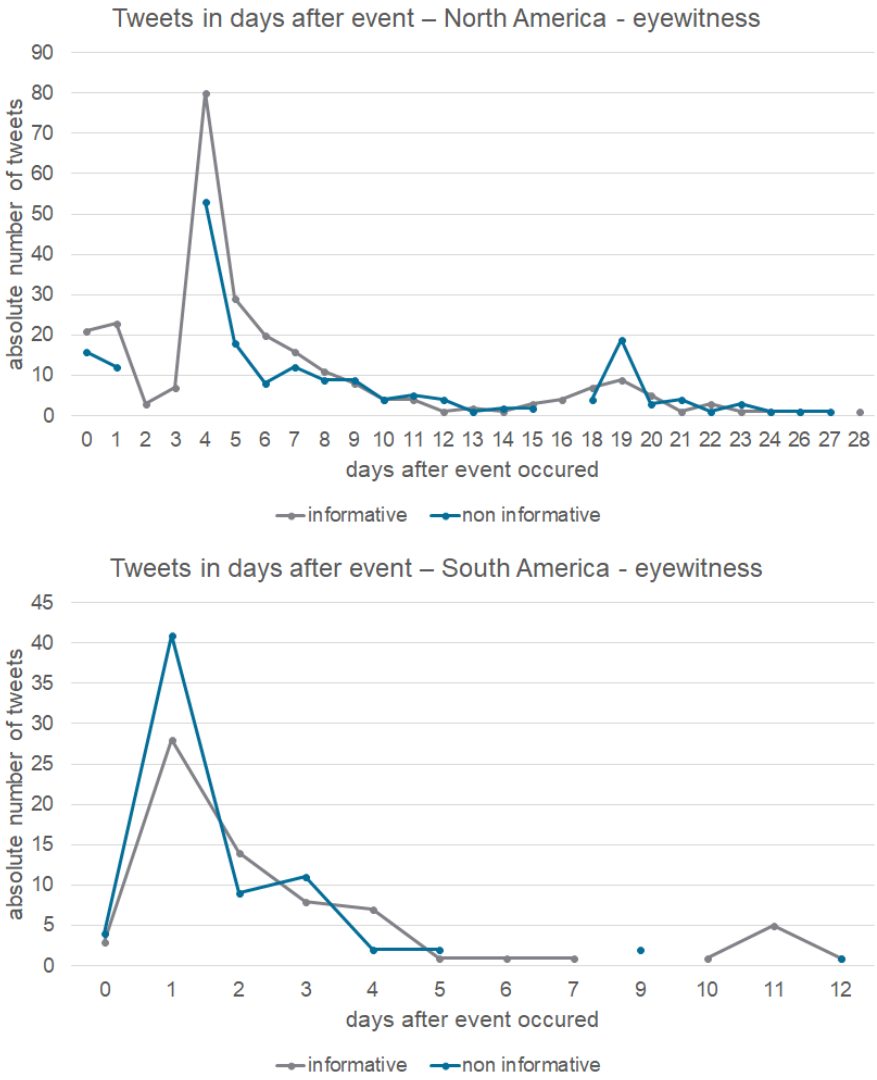


Figure 4.34: Informativeness per geographic location and information source eyewitness (2)

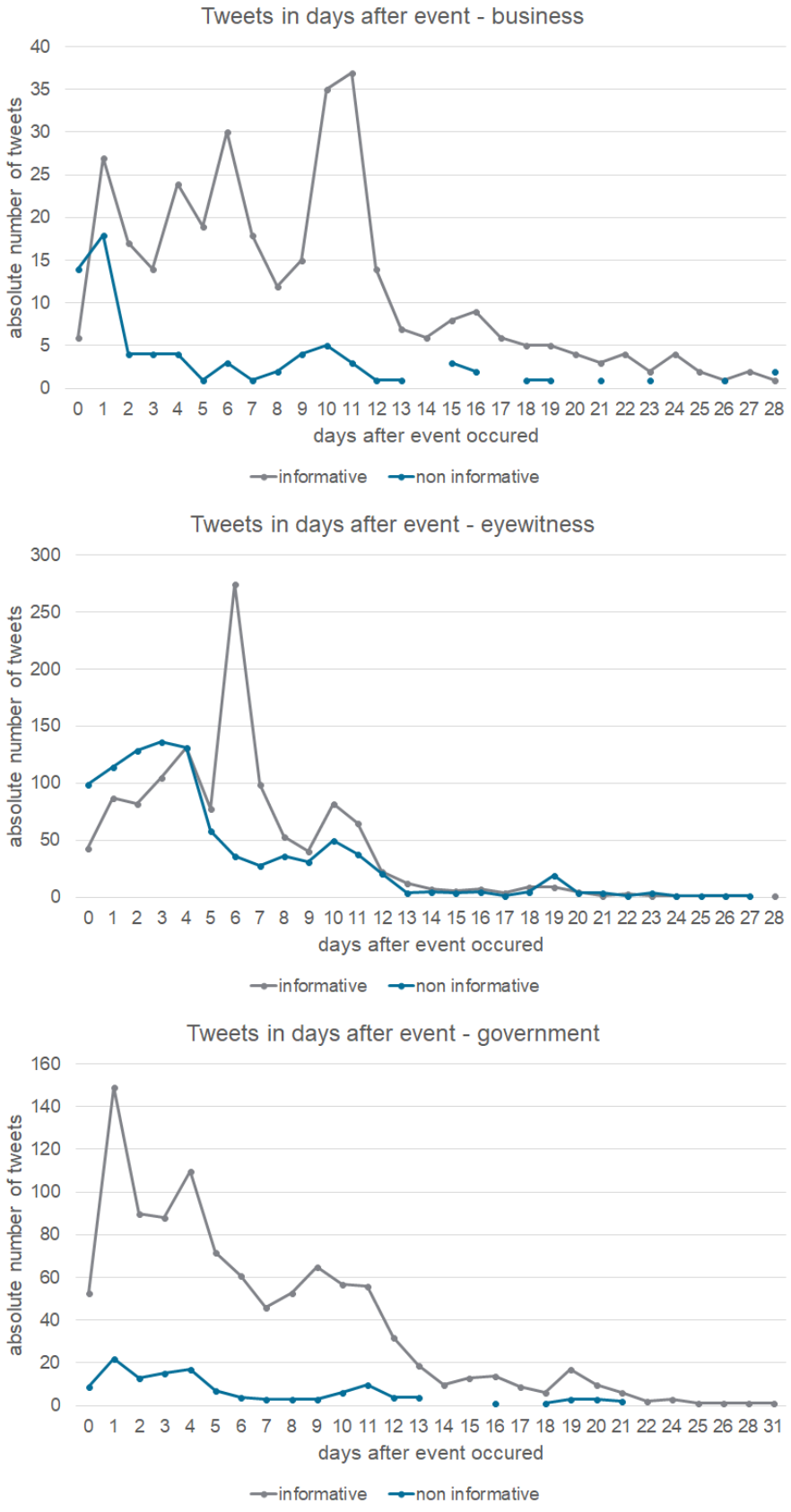


Figure 4.35: Informativeness per information source (1)

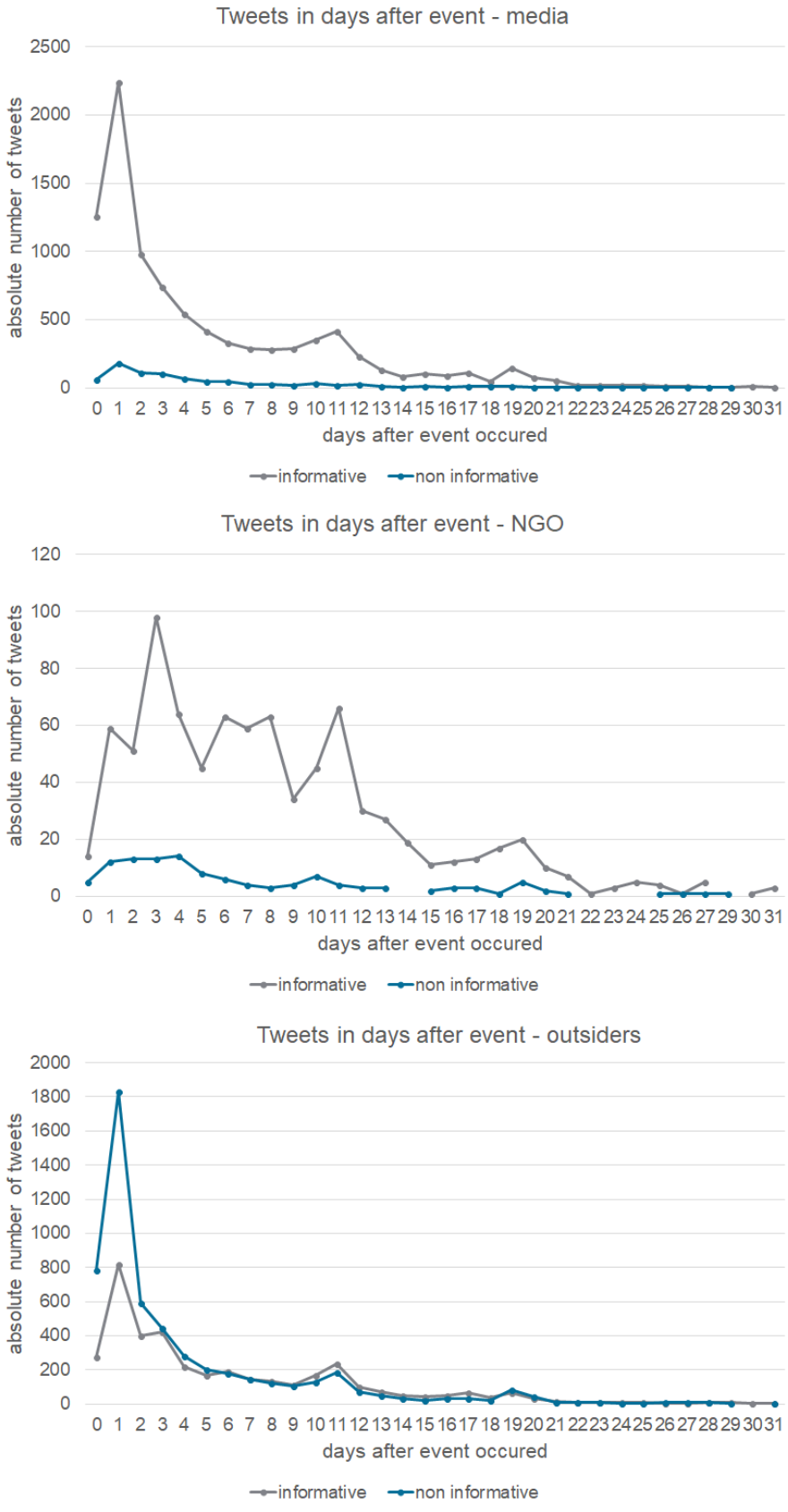


Figure 4.36: Informativeness per information source (2)

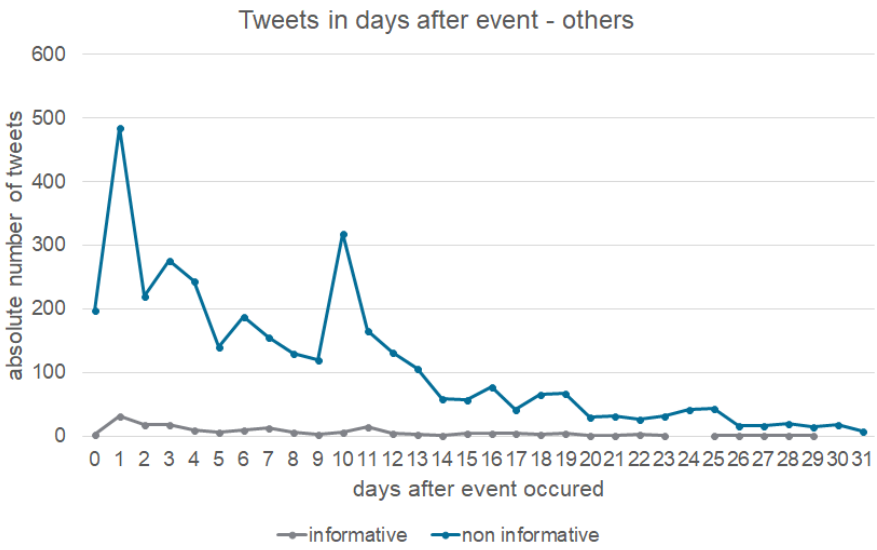


Figure 4.37: Informativeness per information source (3)

4.2.4.5 Findings

Figure 4.38 visualizes informativeness of tweets originating from different sources. Our main findings towards the source dimension can be summarized as follows:

1. Tweets shared by ”business”, ”media”, ”government” and ”NGOs” tend to be *informative*.

2. Tweets shared by ”eyewitness” are informative as well as non informative.

3. Tweets shared by ”outsiders” and ”others” tend to be *non informative*.

4. The amount of tweets shared by each source vary significantly, while 36% of all tweets originate from ”media” and 33% from ”outsiders”, only 4% originate from ”government”, 1.5% from ”business”, 8% from ”eyewitness”, 3.5% from ”NGO” and 14% from ”others”. Thus, analysis results based on these smaller classes might be not representative, especially when the amount being reduced further by considering only single events.

4.3 Event similarity

To achieve accurate prediction results on unseen new data, choosing the ”right” training set to train the classifier is one important variable. Only if new data, i.e., upcoming crisis

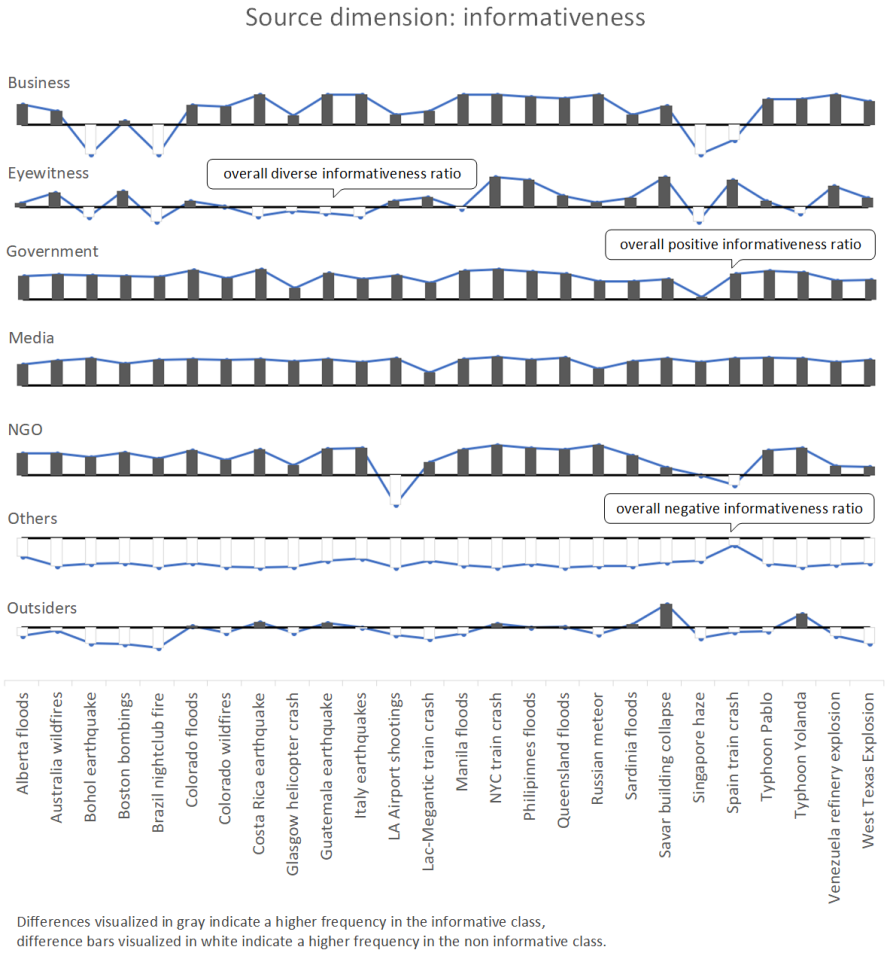


Figure 4.38: Source - informativeness

events, fit the trained model, the classifier is able to predict informativeness accurately. Hence, event similarity, i.e, how similar are events based on certain characteristics, is one possible measure to build appropriate training sets, especially for cross-domain training. New tweets are classified accurately more likely when they are more similar to tweets where the classifier is trained on. For instance, if earthquake events vary significantly from flood events based on characteristics described in the current chapter, it might be not useful to use a earthquake trained classifier on an unseen flood event. This section tries to show event similarity based on hierarchical clustering. In the first experiment, events are clustered based on event dimensions given by Olteanu et al. Clustering results of this experiment are considered as a baseline, which allows comparing further clustering experiments. In the second experiment, events are clustered based on tweet characteristics discussed in the previous sections.

4.3.1 Event similarity based on event dimensions

Considering the dimensions of events and characteristics of each event given in Table 4.1, hierarchical clustering can provide a simplified visualization of similarity in a cluster dendrogram. The clustering algorithm uses the event dimensions duration, geographic location, category and subcategory as well as development and spread as input and calculates the similarity between events. As expected, the dendrogram shows certain event types as clusters. For instance, floods are clustered together as well as earthquakes or wildfires.

The y-axis shows the similarity between certain clusters or events, respectively. Events which are clustered together very early, i.e., on the bottom of the chart, are considered to be very similar. For instance, Philippines floods and Sardinia floods are considered to be very similar based on the cluster dendrogram shown in Figure 4.39.

4.3.2 Event similarity based on tweet characteristics

Previously discussed characteristics can be used as features for hierarchical clustering of events. While the clustering results of Figure 4.39 show event similarity based on static event dimensions, the current clustering tries to visualize intrinsic similarity given by tweet characteristics. Similarity is based on how people write and communicate informative and non informative tweets in certain crisis events, i.e., rather the characteristics of the tweet itself than the characteristics of the event is important. Hence, similarity of events can be interpreted as people's online communication behavior in crisis events. For instance, Colorado floods and Colorado wildfires are considered to be similar, i.e., clustered very early (cf. Figure 4.40), although their event type is different.

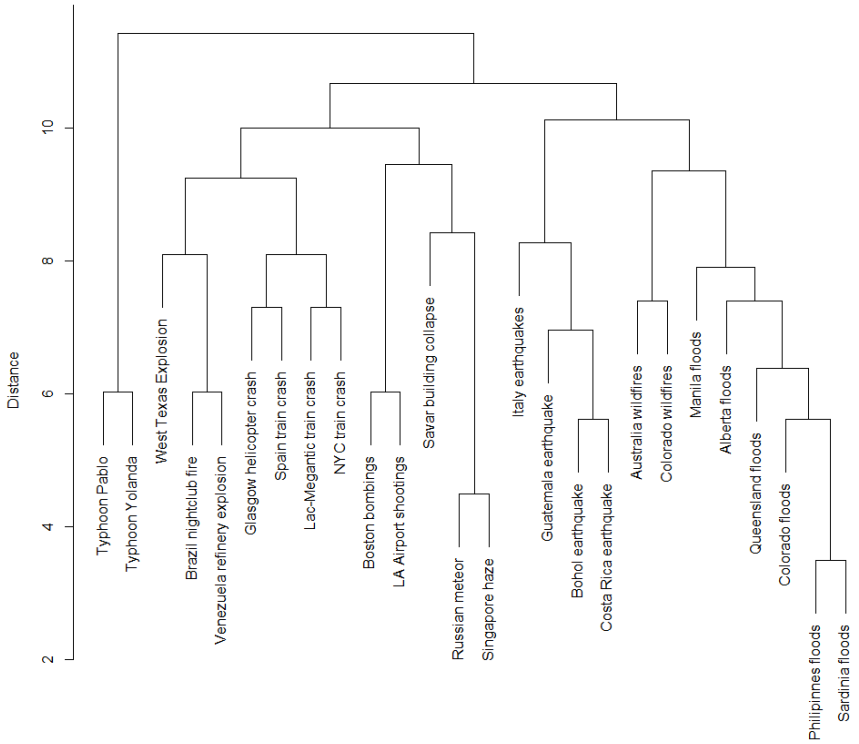


Figure 4.39: Event similarity by hierarchical clustering - event dimensions based on Table 4.1

Data characteristics described above are scaled and used as input for hierarchical clustering. Details are listed in Table 4.6. For the clustering itself, the R package "hclust" applying complete linkage was used to create clusters. For those characteristics which no information is available, because more than 50% of tweets are not in English, they are replaced by mean values of all other events.

Comparing both clustering results, clustering based on event dimensions (cf. Figure 4.39) and clustering based on intrinsic tweet characteristics (cf. Figure 4.40), show some interesting observations. While the dimension based clustering uncovers clusters mainly based on the event type, for instance Philippines floods and Sardinia floods, this is not the case in the tweet characteristic based clustering. Still, there are some event type based clusters, for instance Costa Rica and Guatemala earthquake or Manila and Philippines floods, yet, there are clusters based on tweet characteristics, for instance Italy earthquakes and Sardinia floods, which both took place in Italy, but are from different event types. One further example is Colorado floods and Colorado wildfire, which are clustered together although events are from different types, but same geographic location. Considering the cluster around both Typhoons show that Typhoon Yolanda is more similar to the Colorado cluster than to Typhoon Pablo, since it was clustered in a further step.

Feature type	**Features**
Language	EN ES PT TL
Information source	Business Eyewitness Government Media NGO Outsiders
Avg. number of tokens	informative not informative
Avg. number of nouns	informative not informative
Avg. number of verbs	informative not informative
Avg. number of adjectives	informative not informative
Avg. number of adverbs	informative not informative
Avg. number of ”#”	informative not informative
Avg. number of ”?”	informative not informative
Avg. number of ”!”	informative not informative
Avg. number of URLs	informative not informative
Avg. number of ”@”	informative not informative
Emoticons	positive negative
Sentiment	polarity informative polarity not informative subjectivity informative subjectivity not informative
Time	avg. reply time
Hashtags	informative not informative
Negation terms	informative not informative

Table 4.6: Features used for hierarchical clustering

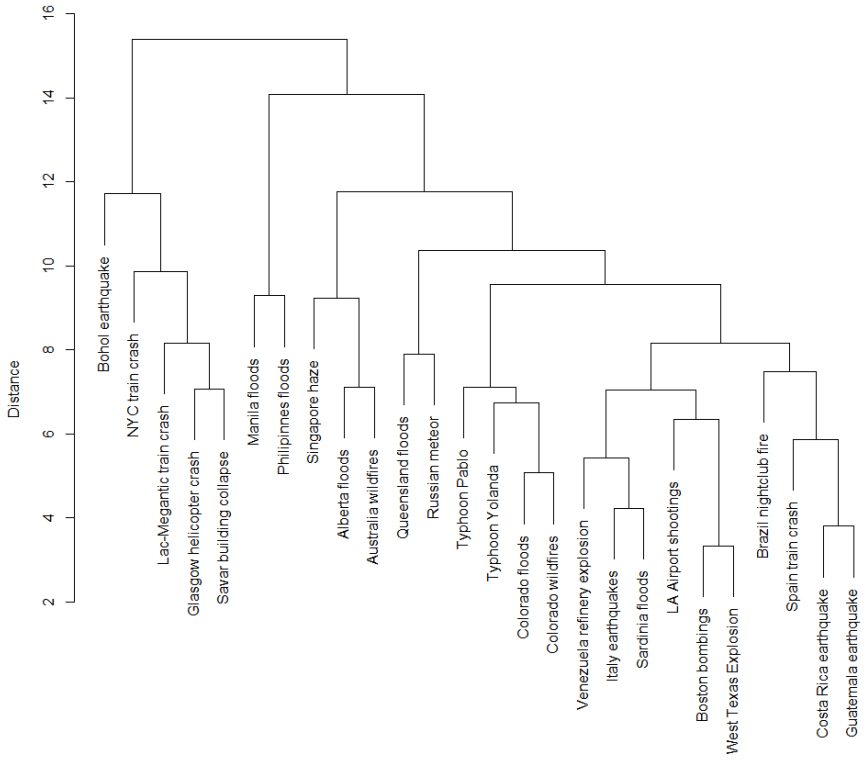


Figure 4.40: Event similarity by hierarchical clustering - tweet characteristics

To sum up, event types are not as important for informativeness classification. Tweet characteristics are considered as similar although they are from different event types. In other words, whether a tweet is informative or not depends on certain characteristics, those characteristics are not considerable different for different event types. Nevertheless, the importance of event types for classification might be different in other crisis related mining tasks, where semantic information is more important.

4.3.3 Tweets of events

Figure 4.41 underlines the statement of the precious section by showing all informative tweets in a scatter plot. No clusters related to event types are recognizable. Tweets from all events are scattered distributed over the entire plot. Since there are no clusters of certain events or event types, tweet characteristics are distributed over all events. For the scatter plot, characteristics described in the sections above were used. Furthermore, Principal Component Analysis (PCA) was used to visualize d-dimensional information in a two dimensional space.

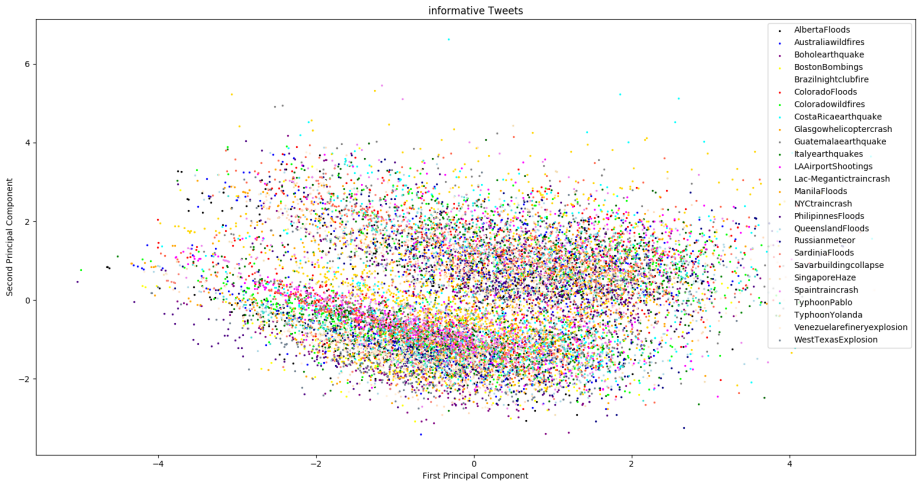


Figure 4.41: Tweets scattered after PCA applied

4.4 Conclusion of data analysis

Detailed analysis of the CrisisLexT26 dataset shows some interesting insights in data and furthermore important aspects for informativeness classification. Firstly, there are some trends of certain characteristics of tweets, which tend more belonging to the informative class or the non informative class. For instance, tweets originating from source media, government or NGOs tend to be more informative than tweets originating from source eyewitness, business or outsiders. Analyzing informativeness over time show eyewitness reports in early stages are not highly informative. Moreover, geographic location does not make significantly difference. While informativeness of tweets in instantaneous crisis events tend to be similar considering tweets in time, this is not the case for progressive events. Yet, more crucial for informativeness classification are linguistic characteristics of tweets. The average informative tweet contains more words, more nouns, less verbs, more adjectives, less adverbs than the average not informative tweet. Moreover, the average informative tweet contains more URL links as well as more ”@” characters. In contrast, the average non informative tweet contains more ”!”, ”?”, which tend to be more emotional, which in turn tend to be not as informative. Regarding sentiment in tweets, positive sentiment tend to non informative tweets and negative sentiment tend to informative tweets. In addition, crisis related hashtags in the tweet are an indicator for informative tweets, whereas negation in tweets is an indicator for non informative tweets.

Beside these tendencies of certain characteristics, results of the hierarchical clustering allow one further conclusion. For informativeness classification of tweets, the event type does not seem to be as important. Considerable more important, whether a tweet is informative or not, are intrinsic characteristics of the tweet itself. In other words, the main reasons

why a tweet is informative or not, is what the tweet is about and what text it contains, and these characteristics with respect to informativeness does not differ significantly cross events. The characteristics discussed in this chapter, try to express characteristics of a tweet in values, which can be used as features for classification.

Finally, main findings along all four dimensions as well as results of hierarchical clustering can be summarized in *two outcomes*: Firstly, analysis of temporal, linguistic and source dimensions uncovers tweet characteristics having impact on informativeness thus being candidates to be used as features, whereas analysis of spatial dimension (per continent) shows no significant correlations with respect to informativeness. Secondly, the fact that analysis does not show significant differences with respect to disaster types and furthermore does not show significant similarities within same disaster types, supports the hypotheses to train an accurate cross-domain classifier cross the 13 different disaster types available in our dataset.

Chapter 5

Classification approach

Based on the analysis presented in the previous chapter suggesting that different disaster types do not show significant differences in terms of informativeness and taking into account indication of informativeness of certain characteristics, this chapter proposes a classification approach which can be used on new events of various disaster types while being at least as accurate in informativeness classification as disaster type specific ones. Data analysis uncovers characteristics, which show correlations to either informative or non informative tweets and thus are suited to be realized as features for classification.

The current chapter proposes a cross-domain informativeness classification approach, visualized in Figure 5.1. In the following, firstly, some important aspects in classifying tweets with respect to informativeness are mentioned. Secondly, features used for classification are described in detail as well as their extraction out of the tweet text. Thirdly, used train- and test sets are discussed.

Informativeness classification in the current thesis can be considered as binary classification of potentially crisis related tweets, which are either informative or non informative. The classification approach proposed in this thesis is designed to be used upon directly on the Twitter message stream (cf. Figure 5.1), i.e., shared tweets are instantly classified by an already trained classifier. In literature this is called ”online” classification. Thereby, classification is mainly based on the information source and characteristics of the tweet content itself as data analysis showed that informativeness does not differ a lot in different event types, rather informativeness correlates with information source and certain tweet characteristics. Thus, it is not essential to distinguish events and different event types for classification if the features used expresses the characteristics of the tweet content. This approach can be used for in-domain, out-domain and cross-domain classification.

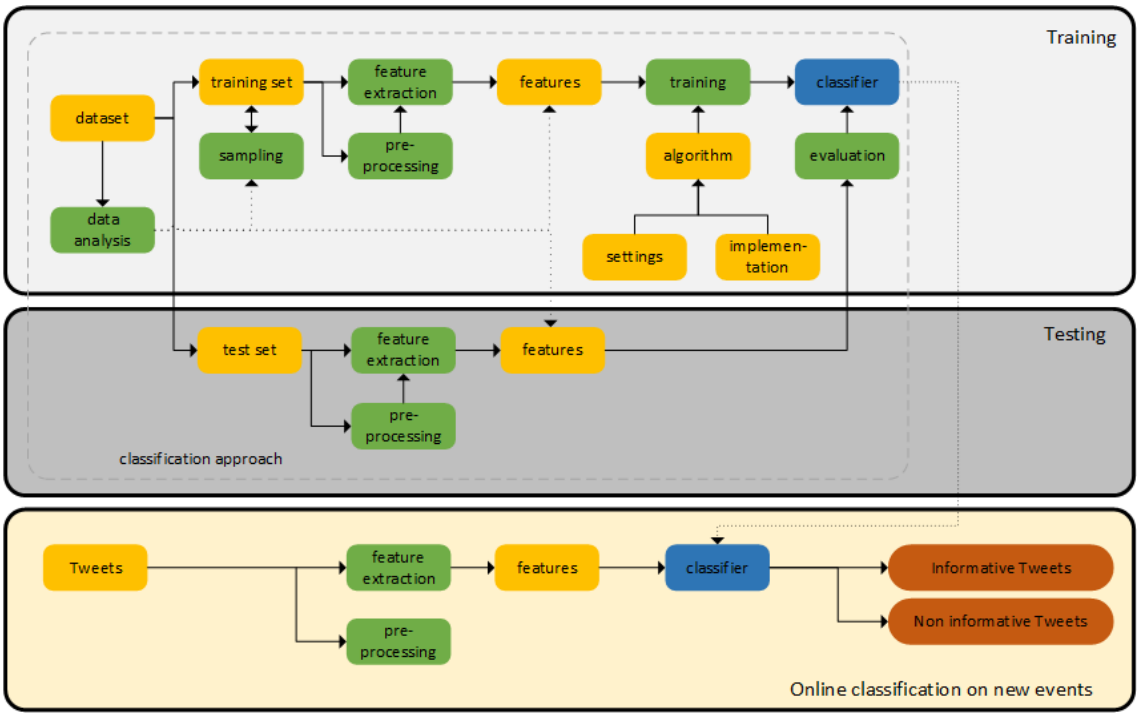


Figure 5.1: Classification approach

5.1 Offline versus online classification

Using Social Media messages as source for classification, literature distinguish between offline and online classification. Offline means applying classification of messages, on an offline available dataset, mainly for research and development purposes. In contrast, online means applying classification directly on online data streams, which is the final use case also for crisis related classification. Finally, a learned classifier classifies messages in real-time on the message stream to filter informative from not informative messages or tweets, respectively. As a consequence of online classification, some information is not available to use for classification, e.g. for features. For instance, it is not possible to use the number of Retweets as feature for online classification, since the final number of retweets is not known at the point in time when a certain message is classified.

The current approach proposed in this thesis is designed to use for online classification. Hence, all used feature can be extracted immediately when classifying the tweet. Only exception is the feature information source, which is not explicitly given by Twitter even though user profiles show the source implicitly. In addition, Supervised Machine Learning can be used to classify tweets towards information source. However, this task is not in the scope of this thesis.

5.2 Processing textual data

In general, processing textual data usually requires some pre-processing steps before features can be extracted out of the tweet text. In the current approach, first, stop words are removed, because they usually do not provide any additional information. Second, Tokenization splits up sentences or words into tokens, which then are processed further. Third, Stemming converts tokens in their basic form, if possible. For instance, the token "supported" is converted to the basic verb form "support". Not all pre-processing steps are suitable for all kind of features, yet, this has to be decided individually for each feature. Details regarding features are explained in the following section.

5.3 Features

Essential for accurate classification results are the features used. Thus, extracting and selecting features, which lead to accurate classification, is a main part of the entire classification process. As discussed in Chapter 3, many different types of features exist and also are used in approaches of related work. Whether an individual feature is suitable for classification or not can be tested using feature evaluation methods. Different feature evaluation methods exist [Pedregosa et al., 2011]. One possible way of feature evaluation is determining the correlation between an individual feature and individual class labels. Features, which have a higher correlation with one of the class labels are more expressive regarding classification. Moreover, correlations between features among themselves allow elimination of one feature of both, because the second feature does not provide any additional information, and more features increase learning complexity. For instance, if feature A has a strong correlation with feature B, then one, either feature A or feature B, can be excluded from classification, since feature B does not provide any additional information to feature A. Which features are used in the current classification approach are listed in Table 6.1. In the following, features are explained in more detail along this table.

5.3.1 Reply time to event

The time span from the begin of a crisis to when a certain tweet is shared online, contains important information. This time span can be an indicator whether it is informative or not. Therefore, time based features are crucial for classification and commonly used in research. The current feature represents the time difference between the start of the crisis to the point in time when the tweet was shared online. The time difference is measured

Nr.	**Dimension**	**Feature**	**Data type**
1	Temporal	Response time	Integer
2-5	Linguistic	Probability of language EN, ES, TL, PT	Float [0,1]
6		Number of tokens	Integer
7-10		Special characters[a] "#", "?", "!", "@"	Integer
11		Links (URLs)[a]	Integer
12-15		POS[a]: nouns, verbs, adjectives, adverbs	Integer
16-17		Positive/negative Emoticons[a]	Integer
18		Sentiment polarity	Float [-1,1]
19		Sentiment subjectivity	Float [0,1]
20		Crisis related Hashtags[a]	Integer
21		Negation terms[a]	Integer
22		Sentence finishes with punctuation	Binary
23	Source	Media	Binary
24		Business	Binary
25		NGO	Binary
26		Government	Binary
27		Eyewitness	Binary
28		Others	Binary
29		Outsiders	Binary

[a]quantity contained in tweet text

Table 5.1: Set of features - grouped by analysis dimensions

in days. Values close to zero mean the tweet was shared in very early stages of the event, higher values mean the tweet was shared later during the event.

$$featureReplyTime = tweetPostDate - eventStarted \tag{5.1}$$

5.3.2 Language

Obviously, language information can be crucial in combination with other feature due to the reason that some feature rely on the used language. Results, for instance Part-of-Speech information, are different in different languages. For this work the top four languages used in the dataset are represented in four features and four feature values respectively. The top four languages includes more than 90% of all tweets. The language of each tweet is determined using an language detection tool based on the tweet text itself. More details about the tool are described in section 6. The values of all four features represents the probability of a certain tweet belongs to the language. Values close to one mean the tweet belongs with a high probability to this particular language.

5.3.3 Number of tokens

Data insights showed that there is a tendency of shorter tweets belonging to the not informative class. Therefore, this feature is based on the tweet text itself and represents the tweet text in number of tokens. Tokens are extracted from tweet text after removing stopwords and applying Tokenization in pre-processing steps. The implementation used for removing stopwords and Tokenization is described in more detail in the following Chapter 6.

5.3.4 Special characters

According to data analysis, frequently used special characters in tweets are suited to distinguish informativeness in tweets. Each character ”#”, ”?”, ”!”, ”@” or character sequence in case of URLs respectively, result in one feature, where its values are the frequency used in raw tweet text. Number of URLs refer to web-links only or ”http” links respectively. In total, special characters in tweet text are represented as five independent features.

5.3.5 POS features

Part-of-Speech information differ regarding informative and not informative tweets. Data analysis showed that nouns and adjectives tend to be more frequently used in informative tweet text, whereas verbs and adverbs tend to be more frequently used in non informative tweet text. The Part-of-Speech information results in four different features: number of nouns, number of verbs, number of adjectives and number of adverbs. Its values represents the frequencies of each Part-of-Speech element in tweet text. Pre-processing in form of removing stopwords, Tokenization and stemming are applied. The implementation used for extracting POS information is described in more detail in the following chapter 6.

5.3.6 Emoticons

Emotion in tweet text is very commonly used to be represented as features in related literature. However, to determine the emotion of tweets there exist a variety of methods, also very complex ones. Acerbo et al. [Acerbo and Rossi, 2017] showed in their work that emotion tend rather to be not informative. This is also true for the statistics shown in data analysis in chapter 4. A very simple form of emotion expression is the usage of Emoticons, for instance, ":-)" or ":-(", which was used in the current work. Two features expresses positive and negative Emoticons. Its values are the amount of positive and negative Emoticons used in tweet text.

5.3.7 Sentiment

Data analysis showed that sentiment in tweet text can possibly be an indicator whether a tweets contains informative information or not. Various sentiment analyzing tools exist, which allow to assess a sentiment score based on the language of the text. For this work a sentiment analyzing tool was used. Hence, two features represent the sentiment in tweet text. The first feature expresses the sentiment polarity, whether the tweet contains positive or negative sentiment. The second feature represents the subjectivity of the sentiment, whether the sentiment polarity value is subjective or rather objective. Both features result from the used sentiment analyzing tool, which is the reason why sentiment is realized as two features for classification. Details about the tool are described in section 6.

5.3.8 Number of hashtags

Additional information whether a tweet contains informative information or not, are the number of hashtags used in the tweet. Data analysis showed that there is a trend towards less crisis related Tags in informative tweets. The current feature value represents the number of crisis related hashtags used in tweets. Crisis related hashtags are defined by the CrisisLexT26 dataset for each crisis.

5.3.9 Negations terms

Data analysis showed, more negations terms are used in not informative tweets than they are used in informative tweets. Thus, the information about negation in tweet text is a potential feature for classifying tweets. The resulted feature represents the number of negation terms used in the tweet text. In the following list contains terms considered as negation terms and it was created during classification experiments. (profane words are spelled with "*" to replace one letter)

Negation terms: not, none, neither, never, no one, nobody, nor, nothing, nowhere, does not, did not, f*ck

5.3.10 Sentence finishes with punctuation

An additional feature used in the work of [Cresci et al., 2015] is whether a tweet sentence finishes with punctuation or not. Punctuation includes the characters ".", "!" an "?". This feature results in a binary feature where its value is one when the tweet text finished with punctuation.

5.3.11 Source

Information source showed considerable difference regarding informativeness classes after analyzing the data. Formal sources, like media, government or NGOs, tend to be much more reliable and informative than not formal sources like eyewitness, outsiders or even business. Hence, representing the information source as a feature leads to an important information to classify tweets. The information source is given by the dataset and was labeled manually by humans. Naturally, this information is not explicitly known using the classifier in an online real scenario, however, the information source is implicitly given

by the user who shares the tweet. Information source results in six binary features for classification. Each source represents one binary feature.

5.4 Training and testing the classifier

The current section describes different training and test sets to use in more details.

5.4.1 Training

In order to evaluate which classifier, a cross-domain one or an event specific one, is more suited for new events, the approach proposed in this thesis considers three different types of experiments regarding the disaster type:

1. **In-domain**: training and test data belong to the *same* disaster type.

2. **Out-domain**: training and test data belong to *different* disaster types.

3. **Cross-domain**: training set consists of tweets of *various* disaster types, the test set is of a disaster type included or not included in the training set.

Additional experimental parameters, which can be used in different ways for classification experiments, allows to realize variations of in-domain, out-domain and cross-domain experiments: i) Over- and under-sampling allows to balance the data, i.e., use same amount of tweets in both informativeness classes. ii) Randomly sampling portions of the training data allows more flexible classification experiments. iii) The training set can be split up into subsets regarding the time when the tweet was shared to use separate classifiers for each time phase during an event (Details are explained in Section 5.4.1.2). iv) Using multiple events for training allows to perform cross-domain training. v) Cross-validation can be used to optimize classification in the development phase. These various implementation-parameter settings can be used to train different classifier and compare their informativeness classification accuracy.

5.4.1.1 Cross-domain training

In particular, this thesis emphasizes cross-domain training, since data analysis showed that differences between event dimensions are not as important as the tweet content itself. Hence, cross-domain training might be more sufficient due to it allows training on a larger

amount of training samples and therefore gets a more heterogeneous classification model to use for unseen tweets. This thesis wants to show how well cross-domain training perform in comparison to in-domain and out-domain training and testing. To apply cross-domain training the ”leave-one-out” method is used. The classifier is trained on n-1 events and tested on the remaining event, which was not used for training. To distinguish further, experiments can be run cross-domain or cross-event, where multiple events from the same event type are used. Thus, comparing their experimental results allows to uncover the factor event type in the classification process.

5.4.1.2 Three-step time dependent classification

Since the period when a tweet has been shared after the event started is crucial for classification [Stowe et al., 2016], the current section introduces a time dependent three-step classifier. Timeline analyses of events showed that informativeness is also time dependent, i.e., the period when a tweet has been shared after the event started. Hence, this time dependent classification approach splits the training as well as the test set into three subdatasets regarding the time period of the event. Therby, time periods are statically or dynamically defined, chosen by parameter settings during experiments. Static time interval means, for all events there are fixed days of the entire crisis event period, which belong to a certain phase. For instance, tweets between day 0 and day 3 belong to the early stage of the event, tweets from day 4 to day 10 belong to the middle stage and greater tweets shared after day 10 belong to the late stage of the event. Thus, each tweet belongs to a particular subset. In dynamic time intervals the event stages are dependent from the duration of the current event. However, the current event duration is not known when the classifier is used for online stream classification. As a results, each stage of the event is classified separately, based on trained classifiers on the particular stage, only.

5.4.2 Test classifier

In order to simulate a real scenario where tweets are classified with respect to informativeness on new events, the trained classifiers are tested on unseen crisis events. To compare their classification results, classifiers are tested on the *same* crisis event.

5.4.3 Evaluation

Evaluation of classification results is important to estimate how accurate a trained classifier perform in real classification scenarios. In the current thesis, classification accuracy is

employed as basis of evaluation, i.e., how many tweets related to the total number of classified tweets are classified as informative or non informative correctly.

5.5 Summary

The informativeness classification approach presented in the current chapter tries to operationalize the outcomes of the previous data analysis chapter. Firstly, certain characteristics, which show correlations towards one of both informativeness classes, are realized as features for classification. Secondly, the current approach allows flexible compositions of training sets including cross-domain training since data analysis show event similarities are not as important for informativeness classification. Therefore, the possibility of using a larger training set by cross-domain training might be sufficient to improve the performance of informativeness classification.

Chapter 6

Implementation

The implementation of a classification prototype described in this chapter is based on the approach discussed in the previous Chapter 5. In contrast to the previous chapter, the current chapter focuses on the actual implementation of classification experiments in more detail, explains used tools and settings made. The full implementation is available on Gitlab[1].

Firstly, this chapter discusses tools and technology used. Secondly, some important Python libraries for feature extracting are listed. Furthermore, this chapter focuses on data processing techniques commonly used for Machine Learning task. Finally, this chapter describes used classification algorithms and its settings.

6.1 Tools and technology

Data analyzing was done using Microsoft Excel including Pivot Tables and Pivot Charts for visualizing and interpreting the data characteristics, which are described in chapter 4. The tool R including R-Studio was used for the hierarchical clustering of events.

For Machine Learning tasks many different implementations of algorithms in different languages exist. Some examples of popular languages for scientific research are WEKA, R and Python. Python supports the very frequently used Machine learning library Scikit-Learn [Pedregosa et al., 2011]. Scikit additionally supports the entire data processing process, from pre-processing steps to evaluation metrics and provides a variety of state-of-the-art Machine Learning algorithms. Scikit-Learn is built on NumPy[2], SciPy[3], and matplotlib[4]

[1] https://gitlab.com/davidgraf-mathesis/crisisInfoMining.git
[2] http://www.numpy.org/
[3] https://www.scipy.org/
[4] https://matplotlib.org/

and it provides "simple and efficient tools for data mining and data analysis"[5]. In addition, it provides an extensive online documentation as well as some general information about Machine Learning techniques. Furthermore, Python was already used in a previous project, parts of the code of [Graf, 2017] were reused as a starting point for the current implementation. Thus, the main implementation of this classification prototype is realized in Python.

For storing and accessing tweet data a relational Postgres[6] database is used for which the python library Psycopg2[7] allows convenient access in Python.

The following subsections deal with special Python libraries used for certain tasks in the implementation.

6.1.1 NLTK

NLTK stands for Natural Language Toolkit[8] and provides support for Natural Language processing in Python programs. In the current work, an English stopword lexicon, a Tokenization function and the PorterStemmer for stemming, all provided by the NLTK library, are used. In addition, a Part-of-Speech tagger function "pos_tag" of the NLTK library is used to annotate English nouns, verbs, adjectives and adverbs in tweet sentences.

6.1.2 TextBlob

Like NLTK, TextBlob[9] is a Python library for Natural Language Processing and is dealing with textual data. Particularly, TextBlob provides additionally a sentiment analysis module, which is easy to use. Based on a pattern analyzer the sentiment analysis module's output is an estimation of the popularity and the subjectivity of a sentence or a particular text. Popularity values are in an interval from minus one to plus one, where negative values is related to negative sentiment and positive values are related to positive sentiment. Additionally, the subjectivity value estimates the objectivity of the sentiment popularity, where zero is a totally objective and 1 is a totally subjective assessment.

[5] http://scikit-learn.org
[6] https://www.postgresql.org/
[7] https://pypi.python.org/pypi/psycopg2
[8] https://www.nltk.org/
[9] http://textblob.readthedocs.io/en/dev/index.html

6.1.3 Language detection

Since there is no language information beside the tweet text itself, a language detection tool was used to assess the language based on the tweet text. The Python library Langdetect[10] is ported from Google's language-detection and supports 55 different languages. In the current implementation Langdetect was used to calculate the probability of a certain tweets is written in English, Spanish, Portuguese and Tagalog, a language used in the Philippines. A probability close to one means a certain piece of text is certainly written in this language.

6.2 Data processing

Before data and extracted features can be used for training the classifier some pre-processing steps are necessary. This section describes these steps in more detail.

6.2.1 Data balancing

Class distribution and data balancing are crucial aspects to consider in Machine Learning applications. Some classification algorithms are vulnerable [Pedregosa et al., 2011] against unbalanced data, which lead to biased classification models and results respectively. Unbalanced data mean samples are not equally distributed between class labels, i.e., the number of samples belonging to each class are significantly different. For instance, only 20% of all samples belong to the informative class and 80% of all tweets belong to the non informative class. Informativeness distribution of all events is visualized in Figure 4.2 in Chapter 4. Hence, this class distribution of samples might bias the classifier. In literature as well as in online documentations, e.g. of Scikit-Learn[11] there exist several strategies against unbalanced data. One possible option is sampling to gain equal distribution over classes. In this work under and over sampling, explained in the following sections in more detail, are used to balance data.

6.2.1.1 Data under-sampling

Under-sampling is a simple strategy to gain equal class distribution. The same amount of samples are used for both classes, thus leaving out some of the available samples from the larger class. This results in classes with same amount of samples. Yet, one major

[10] https://pypi.python.org/pypi/langdetect
[11] http://contrib.scikit-learn.org/imbalanced-learn/stable/introduction.html

disadvantage of under sampling is the reduced dataset size, i.e., there are in total less training samples for the final dataset. This might not be a problem in some cases, however, in the current thesis under sampling is not suitable.

6.2.1.2 Data over-sampling

If the final data size matters, over sampling is probably the option more suitable to balance the data. Over-sampling means replicating samples from the smaller class until the former smaller class has become the same size as the former larger class. A disadvantage of over sampling is that there are copies of training samples included in the data, which might lead to overfitting of the classification model.

For this work in the final classification approach over-sampling is used, since over-sampling does not reduce the amount of samples.

6.2.2 Feature scaling

Extracted features values usually have different units. Hence, scaling these values to common units is necessary to prevent classification bias. Scikit-Learn provides several feature scaling methods, whereas it turned out that the Standard Scaler[12] work best for this classification approach. The Standard Scaler scales all feature values independently from each other and "Standardizes features by removing the mean and scaling to unit variance".

6.3 Feature evaluation

Whether a feature is suitable for the final classification or not can be measured by feature evaluation metrics. Several of them are provided by the Scikit-Learn framework. Basically, the correlation between single features and the class labels are determined. Strong correlation of a single feature to a certain class expresses an important feature for classification. The feature evaluation metric used in this prototype implementation is "mutual information classification", which measures the dependency between the feature and a certain class label. Thereby, value close to zero means no dependency, a higher value means higher dependency. Feature evaluation is an important step to select those features which provide additional information regarding the classification. Table 6.1 shows mutual information classification values on average over all events, sorted in descending order. Higher

[12] http://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

Feature	Dependency to class
sourceMedia	0.134248132
numberOfURL	0.058554753
sourceOut	0.053998574
sentimentpolarity	0.048314646
numberOfNouns	0.046057527
sentimentsubjectivity	0.038255962
numberOfHash	0.037680616
replyTimeToEvent	0.031614661
numberOfTokens	0.025143307
numberOfContainedTags	0.022630527
langEN	0.021173189
numberOfAdj	0.016483257
numberOfEMark	0.015189369
numberOfQMark	0.015153933
numberOfAt	0.015095157
numberOfAdv	0.012776012
langTL	0.012484242
sourceEyew	0.012277692
sourceGov	0.011127693
numberOfVerbs	0.010531956
finishesWithPunct	0.009901169
sourceNGO	0.009665882
langES	0.008910637
langPT	0.008864438
sourceBus	0.006817192
negationWordsCount	0.006486926
posEmoji	0.00613398
negEmoji	0.005494237

Table 6.1: Mutual information classification - on average over all events

values mean more important for classification. For instance, the feature source media is highly important for classification.

6.4 Classification algorithms

State-of-the-art informativeness classification employ standard algorithms, like Support Vector Machines (SVM) [Cresci et al., 2015][Stowe et al., 2016][Khare et al., 2017] [Mohammad et al., 2013][Pekar et al., 2016], Naive Bayes classification, Maximum Entropy Models [Stowe et al., 2016] [Verma et al., 2011][Pekar et al., 2016] or Random Forest classification [Acerbo and Rossi, 2017][Imran et al., 2016] [Longhini et al., 2017] as well as deep learning [Ning et al., 2017], using a Convolutional Neural Network (CNN). Classification algorithms provided by Scikit-Learn, such as SVM, Naive Bayes, AdaBoost (an ensemble method), Random Forest, and a Multilayer Perceptron (a neural network)

have been used for the presented experiments. However, experiments showed that SVM using an RBF kernel by applying Scikit-Learn default settings (C=1.0, gamma='auto') work best out of all other algorithms being therefore the first choice for all of our experiments.

Chapter 7

Evaluation

The current chapter describes various informativeness classification experiments and compares their results against each other as well as against results of approaches of other related works. For a systematic evaluation of the classification approach, described in the two previous chapters, two orthogonal dimensions for experiments are defined:

1. **Event specifity**: To clarify the question whether disaster type specific classifiers or more generic cross-domain classifiers are more beneficial, training data containing tweets of only one type of disaster (i.e., deep event specifity) from training data including tweets of multiple different types of disasters (broad event specifity) are considered.

2. **Training size**: To clarify the impact of sample size on classifier performance in relation to event specifity, the amount of tweets used for training from 1K (small training size) to 28K (large training size) is distinguished.

Hence, experiments are categorized into four groups as shown in Figure 7.1. However, experiments using a large amount of tweets for training and a deeper event specificity had to be ruled out, since a comparable amount of tweets of the same disaster type is not available in the CrisisLexT26 data set. By comparing classification results based on the three remaining groups, evaluation of this thesis is able to show the impact of the disaster type, i.e., event specifity, as well as the impact of the amount of tweets used for training the classifier on classification accuracy. Thus, evaluation can show, how accurate the proposed cross-domain classifier performs on various events of different disaster types against classifiers trained on events of the same disaster type (e.g., trained on an earthquake event ”Guatemala earthquake” and tested on an earthquake event ”Costa Rica earthquake”).

The evaluation of different experiments is based on classification accuracy, i.e. how many Tweets are classified correctly. Hence, several values shown in tables of this chapter contain

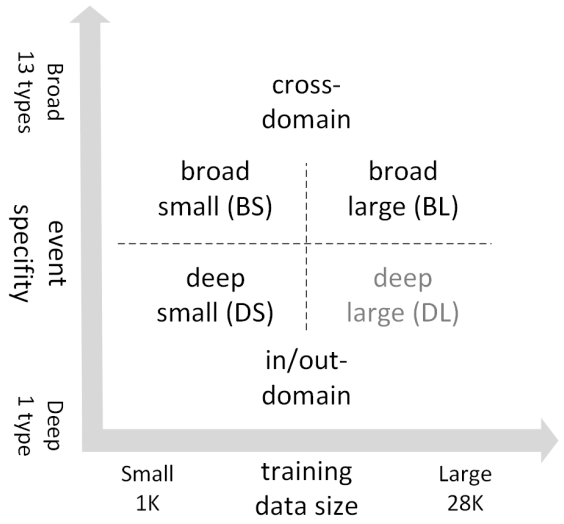


Figure 7.1: Evaluation dimensions

accuracy classification values. An accuracy values of one means every single Tweet of the test set was classified correctly, thus the aim to achieve for classification approaches are close to one. For a two class, informative and non informative, the simplest baseline accuracy value is 0.5, because statistically randomly guessing would achieve a classification accuracy of 0.5 if having equal prior probabilities.

7.1 Deep | Small experiments (DS)

In order to get an evaluation baseline for classification, first systematic in-domain and out-domain classification experiments by using all possible train/test-set combinations are considered, which results in a 26 times 26 event matrix shown in Figure 7.2. Regarding training data size, this kind of experiments use a rather small amount of approximately 1K tweets to train the classifier, since limited by the available CrisisLexT26 dataset. For those disaster types in the dataset which contain data about more than one event, crisis events of the same disaster type are used for training and testing the classifier, i.e., in-domain classification. For instance, for disaster type ”floods”, the event Alberta floods is used to train the classifier and the event Sardinia floods is used to test the classifier. In total, the CrisisLexT26 dataset allows 55 in-domain experiments, by using all possible train/test-set combinations of crisis events of the same disaster type, considering those types identified by Olteanu et al. [Olteanu et al., 2015]. Blue labels in figure 7.3 show in-domain results. Using all possible train/test-set combinations of crisis events of different disaster types result in additional 495 out-domain experiments, which are marked by red labels in figure 7.3. Figure 7.4 shows the average classification accuracy of all 650 in- and out-domain experiments with respect to one particular crisis event used for testing the classifier. All DS classification results are listed in Table7.1. In-domain experiments result on average in

Test \ Train	Alberta floods	Australia wildfires	Bohol earthquake	Boston bombings	Brazil nightclub fire	Colorado floods	Colorado wildfires	Costa Rica earthquake	Glasgow helicopter crash	Guatemala earthquake	Italy earthquakes	LA Airport shootings	Lac-Megantic train crash	Manila floods	NYC train crash	Philipinnes floods	Queensland floods	Russian meteor	Sardinia floods	Savar building collapse	Singapore haze	Spain train crash	Typhoon Pablo	Typhoon Yolanda	Venezuela refinery explosion	West Texas Explosion
Alberta floods	n.a.	0.72	0.62	0.60	0.50	0.74	0.65	0.49	0.58	0.65	0.66	0.63	0.64	0.73	0.71	0.73	0.71	0.52	0.61	0.47	0.50	0.68	0.71	0.64	0.64	0.57
Australia wildfires	0.77	n.a.	0.75	0.76	0.68	0.79	0.80	0.69	0.76	0.76	0.76	0.80	0.75	0.75	0.63	0.80	0.85	0.71	0.75	0.66	0.66	0.70	0.82	0.78	0.83	0.75
Bohol earthquake	0.84	0.84	n.a.	0.85	0.81	0.75	0.81	0.79	0.82	0.74	0.71	0.83	0.81	0.77	0.50	0.72	0.83	0.84	0.77	0.57	0.80	0.74	0.81	0.75	0.82	0.86
Boston bombings	0.78	0.80	0.81	n.a.	0.80	0.73	0.81	0.75	0.79	0.65	0.71	0.80	0.81	0.75	0.51	0.67	0.80	0.78	0.72	0.56	0.79	0.71	0.76	0.73	0.78	0.82
Brazil nightclub fire	0.85	0.84	0.85	0.86	n.a.	0.78	0.86	0.75	0.83	0.72	0.83	0.83	0.82	0.83	0.61	0.64	0.83	0.85	0.72	0.61	0.86	0.81	0.85	0.71	0.85	0.86
Colorado floods	0.77	0.80	0.68	0.70	0.66	n.a.	0.75	0.67	0.70	0.77	0.76	0.74	0.70	0.81	0.78	0.81	0.82	0.65	0.74	0.69	0.63	0.73	0.80	0.77	0.74	0.69
Colorado wildfires	0.78	0.83	0.74	0.77	0.75	0.80	n.a.	0.74	0.78	0.78	0.79	0.78	0.80	0.79	0.62	0.74	0.83	0.77	0.79	0.70	0.74	0.70	0.82	0.79	0.81	0.78
Costa Rica earthquake	0.84	0.88	0.79	0.83	0.86	0.84	0.86	n.a.	0.87	0.88	0.83	0.87	0.82	0.84	0.53	0.79	0.88	0.85	0.85	0.88	0.84	0.65	0.86	0.87	0.87	0.86
Glasgow helicopter crash	0.74	0.82	0.73	0.76	0.73	0.68	0.76	0.75	n.a.	0.72	0.76	0.80	0.74	0.71	0.62	0.68	0.83	0.81	0.75	0.70	0.72	0.64	0.74	0.81	0.78	0.74
Guatemala earthquake	0.75	0.80	0.74	0.70	0.75	0.78	0.76	0.82	0.76	n.a.	0.77	0.76	0.68	0.75	0.73	0.77	0.82	0.74	0.83	0.78	0.71	0.75	0.82	0.82	0.80	0.76
Italy earthquakes	0.70	0.67	0.68	0.64	0.63	0.70	0.69	0.63	0.66	0.70	n.a.	0.64	0.69	0.70	0.64	0.68	0.72	0.65	0.71	0.68	0.62	0.63	0.71	0.74	0.68	0.68
LA Airport shootings	0.80	0.85	0.79	0.83	0.80	0.80	0.84	0.79	0.84	0.81	0.81	n.a.	0.80	0.78	0.73	0.77	0.85	0.81	0.78	0.77	0.79	0.78	0.83	0.84	0.83	0.83
Lac-Megantic train crash	0.71	0.70	0.67	0.69	0.66	0.69	0.72	0.64	0.68	0.68	0.68	0.69	n.a.	0.70	0.63	0.65	0.69	0.67	0.69	0.60	0.65	0.66	0.72	0.66	0.71	0.69
Manila floods	0.79	0.73	0.70	0.65	0.58	0.75	0.72	0.55	0.64	0.76	0.76	0.63	0.75	n.a.	0.65	0.77	0.70	0.61	0.76	0.56	0.55	0.68	0.79	0.77	0.69	0.67
NYC train crash	0.76	0.78	0.63	0.74	0.77	0.85	0.80	0.75	0.75	0.86	0.75	0.81	0.73	0.69	n.a.	0.82	0.82	0.66	0.65	0.82	0.70	0.88	0.80	0.81	0.76	0.65
Philipinnes floods	0.67	0.67	0.50	0.44	0.39	0.65	0.51	0.42	0.42	0.59	0.59	0.48	0.52	0.74	0.76	n.a.	0.71	0.42	0.58	0.42	0.40	0.72	0.75	0.53	0.51	0.48
Queensland floods	0.79	0.88	0.76	0.78	0.71	0.82	0.82	0.72	0.78	0.76	0.80	0.80	0.78	0.80	0.66	0.83	n.a.	0.76	0.79	0.73	0.70	0.74	0.83	0.81	0.84	0.81
Russian meteor	0.74	0.78	0.77	0.77	0.75	0.72	0.75	0.75	0.77	0.73	0.71	0.76	0.74	0.71	0.53	0.69	0.77	n.a.	0.73	0.68	0.73	0.65	0.74	0.74	0.77	0.77
Sardinia floods	0.64	0.70	0.66	0.64	0.61	0.65	0.65	0.65	0.66	0.71	0.67	0.62	0.64	0.69	0.64	0.68	0.70	0.65	n.a.	0.68	0.60	0.65	0.69	0.71	0.67	0.66
Savar building collapse	0.68	0.82	0.73	0.69	0.68	0.74	0.71	0.85	0.82	0.85	0.73	0.78	0.69	0.75	0.62	0.71	0.80	0.80	0.83	n.a.	0.67	0.59	0.79	0.87	0.82	0.76
Singapore haze	0.74	0.77	0.71	0.78	0.80	0.71	0.79	0.79	0.78	0.67	0.78	0.76	0.79	0.71	0.49	0.58	0.75	0.77	0.75	0.78	n.a.	0.65	0.69	0.71	0.76	0.79
Spain train crash	0.76	0.79	0.72	0.76	0.76	0.79	0.77	0.76	0.76	0.79	0.71	0.78	0.71	0.75	0.81	0.78	0.78	0.73	0.76	0.69	0.75	n.a.	0.76	0.77	0.79	0.72
Typhoon Pablo	0.78	0.84	0.75	0.73	0.65	0.78	0.80	0.62	0.71	0.79	0.76	0.75	0.76	0.80	0.73	0.78	0.83	0.69	0.76	0.64	0.64	0.76	n.a.	0.78	0.79	0.74
Typhoon Yolanda	0.60	0.71	0.67	0.54	0.50	0.67	0.57	0.59	0.66	0.77	0.66	0.65	0.58	0.67	0.75	0.69	0.74	0.67	0.74	0.73	0.48	0.66	0.74	n.a.	0.61	0.62
Venezuela refinery explosion	0.72	0.76	0.73	0.72	0.69	0.68	0.74	0.68	0.71	0.74	0.72	0.71	0.71	0.73	0.59	0.66	0.74	0.71	0.73	0.58	0.67	0.74	0.76	0.67	n.a.	0.74
West Texas Explosion	0.81	0.82	0.82	0.83	0.80	0.78	0.82	0.79	0.78	0.71	0.70	0.80	0.80	0.78	0.51	0.70	0.82	0.80	0.74	0.60	0.79	0.76	0.78	0.72	0.81	n.a.

Figure 7.2: In-domain and out-domain results

an informativeness classification accuracy of 75% (including a standard deviation of 5%). Out of 55 in-domain experiments, the best classification accuracy of 88% achieved the Costa Rica earthquake using the Guatemala Earthquake as training event, which might be due to obvious similarities between these two events. The worst result of 58% gives the event Philippines flood using Sardinia floods as training set. Figure 7.2 visualizes results of all DS classification experiments. Red cells indicate a higher classification accuracy, blue cells a lower classification accuracy. In order to verify the hypothesis that a cross-domain classifier leads to at least as accurate informativeness classification as an in-domain classifier, i.e., a more specific one, the mentioned classification results serve as a baseline.

7.2 Broad | Small experiments (BS)

In order to eliminate the impact of training data size on classification accuracy when comparing results against DS experiments, in the second group of experiments the same

Test \ Train	Alberta floods	Australia wildfires	Bohol earthquake	Boston bombings	Brazil nightclub fire	Colorado floods	Colorado wildfires	Costa Rica earthquake	Glasgow helicopter crash	Guatemala earthquake	Italy earthquakes	LA Airport shootings	Lac-Megantic train crash	Manila floods	NYC train crash	Philipinnes floods	Queensland floods	Russian meteor	Sardinia floods	Savar building collapse	Singapore haze	Spain train crash	Typhoon Pablo	Typhoon Yolanda	Venezuela refinery explosion	West Texas Explosion
Alberta floods	0.84	0.72	0.62	0.60	0.50	0.74	0.65	0.49	0.58	0.65	0.66	0.63	0.64	0.73	0.71	0.73	0.71	0.52	0.61	0.47	0.50	0.68	0.71	0.64	0.64	0.57
Australia wildfires	0.77	0.90	0.75	0.76	0.68	0.79	0.80	0.69	0.76	0.76	0.76	0.80	0.75	0.75	0.63	0.80	0.85	0.71	0.75	0.66	0.66	0.70	0.82	0.78	0.83	0.75
Bohol earthquake	0.84	0.84	0.91	0.85	0.81	0.75	0.81	0.79	0.82	0.74	0.71	0.83	0.81	0.77	0.50	0.72	0.83	0.84	0.77	0.57	0.80	0.74	0.81	0.75	0.82	0.86
Boston bombings	0.78	0.80	0.81	0.86	0.80	0.73	0.81	0.75	0.79	0.65	0.71	0.80	0.81	0.75	0.51	0.67	0.80	0.78	0.72	0.56	0.79	0.71	0.76	0.73	0.78	0.82
Brazil nightclub fire	0.85	0.84	0.85	0.86	0.90	0.78	0.86	0.75	0.83	0.72	0.83	0.83	0.82	0.83	0.61	0.64	0.83	0.85	0.72	0.61	0.86	0.81	0.85	0.71	0.85	0.86
Colorado floods	0.77	0.80	0.68	0.70	0.66	0.91	0.75	0.67	0.70	0.77	0.76	0.74	0.70	0.81	0.78	0.81	0.82	0.65	0.74	0.69	0.63	0.73	0.80	0.77	0.74	0.69
Colorado wildfires	0.78	0.83	0.74	0.77	0.75	0.80	0.87	0.74	0.78	0.78	0.79	0.78	0.80	0.79	0.62	0.74	0.83	0.77	0.79	0.70	0.74	0.70	0.82	0.79	0.81	0.78
Costa Rica earthquake	0.84	0.88	0.79	0.83	0.86	0.84	0.86	0.93	0.87	0.88	0.83	0.87	0.82	0.84	0.53	0.79	0.88	0.85	0.85	0.88	0.84	0.65	0.86	0.87	0.87	0.86
Glasgow helicopter crash	0.74	0.82	0.73	0.76	0.73	0.68	0.76	0.75	0.89	0.72	0.76	0.80	0.74	0.71	0.62	0.68	0.83	0.81	0.75	0.70	0.72	0.64	0.74	0.81	0.78	0.74
Guatemala earthquake	0.75	0.80	0.74	0.70	0.75	0.78	0.76	0.82	0.76	0.91	0.77	0.76	0.68	0.75	0.73	0.77	0.82	0.74	0.83	0.78	0.71	0.75	0.82	0.82	0.80	0.76
Italy earthquakes	0.70	0.67	0.68	0.64	0.63	0.70	0.69	0.63	0.66	0.70	0.84	0.64	0.69	0.70	0.64	0.68	0.72	0.65	0.71	0.68	0.62	0.63	0.71	0.74	0.68	0.68
LA Airport shootings	0.80	0.85	0.79	0.83	0.80	0.80	0.84	0.79	0.84	0.81	0.81	0.90	0.80	0.78	0.73	0.77	0.85	0.81	0.78	0.77	0.79	0.78	0.83	0.84	0.83	0.83
Lac-Megantic train crash	0.71	0.70	0.67	0.69	0.66	0.69	0.72	0.64	0.68	0.68	0.68	0.69	0.78	0.70	0.63	0.65	0.69	0.67	0.69	0.60	0.65	0.66	0.72	0.66	0.71	0.69
Manila floods	0.79	0.73	0.70	0.65	0.58	0.75	0.72	0.55	0.64	0.76	0.76	0.63	0.75	0.88	0.65	0.77	0.70	0.61	0.76	0.56	0.55	0.68	0.79	0.77	0.69	0.67
NYC train crash	0.76	0.78	0.63	0.74	0.77	0.85	0.80	0.75	0.75	0.86	0.75	0.81	0.73	0.69	0.95	0.82	0.82	0.66	0.65	0.82	0.70	0.88	0.80	0.81	0.76	0.65
Philipinnes floods	0.67	0.67	0.50	0.44	0.39	0.65	0.51	0.42	0.42	0.59	0.59	0.48	0.52	0.74	0.76	0.92	0.71	0.42	0.58	0.42	0.40	0.72	0.75	0.53	0.51	0.48
Queensland floods	0.79	0.88	0.76	0.78	0.71	0.82	0.82	0.72	0.78	0.76	0.80	0.80	0.78	0.80	0.66	0.83	0.93	0.76	0.79	0.73	0.70	0.74	0.83	0.81	0.84	0.81
Russian meteor	0.74	0.78	0.77	0.77	0.75	0.72	0.75	0.75	0.77	0.73	0.71	0.76	0.74	0.71	0.53	0.69	0.77	0.83	0.73	0.68	0.73	0.65	0.74	0.74	0.77	0.77
Sardinia floods	0.64	0.70	0.66	0.64	0.61	0.65	0.65	0.65	0.66	0.71	0.67	0.62	0.64	0.69	0.64	0.68	0.70	0.65	0.81	0.68	0.60	0.65	0.69	0.71	0.67	0.66
Savar building collapse	0.68	0.82	0.73	0.69	0.68	0.74	0.71	0.85	0.82	0.85	0.73	0.78	0.69	0.75	0.62	0.71	0.80	0.80	0.83	0.94	0.67	0.59	0.79	0.87	0.82	0.76
Singapore haze	0.74	0.77	0.71	0.78	0.80	0.71	0.79	0.79	0.78	0.67	0.78	0.76	0.79	0.71	0.49	0.58	0.75	0.77	0.75	0.78	0.85	0.65	0.69	0.71	0.76	0.79
Spain train crash	0.76	0.79	0.72	0.76	0.76	0.79	0.77	0.76	0.76	0.79	0.71	0.78	0.71	0.75	0.81	0.78	0.78	0.73	0.76	0.69	0.75	0.86	0.76	0.77	0.79	0.72
Typhoon Pablo	0.78	0.84	0.75	0.73	0.65	0.78	0.80	0.62	0.71	0.79	0.76	0.75	0.76	0.80	0.73	0.78	0.83	0.69	0.76	0.64	0.64	0.76	0.91	0.78	0.79	0.74
Typhoon Yolanda	0.60	0.71	0.67	0.54	0.50	0.67	0.57	0.59	0.66	0.77	0.66	0.65	0.58	0.67	0.75	0.69	0.74	0.67	0.74	0.73	0.48	0.66	0.74	0.90	0.61	0.62
Venezuela refinery explosion	0.72	0.76	0.73	0.72	0.69	0.68	0.74	0.68	0.71	0.74	0.72	0.71	0.71	0.73	0.59	0.66	0.74	0.71	0.73	0.58	0.67	0.74	0.76	0.67	0.83	0.74
West Texas Explosion	0.81	0.82	0.82	0.83	0.80	0.78	0.82	0.79	0.78	0.71	0.70	0.80	0.80	0.78	0.51	0.70	0.82	0.80	0.74	0.60	0.79	0.76	0.78	0.72	0.81	0.89

Figure 7.3: In-domain (blue) vs. out-domain (red)

Event	Cross-Val. Acc.	Avg. Acc.	Min. Acc.	Max Acc.	# Tweets
Alberta floods	0.75	0.63	0.47	0.74	1376
Australia wildfires	0.84	0.75	0.63	0.85	1408
Bohol earthquake	0.86	0.78	0.50	0.86	1150
Boston bombings	0.80	0.74	0.51	0.82	1180
Brazil nightclub fire	0.87	0.79	0.61	0.86	1120
Colorado floods	0.86	0.73	0.63	0.82	1536
Colorado wildfires	0.82	0.77	0.62	0.83	1370
Costa Rica earthquake	0.89	0.83	0.53	0.88	1484
Glasgow helicopter crash	0.82	0.74	0.62	0.83	1156
Guatemala earthquake	0.88	0.77	0.68	0.83	1508
Italy earthquakes	0.77	0.67	0.62	0.74	1254
LA Airport shootings	0.86	0.81	0.73	0.85	1364
Lac-Megantic train crash	0.70	0.68	0.60	0.72	1118
Manila floods	0.84	0.69	0.55	0.79	1256
NYC train crash	0.91	0.76	0.63	0.88	1884
Philipinnes floods	0.88	0.55	0.39	0.76	1522
Queensland floods	0.88	0.78	0.66	0.88	1456
Russian meteor	0.78	0.73	0.53	0.78	1602
Sardinia floods	0.74	0.66	0.60	0.71	1264
Savar building collapse	0.90	0.75	0.59	0.87	1676
Singapore haze	0.81	0.73	0.49	0.80	1087
Spain train crash	0.85	0.76	0.69	0.81	1458
Typhoon Pablo	0.86	0.75	0.62	0.84	1356
Typhoon Yolanda	0.84	0.65	0.48	0.77	1530
Venezuela refinery explosion	0.77	0.71	0.58	0.76	1092
West Texas Explosion	0.83	0.76	0.51	0.83	1258
Total	**0.83**	**0.73**	**0.39**	**0.88**	**35456**

Table 7.1: In-domain and out-domain summery

amount of tweets as before is used, for training. These 1K tweets out of all crisis events are sampled, excluding the one crisis event used for testing the classifier, to address a broad event specificity. In general, cross-domain classification experiments follows a "leave one out" strategy, 25 crisis events are used for training and the remaining 26^{th} event was used for testing. The average classification accuracy over all 26 experiments is 79% (including a standard deviation of 7%), which is 4% higher than the average of in-domain experiments (cf. Figure 7.4, column "BS"). In other words, a *classifier trained on random sampled tweets* from various crisis events of different disaster types *achieves a 4% higher informativeness classification accuracy* than using a classifier *trained on the same disaster type* as the actual crisis is, even *using the same size* of tweets for training.

7.3 Broad | Large experiments (BL)

Since using tweets of different crisis events of different types of disasters for training lead to more accurate classification on average than in-domain training, in the third group of experiments evaluation wants to figure out the impact of the training size on classification accuracy. By applying a "leave one out" strategy, 26 experiments use all available 28K tweets, again excluding those used for testing, to train the classifier. Classification results

Cross-domain (leave-one-out)	**# Samples**		
Event	**10K**	**2K**	**1K**
Alberta floods	0.72	0.73	0.71
Australia wildfires	0.85	0.85	0.84
Bohol earthquake	0.86	0.86	0.86
Boston bombings	0.81	0.79	0.82
Brazil nightclub fire	0.86	0.87	0.85
Colorado floods	0.83	0.83	0.80
Colorado wildfires	0.82	0.83	0.82
Costa Rica earthquake	0.89	0.86	0.88
Glasgow helicopter crash	0.82	0.80	0.80
Guatemala earthquake	0.84	0.84	0.82
Italy earthquakes	0.74	0.72	0.73
LA Airport shootings	0.86	0.85	0.83
Lac-Megantic train crash	0.71	0.72	0.72
Manila floods	0.82	0.79	0.82
NYC train crash	0.84	0.82	0.82
Philipinnes floods	0.60	0.58	0.62
Queensland floods	0.86	0.87	0.85
Russian meteor	0.78	0.77	0.77
Sardinia floods	0.70	0.69	0.69
Savar building collapse	0.83	0.78	0.79
Singapore haze	0.72	0.75	0.77
Spain train crash	0.79	0.78	0.78
Typhoon Pablo	0.84	0.84	0.82
Typhoon Yolanda	0.75	0.70	0.63
Venezuela refinery explosion	0.76	0.76	0.77
West Texas Explosion	0.83	0.83	0.82
Total	**0.80**	**0.79**	**0.79**

Table 7.2: Cross-domain training with samples

			Accuracy		
Event	**Acc.**	**# Tweets**	**Δ avg.**	**Δ min.**	**Δ max.**
Alberta floods	0.73	32430	10.02	25.50	-1.39
Australia wildfires	0.86	32398	11.16	22.75	1.42
Bohol earthquake	0.87	32956	9.36	37.20	1.00
Boston bombings	0.82	32970	7.40	30.36	0.30
Brazil nightclub fire	0.87	32922	7.26	25.65	0.20
Colorado floods	0.85	32270	11.12	21.70	2.80
Colorado wildfires	0.83	32436	6.27	21.08	0.33
Costa Rica earthquake	0.89	32322	6.20	36.40	0.71
Glasgow helicopter crash	0.83	32650	8.97	20.91	0.55
Guatemala earthquake	0.84	32298	7.89	16.67	1.14
Italy earthquakes	0.74	32552	6.26	11.28	-0.40
LA Airport shootings	0.87	32442	6.06	13.47	1.55
Lac-Megantic train crash	0.71	32688	2.86	10.96	-1.59
Manila floods	0.83	32550	14.06	28.30	3.80
NYC train crash	0.87	31922	10.38	24.00	-1.24
Philipinnes floods	0.62	32284	6.65	22.70	-14.30
Queensland floods	0.87	32350	9.23	21.67	-0.83
Russian meteor	0.78	32522	4.83	24.81	0.28
Sardinia floods	0.70	32542	3.90	10.09	-1.30
Savar building collapse	0.84	32130	9.18	25.66	-2.55
Singapore haze	0.71	32884	-2.19	22.50	-9.40
Spain train crash	0.81	32348	4.76	11.60	-0.70
Typhoon Pablo	0.85	32450	10.39	22.60	1.10
Typhoon Yolanda	0.76	32276	10.74	27.96	-0.86
Venezuela refinery explosion	0.76	32714	5.90	17.98	0.60
West Texas Explosion	0.83	32844	6.56	31.71	-0.45
Total	**0.80**	**845150**	**7.51**	**22.52**	**-0.74**

Table 7.3: Cross-domain results "leave-one-out"

are visualized in Figure 7.4, column "BL". On average, classification accuracy over all cross-domain experiments is 80% (including a standard deviation of 7%), which is significantly higher compared to in- and out-domain experiments. Comparing the average classification results using 28K tweets for training against using 1K tweets, show a slightly, 1%, higher classification accuracy. Thus, the interesting finding here is that the size of the dataset used for training seems not to be primary relevant for accurate classification of informativeness.

7.4 Interpretation of evaluation

To sum up, experimentation results allow us the following *conclusions*:

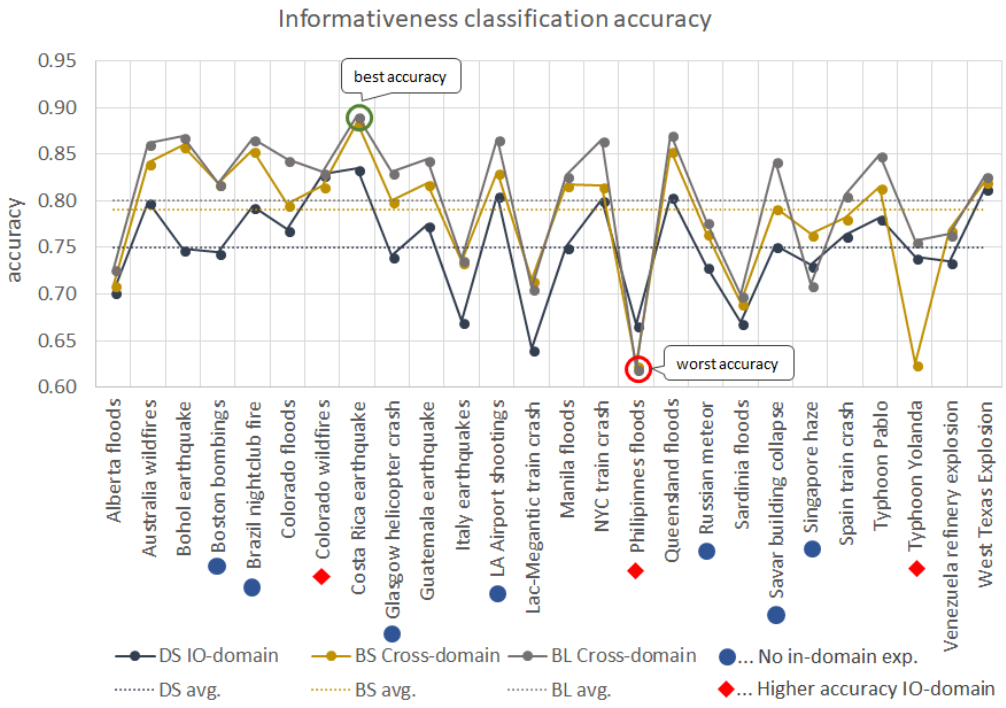


Figure 7.4: Classification accuracy results

1. Using a classifier trained on *various* events cross different types of disasters outperforms in 23 cases out of 26 (cf. red diamonds in Figure 7.4) more specific classifiers, trained on the same disaster type, in classification accuracy of 4% (avg.). This even with the same amount of tweets in the training data.

2. Using a more specifically trained classifier may tend to overfit and therefore leads to less accurate informativeness classification of unseen crisis events.

3. Increasing the amount of training samples on average lead to slightly more accurate classification only (cf. Figure 7.4).

7.5 Classification performance

Finally, when considering informativeness classification results of other closely related approaches, which are discussed in more detail in the following related work section, as a baseline, results show that the cross-domain classifier presented in the current work achieved a higher accuracy on average as well as regarding best accuracy values (cf. Table 4). In particular, compared to cross-domain trained classifiers of [Acerbo and Rossi, 2017] as well as to the in-domain classifier of [Longhini et al., 2017], our cross-domain classifier is 4% (avg.) more accurate in informativeness classification, and compared to [Imran et al., 2016] 1.3%

more accurate. In the current work, the best average accuracy is 80% (including a standard deviation of 7%) using cross-domain training over all events, best accuracy classification result is 89%, which is compared against [Imran et al., 2016] an accuracy improvement of 2% and 12% compared against [Longhini et al., 2017]. The worst result was classifying Philippines flood with an accuracy of 62%.

7.6 Time-dependent classification

As introduced in Chapter 5 in Section 5.4.1.2, the current implementation includes settings to split classification regarding the time. In other words, instead using a single trained classifier to classify tweets of the test set, three different trained classifier are used to classify tweets of the three-times-split test set. The aim of this experiment is to emphasize the aspect of time during crisis events. The experimental setting includes two runs over all events, by using leave-one-out. Training events as well as the test event are split into three parts, which leads to three different classifiers. One for early time phases during the event, one for middle time phases and one for late time phases. Intervals, for instance from day 0 to day 3, are static defined since for online classification no dynamic time intervals are possible. Results shown in Table 7.4 lead to some interesting conclusions. Accuracy values are averaged weighted accuracy scores of all 3 time dependent classification stages.

On average, the three way time dependent classification approach does not lead to significant improvements over single trained cross-domain classifier. However, for events with low classification performance, for instance Philippines floods or the Lac-Megantic train crash, the 3-step-time dependent classification approach leads to significant improvements. In the case of classifying tweets from the Philippines flood, classification improved from 62% using single classification to 83% using time dependent classification. However, for many events the single classification perform better. Using more sophisticated time intervals or further classification steps, i.e., mapping selected tweets to the best suited classifier, might lead to overall improvements and might be suggestions for future work.

To sum op, time dependent classification can lead to accuracy improvements, especially for data, which do not perform very well using single classification or extraordinary events, which does not fit very well to the training data. More future work in this field might lead to multi-step classification where for each tweet the best suitable classifier for prediction is selected.

Event	**Exp. 1**[a]	**Δ Acc. table 7.3**	**Exp. 2** [b]	**Δ Acc. table 7.3**
Alberta floods	0.73	0.00	0.70	2.49
Australia wildfires	0.85	0.83	0.85	1.08
Bohol earthquake	0.86	1.00	0.86	0.50
Boston bombings	0.80	1.39	0.80	1.79
Brazil nightclub fire	0.86	1.10	0.86	1.00
Colorado floods	0.84	0.32	0.84	0.62
Colorado wildfires	0.81	2.08	0.82	0.75
Costa Rica earthquake	0.90	-0.71	0.90	-0.42
Glasgow helicopter crash	0.82	1.00	0.82	1.36
Guatemala earthquake	0.86	-1.52	0.86	-1.52
Italy earthquakes	0.74	-0.30	0.76	-2.40
LA Airport shootings	0.87	-0.10	0.87	0.10
Lac-Megantic train crash	0.69	1.49	0.70	0.80
Manila floods	0.81	1.40	0.79	3.80
NYC train crash	0.85	1.24	0.85	1.43
Philippines floods	0.79	-16.70	0.83	-20.90
Queensland floods	0.87	0.58	0.88	-0.83
Russian meteor	0.78	-0.62	0.78	0.21
Sardinia floods	0.72	-2.30	0.71	-1.50
Savar building collapse	0.88	-3.27	0.87	-2.39
Singapore haze	0.70	1.10	0.71	-0.40
Spain train crash	0.82	-1.30	0.83	-1.90
Typhoon Pablo	0.84	1.20	0.84	0.80
Typhoon Yolanda	0.83	-6.87	0.86	-9.83
Venezuela refinery explosion	0.76	0.00	0.77	-0.30
West Texas Explosion	0.82	0.36	0.83	-0.09
Total	**0.81**	**-0.71**	**0.81**	**-0.99**

Table 7.4: Time-dependent classification summery

[a] time dependent split according intervals [0<=x<3][3<=x<10][10<=x]
[b] time dependent split according intervals [0<=x<2][2<=x<5][5<=x]

Chapter 8

Lessons learned and open issues

This master thesis deals with informativeness classification of potentially crisis-related tweets. Previous chapters discuss the main contributions of this work: Firstly, a detailed analyses of data with respect to informativeness, by investigating temporal, spatial, linguistic and source aspects of tweets included in the CrisisLexT26 dataset. Secondly, this thesis operationalized the outcomes of data analysis by proposing a cross-domain informativeness classifier, providing the following benefits:

1. It is usable on various events of various types of disasters so that a single classifier is applicable for any event.

2. It achieves 4% (avg.) higher classification accuracy than disaster-type specific classifiers using the same size of training data.

3. It increases the amount of available training data since being not limited to one type of disaster.

The current chapter wants to summarize the most important aspects and results discussed in this thesis. Moreover, the open issues section provides an outlook for possible further research in this area.

8.1 Lessons learned

Situation awareness during disaster situations is crucial for organizations and authorities, since it is the basis of an effective crisis management and allows appropriate decisions and reactions to support affected people. Social Media services like Twitter can contribute to situation awareness, because people communicate a lot even in disaster situations via Social Media. Information about affected people, damaged infrastructure, affected ares are

shared by people during crisis events. However, online communication includes thoughts, prayers, gratitude as well as lot of emotional messages too, which do not contribute to situation awareness at all. Hence, filtering only informative tweets, i.e., informative in the sense of information, which can contribute to gaining situation awareness, is a major task. Due to the huge amount of data, this kind of filtering task is not manageable by humans, instead supervised machine learning algorithms are used to classify messages in whether they contain informative or non informative content.

While many research focus on feature-engineering, i.e., inventing and evaluating new features, [Khare et al., 2018][Khare et al., 2017][Longhini et al., 2017] or classification itself [Pekar et al., 2016][Stowe et al., 2016][Verma et al., 2011], only few research deal with crisis data analysis *with respect to informativeness*. Knowing the underlying data is essential in machine learning to gain reliable classification results. This thesis analyzed a large dataset containing 26 different crisis events regarding temporal, spatial, linguistic and source aspects with respect to informativeness. In addition, the CrisisLexT26 dataset used in this thesis is one of the largest collection of crisis events, including a variety of different event types, available, compared to other work in this field. Firstly, data analysis uncovers tweet characteristics having impact on informativeness thus being candidates to be used as features. Secondly, analysis does not show significant differences with respect to event types and furthermore does not show significant similarities within same event types, which allows accurate cross-domain classification thus it is not necessary to train a classifier for each event type.

However, these conclusions are related to informativeness classification and might not be the case for further information mining of crisis related tweets, for instance entity annotation in tweet text. Informativeness does not focus primary on semantics in tweets, rather informativeness classification use statistical properties or characteristics to classify tweets in informative and not informative messages. This is one reason why for informativeness classification the event type is not considerable important.

Overall, this thesis uses the insights gained through detailed data analysis to propose a cross-domain classification approach which can be used for various event types. Features used are mainly based on linguistic characteristics of tweets, which differ not significantly over various event types. Classification experiments showed that cross-domain training leads to significant better classification results than in- and out-domain training. In addition, data size matters, a larger amount of training samples lead to slightly better classification results even if there are a diversity of event types included. People's behavior is similar in crisis situations, although they are different particular events. Therefore, a more heterogeneous training set in the sense of variations of tweets, provided by different types of events, is more suitable to classify new and unseen events than using event type adapted training.

8.2 Critical reflection

However, those conclusions made in the previous section, in particular analysis and experiments realized in the current thesis, require a critical reflection. Firstly, the comparison of classification performance of different approaches is based on average classification accuracy values. However, average values do not show the distribution of classification results, which measures of dispersion, such as variance and standard deviation, do. For instance, Figure 7.4 shows average classification accuracy values of events of in-domain and out-domain experiments, yet classification accuracy of some events is better than of others, which is shown in more detail in Figure 7.2. Secondly, there might be further possible improvements of classification accuracy when additionally experimenting with different classification algorithms and their settings. Although in the current thesis several different algorithms are tried out, there might be other classification algorithms, such as Deep Neural Networks, as well as different parameter settings, which might lead to better results. Thirdly, part of features used, like POS or sentiment of tweets, are designed for English, yet applied on all tweets of the dataset, which might bias informativeness classification of tweets of other languages than English. Lastly, sampling strategies used in the current thesis for experiments might bias classification accuracy as well, since tweets used for learning the classifier are different when new tweets are sampled out of the entire dataset.

8.3 Limitations and open issues

This thesis focuses on informativeness classification, only. However, this is only one early step in using Social Media services to support disaster management and the decision making process.

Since, however, evaluation of classification is based on the 26 crisis events and 13 disaster types included in the CrisisLexT26 data set, only. Hence, results might be different when classifying new disaster types. In addition, since a majority of tweets in the CrisislexT26 data set are in English, it can be expected that classification is not as accurate in classifying tweets of other languages thus limiting the classifiers applicability. Classifying tweets of other languages requires to adjust linguistic features (e.g. POS, sentiment) to particular languages. Moreover, the CrisislexT26 data set contained tweets stemming from the years 2012 to 2013. Yet, communication of people might change over years and thus informativeness classification accuracy using classifiers trained on past events happening years ago might decrease over time. Since the CrisisLexT26 dataset does not provide any user related information, user based features are not used for classification. Yet, user-related information like user meta data or geo-location are expected to further improve the overall

quality of classification, but has not been considered in the current work since not being included in the CrisisLexT26 dataset. In our work we use mainly linguistic, emotional and sentimental features. Experiments including additional (semantic) features as well as geo-location information, which could help increase the accuracy of the classifier, is subject to work. Another line of intended research is the consideration of other languages, crisis data sets and finally also the expansion of our evaluation towards other domains like the classification of news articles or web documents realizing a domain-generic informativeness classification approach.

Moreover, information source is not explicitly given by Twitter even though the source is implicitly known by the user or the user profile, respectively. Since the current classification approach is designed for online classification, it is important to determine the source of the tweet immediately when classifying the tweet. Supervised Machine Learning is one solution to deal with this issue and might be necessary for future works.

Finally, the time dependent classification approach introduced in previous chapters point out that informativeness classification is not limited to state-of-the-art classification approaches. Using multiple classifiers for different stages during a crisis and additionally use an intelligent way to map portions of tweets to the best suitable classifier might lead to novel designs of classification approaches. For future works this seems to be a solid basis to build on.

Bibliography

[Acerbo and Rossi, 2017] Acerbo, F. and Rossi, C. (2017). Filtering Informative Tweets During Emergencies: A Machine Learning Approach. In *Proc. of the 1st CoNEXT Workshop on ICT Tools for Emergency Networks and Disaster Relief*, I-TENDER '17, pages 1–6, New York, NY, USA. ACM.

[Avvenuti et al., 2016] Avvenuti, M., Cimino, M., Cresci, S., Marchetti, A., and Tesconi, M. (2016). A framework for detecting unfolding emergencies using humans as sensors. *SpringerPlus*, 5.

[Bishop, 2006] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

[Cameron et al., 2012] Cameron, M., Power, R., Robinson, B., and Yin, J. (2012). Emergency Situation Awareness from Twitter for Crisis Management. In *Proc. of the 21st Int. Conf. on World Wide Web*, WWW '12, pages 695–698, New York, NY, USA. ACM.

[Cresci et al., 2015] Cresci, S., Tesconi, M., Cimino, A., and Dell'Orletta, F. (2015). A linguistically-driven approach to cross-event damage assessment of natural disasters from social media messages. In *Proc. of the 24th Int. Conf. on World Wide Web*, WWW '15, pages 1195–1200. ACM.

[Dai et al., 2007] Dai, W., Xue, G., Yang, Q., and Yu, Y. (2007). Transferring naive bayes classifiers for text classification. In *Proc. of the 22nd Int. Conf. on Association for the Advancement of Artificial Intelligence*, volume 7 of *AAAI '07*, pages 540–545.

[Derczynski et al., 2018] Derczynski, L., Meesters, K., Bontcheva, K., and Maynard, D. (2018). Helping Crisis Responders Find the Informative Needle in the Tweet Haystack. *arXiv preprint arXiv:1801.09633*.

[Girtelschmid et al., 2016] Girtelschmid, S., Salfinger, A., Pröll, B., Retschitzegger, W., and Schwinger, W. (2016). Near real-time detection of crisis situations. In *Proc. of 39th Int. Conv. on Information and Communication Technology, Electronics and Microelectronics*, MIPRO '16, pages 247–252. IEEE.

[Graf, 2017] Graf, D. (2017). Semeval-2017-t8.

[Graf et al., 2018] Graf, D., Retschitzegger, W., Schwinger, W., Pröll, B., and Kapsammer, E. (2018). Cross-domain Informativeness Classification for Disaster Situations. In *Proc. of the 10th Int. Conf. on Management of Digital Ecosystems*, accepted for MEDES '18. ACM.

[Gundecha and Liu, 2012] Gundecha, P. and Liu, H. (2012). Mining social media: a brief introduction. In *New Directions in Informatics, Optimization, Logistics, and Production*, pages 1–17. Informs.

[Horn et al., 2013] Horn, C., Zhila, A., Gelbukh, A., Kern, R., and Lex, E. (2013). Using factual density to measure informativeness of web documents. In *Proc. of the 19th Nordic Conf. of Computational Linguistics*, NODALIDA '13, pages 227–238.

[Imran et al., 2015] Imran, M., Castillo, C., Diaz, F., and Vieweg, S. (2015). Processing social media messages in mass emergency: a survey. *ACM Computing Surveys (CSUR)*, 47.

[Imran et al., 2013] Imran, M., Elbassuoni, S., Castillo, C., Diaz, F., and Meier, P. (2013). Extracting information nuggets from disaster-Related messages in social media. In *Proc. of the 10th Conf. for Information Systems for Crisis Response and Management*, ISCRAM '13.

[Imran et al., 2016] Imran, M., Mitra, P., and Srivastava, J. (2016). Cross-language domain adaptation for classifying crisis-related short messages. *arXiv preprint arXiv:1602.05388*.

[Kanhabua and Nejdl, 2013] Kanhabua, N. and Nejdl, W. (2013). Understanding the Diversity of Tweets in the Time of Outbreaks. In *Proc. of the 22Nd Int. Conf. on World Wide Web*, WWW '13, pages 1335–1342, New York, NY, USA. ACM.

[Khare et al., 2018] Khare, P., Burel, G., and Alani, H. (2018). Classifying crises-information relevancy with semantics. In *Proc. of the 15th European Semantic Web Conference*, ESWC '18, pages 367–383. Springer.

[Khare et al., 2017] Khare, P., Fernandez, M., and Alani, H. (2017). Statistical Semantic Classification of Crisis Information. *1st workshop of Hybrid Statistical Semantic Understanding and Emerging Semantics (HSSUES), 16th Int. Semantic Web Conf.*

[Kireyev, 2009] Kireyev, K. (2009). Semantic-based estimation of term informativeness. In *Proc. of the 2009 Conf. of the North American Chapter of the Association for Com-*

putational Linguistics, NAACL-HLT '09, pages 530–538. Association for Computational Linguistics.

[Kwak et al., 2010] Kwak, H., Lee, C., Park, H., and Moon, S. (2010). What is Twitter, a social network or a news media? In *Proc. of the 19th Int. Conf. on World wide web*, WWW '10, pages 591–600. ACM.

[Li et al., 2015] Li, H., Guevara, N., Herndon, N., Caragea, D., Neppalli, K., Caragea, C., Squicciarini, A., and Tapia, A. (2015). Twitter Mining for Disaster Response: A Domain Adaptation Approach. In *Proc. of the 12th Conf. for Information Systems for Crisis Response and Management*, ISCRAM '15.

[Lloret and Palomar, 2016] Lloret, E. and Palomar, M. (2016). Analysing and evaluating the task of automatic tweet generation: knowledge to business. *Computers in Industry*, 78:3–15.

[Longhini et al., 2017] Longhini, J., Rossi, C., Casetti, C., and Angaramo, F. (2017). A language-agnostic approach to exact informative tweets during emergency situations. In *Int. Conf. on Big Data*, Big Data '17, pages 3475–3739. IEEE.

[Mohammad et al., 2013] Mohammad, S., Kiritchenko, S., and Zhu, X. (2013). NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. *arXiv preprint arXiv:1308.6242*.

[Munro and Manning, 2012] Munro, R. and Manning, C. D. (2012). Short message communications: users, topics, and in-language processing. In *Proc. of the 2nd ACM Symposium on Computing for Development*, page 4. ACM.

[Ning et al., 2017] Ning, X., Yao, L., Wang, X., and Benatallah, B. (2017). Calling for Response: Automatically Distinguishing Situation-Aware Tweets During Crises. In *Proc. of the 13th Int. Conf. on Advanced Data Mining and Applications*, ADMA '17, pages 195–208. Springer.

[Olteanu, 2016] Olteanu, A. (2016). Probing the limits of social data.

[Olteanu et al., 2014] Olteanu, A., Castillo, C., Diaz, F., and Vieweg, S. (2014). CrisisLex: A Lexicon for Collecting and Filtering Microblogged Communications in Crises. In *Proc. of the 8th Conf. on Weblogs and Social Media*, ICSWM '14.

[Olteanu et al., 2015] Olteanu, A., Vieweg, S., and Castillo, C. (2015). What to expect when the unexpected happens: Social media communications across crises. In *Proc. of*

the 18th ACM Conf. on Computer Supported Cooperative Work & Social Computing, CSCW '15, pages 994–1009. ACM.

[Pedregosa et al., 2011] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine Learning in {P}ython. *Journal of Machine Learning Research*, 12:2825–2830.

[Pekar et al., 2016] Pekar, V., Binner, J., Najafi, H., and Hale, C. (2016). Selecting Classification Features for Detection of Mass Emergency Events on Social Media. In *Proc. of the 15th Int. Conf. on Security and Management*, SAM '16, page 192. The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp).

[Pogrebnyakov and Maldonado, 2017] Pogrebnyakov, N. and Maldonado, E. (2017). Identifying emergency stages in facebook posts of police departments with convolutional and recurrent neural networks and support vector machines. In *Int. Conf. on Big Data*, Big Data '17, pages 4343–4352. IEEE.

[Ren et al., 2017] Ren, X., Wu, Z., He, W., Qu, M., Voss, C., Ji, H., Abdelzaher, T., and Han, J. (2017). CoType: Joint extraction of typed entities and relations with knowledge bases. In *Proc. of the 26th Int. Conf. on World Wide Web*, WWW '17, pages 1015–1024. Int. World Wide Web Conf. Steering Committee.

[Russell and Norvig, 2016] Russell, S. J. and Norvig, P. (2016). *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited,.

[Salfinger, 2016] Salfinger, A. (2016). *Staying Aware in an Evolving World*. PhD thesis, Johannes Kepler University of Linz.

[Salfinger et al., 2016a] Salfinger, A., Salfinger, C., Pröll, B., Retschitzegger, W., and Schwinger, W. (2016a). Pinpointing the Eye of the Hurricane-Creating A Gold-Standard Corpus for Situative Geo-Coding of Crisis Tweets Based on Linked Open Data. In *LDL 2016 5th Workshop on Linked Data in Linguistics: Managing, Building and Using Linked Language Resources*, page 27.

[Salfinger et al., 2016b] Salfinger, A., Schwinger, W., Retschitzegger, W., and Pröll, B. (2016b). Mining the disaster hotspots-situation-adaptive crowd knowledge extraction for crisis management. In *Proc. of the 2016 Multi-Disciplinary Int. Conf. on Cognitive Methods in Situation Awareness and Decision Support*, CogSIMA '16, pages 212–218. IEEE.

[Stowe et al., 2016] Stowe, K., Paul, M., Palmer, M., Palen, L., and Anderson, K. (2016). Identifying and categorizing disaster-related tweets. In *Proc. of The 4th Int. Workshop on Natural Language Processing for Social Media*, pages 1–6.

[Verma et al., 2011] Verma, S.and Vieweg, S., Corvey, W., Palen, L., Martin, J., Palmer, M., Schram, A., and Anderson, K. (2011). Natural Language Processing to the Rescue? Extracting" Situational Awareness" Tweets During Mass Emergency. In *Proc. of the 5th Conf. on Weblogs and Social Media*, ICWSM '11.

[Vieweg, 2012] Vieweg, S. (2012). *Situational awareness in mass emergency: A behavioral and linguistic analysis of microblogged communications*. PhD thesis, University of Colorado at Boulder.

[Vieweg et al., 2010] Vieweg, S., H., A. L., Starbird, K., and Palen, L. (2010). Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In *Proc. of the Conf. on Human Factors in Computing Systems*, CHI '10, pages 1079–1088. ACM.

[Wong and Kit, 2011] Wong, B. and Kit, C. (2011). Comparative evaluation of term informativeness measures for machine translation evaluation metrics. In *Proc. of the 13th Conf. of Machine Translation Summit*, volume 2011, pages 537–544.

[Wu and Giles, 2013] Wu, Z. and Giles, C. (2013). Measuring term informativeness in context. In *Proc. of the 2013 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '13, pages 259–269.

[Zubiaga et al., 2015] Zubiaga, A., Liakata, M., Procter, R., Bontcheva, K., and Tolmie, P. (2015). Towards detecting rumours in social media. *AAAI Workshop: AI for Cities*.

Curriculum Vitae

David Graf, BSc

Personal Information

Date of Birth: 30. Juli 1988

Citizenship: Austria

Education

since 08/2016 Master studies in Computer Science at Johannes Kepler University, Linz, Austria.

08/2017 – 12/2017 Exchange semester in Computer Science at Northwest Missouri State University, Maryville-MO, United States.

10/2013 – 08/2016 Bachelor of Science in Business Informatics at Johannes Kepler University, Linz, Austria.

09/2002 – 06/2007 Technical school for Electronics and Informatics at HTBLA, Steyr, Austria.

Professional Experience

since 11/2008 Computer Scientist at Meierhofer Österreich GmbH, St.Valentin, Austria.

11/2017 – 12/2017 Research Assistant at Department of Computer Science, Northwest Missouri State University, Maryville-MO, United States.

since 10/2007 Paramedic at Austrian Red Cross, St.Valentin, Austria.

Publications

2018 Graf, D., Retschitzegger, W., Schwinger, W., Pröell, B., and Kapsammer, E. (2018). Cross-domain Informativeness Classification for Disaster Situations. accepted for: *In Proc. of the 10th Int. Conf. on Management of Digital Ecosystems*, MEDES' 18. ACM.

2017 Graf, D., Oppl, S. and Eckmaier, A. (2017). Towards BPM Skill Assessment Using Computerized Adaptive Testing, *In Proc. of the 9th Conf. on Subject-oriented Business Process Management*. S-BPM ONE'17. ACM.

30$^{\text{th}}$ of August, 2018 Linz