UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN

A DOMAIN-INDEPENDENT AND LANGUAGE-AGNOSTIC APPROACH FOR CRISIS EVENT DETECTION AND UNDERSTANDING

TESIS PARA OPTAR AL GRADO DE
DOCTOR EN COMPUTACIÓN

HERNÁN ANDRÉS SARMIENTO ALBORNOZ

PROFESOR GUÍA:
FELIPE BRAVO MARQUEZ
PROFESORA GUÍA 2:
BÁRBARA POBLETE LABRA

MIEMBROS DE LA COMISIÓN:
YELENA MEJOVA
MARCELO MENDOZA ROCHA
JOSÉ PINO URTUBIA

Este trabajo ha sido parcialmente financiado por ANID - Millennium Science Initiative Program - Code ICN17 002 y ANID/Scholarship Program/DOCTORADO BECAS CHILE/2020 - 21201101

SANTIAGO DE CHILE
2023

Resumen

Un Enfoque Independiente del Dominio y Agnóstico al Idioma para la Detección y Comprensión de Eventos de Crisis

Durante una crisis, los usuarios de medios sociales comparten actualizaciones que pueden ayudar a mejorar el conocimiento del evento. Esto ha motivado a investigadores de múltiples campos de emergencias a estudiar el comportamiento de la propagación de la información en línea. Sin embargo, existen varios retos relacionados con las características de los medios sociales, por ejemplo, los datos no estructurados y ruidosos, el procesamiento de grandes colecciones de mensajes, rumores y información falsa, entre otros.

La mayoría de los estudios realizados se han centrado en la caracterización y detección de crisis a través de redes sociales. Por lo general, han analizado eventos específicos en lugar de estudiar los patrones transversales que surgen de las conversaciones durante las crisis. Además, los estudios han considerado los mensajes en inglés como idioma principal debido a la disponibilidad de recursos y datos. Sin embargo, las crisis suelen producirse en países (por ejemplo, Chile e Italia) en los que las lenguas no inglesas son el idioma principal (español e italiano, respectivamente). Por lo tanto, existe una brecha en comprender las crisis en diferentes dimensiones, como los idiomas, dominios y ubicaciones geográficas.

En esta tesis, presentamos un estudio a gran escala de las crisis debatidas en medios sociales. Nuestro objetivo es descubrir y comprender patrones de comunicación relacionados a crisis en diferentes tipos de eventos, ubicaciones e idiomas. En esta línea, investigamos en tres áreas: 1) proponemos metodologías para caracterizar y descubrir patrones generales de mensajes de medios sociales en una diversidad de crisis y que han ocurrido en diferentes localidades e idioma. 2) realizamos un estudio para clasificar mensajes de crisis teniendo en cuenta una evaluación experimental entre idiomas y dominios. 3) implementamos un método para detectar eventos de crisis que es agnóstico del tipo de evento e idioma de los mensajes.

Los principales resultados de este trabajo son: 1) usando representaciones compactas de mensajes para múltiples crisis, podemos diferenciarlas con una precisión del 75%. 2) implementamos un método para identificar y analizar discusiones - con poca intervención humana - que surgen en crisis de larga duración. 3) demostramos que es posible aprovechar datos de idiomas con altos recursos (ej: inglés) para clasificar los mensajes de otros idiomas (de pocos recursos) con un F1-score promedio de 80%. Al introducir mensajes de un nuevo dominio de crisis, la clasificación alcanza un F1-score de 82%. 4) implementamos un método de detección de crisis que los identifica basándose en anomalías en la actividad de localidades en medios sociales, detectando hasta 80% independiente del dominio y agnóstico al idioma.

Abstract

During a crisis, social media users share timely updates that can help improve situational awareness. This has motivated researchers from multiple emergency-related fields to study the behavior of online information propagation. However, there are several challenges related to intrinsic characteristics of social media, for instance, unstructured and noisy data, processing large collections of messages, rumors and false information, among others.

Most studies until now have focused on characterizing and detecting crises using social media. They have generally analyzed specific events instead of studying transversal patterns that emerge from online conversations during crises. Additionally, studies have considered English messages as the main language because of the availability of resources (e.g., Natural Language Processing tools) and data. However, crises usually occur in countries (e.g., Chile and Italy) where non-English languages are the primary language (Spanish and Italian, respectively). Nevertheless, there is a gap in understanding crisis communications across several dimensions such as languages, domains, and geographic locations.

In this thesis, we present a large-scale study of crises discussed through social media platforms. Our goal is to discover and understand universal communication patterns related to crises across different types of events, geographic locations, and languages. Along this line, we research in three areas: 1) we propose methodologies to characterize and discover general patterns of social media messages in a diversity of crises. 2) we conduct an extensive study to classify crisis-related messages by considering a cross-lingual and cross-domain experimental evaluation. 3) we implement a method for detecting crisis events that is agnostic to the type of event and the language of the message.

The main findings obtained from this work are: 1) using compact representations of crisis-related messages, we can differentiate them with 75% accuracy; 2) we implement a method to identify and analyze discussions - with weak human intervention - that emerge in long-term crises. This method can be generalized for any language and type of event; 3) we demonstrate that it is possible to leverage data from high-resource languages (i.e., English) to classify messages from other (low-resource) languages with an average F1-score of 80%. We also noted that by introducing messages from a different new (previously unseen) crisis domain, the classification performance is improved, reaching an F1-score of 82%; 4) we implement a crisis detection method that identifies events based on location-based user activity anomalies by identifying up to 80% of events from a domain-independent and language-agnostic perspective.

A todas las personas que fueron parte del proceso.
En especial, a mi familia.

Acknowledgments

I find it challenging to express my gratitude to everyone who has contributed to my academic journey, which began with my master's degree and partially culminated in my Ph.D. degree. Therefore, I would like to begin by thanking everyone who has interacted with me, shared their knowledge and experiences, and provided advice and support during this journey.

I am extremely grateful to my family for their unwavering support, encouragement, and patience. They have supported me through all my endeavors, even during the most challenging times, and have always believed in me. I realize that pursuing a Ph.D. can be difficult for them to understand, but they have never questioned my decisions and have always been there for me. I would also like to express my appreciation to Catalina, my girlfriend, for her emotional intelligence, understanding, and unwavering support throughout my academic journey. Her encouragement and motivation have been crucial, especially during times of uncertainty and disappointment.

I am grateful to my supervisors, Bárbara and Felipe, for their invaluable guidance, insight, and mentorship. Their ideas, reviews, and discussions have been instrumental in keeping me motivated and focused, especially during the early days of the pandemic. They are not only exceptional researchers and educators but also wonderful people. I also want to thank all the professors who have contributed to my Ph.D. in various capacities, including Claudio Gutierrez, whose wise conversations and exceptional qualities as a person have been invaluable, and Aidan Hogan, whose helpfulness and sympathy are greatly appreciated. I also thank Christian Bridevaux for giving me the opportunity to teach multiple times.

I would like to thank the Institute for Foundational Research on Data (IMFD) for its financial support, which allowed me to attend conferences and events. I also thank all those who are part of the IMFD. I am grateful to ANID for providing the scholarship that funded my Ph.D. degree.

I would like to extend my thanks to Sandra and Angélica for their support and assistance. I also want to thank Sergio Aguilera for his help with hardware, networks, and servers, which I could not have managed alone. Lastly, I am grateful to all the staff members of the DCC for providing an excellent environment to work and learn.

Finally, I would like to express my gratitude to my friends and colleagues. Although they may not fully understand what I am doing or why I am researching, they have provided me with emotional support and excellent moments that have kept me going.

Table of Content

List of Tables

List of Figures

Chapter 1

Introduction

The United Nations Department of Humanitarian Affairs describes a *disaster* as a severe disruption of the functioning of society, causing widespread human, material, or environmental losses, which exceed the ability of the affected society to cope, using only its resources [164]. Additionally, they define an *emergency or crisis* as a sudden and usually unforeseen event that calls for immediate measures to minimize its adverse consequences. Hence, the most significant difference between disaster and emergency is the damage and the effects on people. Crises or disasters are often classified into two primary groups[1]. *Natural* hazards - events not caused by human activities - are naturally occurring physical phenomena, for instance, earthquakes, pandemics, hurricanes, among others. And *human-induced (or man-made)* hazards, events that are caused by humans such as conflicts, industrial accidents, transport accidents, pollution, and so forth.

According to the 2020 World Disaster Report, between 2010 and 2019, there have been 2,850 disasters triggered by natural hazards, which have killed ten or more people or affected at least 100 people [165]. These catastrophes affected some 1.8 billion people, causing many of them injury, homelessness, or leaving them without a means of subsistence. The Economic and Social Commission for Asia and the Pacific (ESCAP) has determined that these regions account for 57% of global fatalities from disasters and 87% of the global affected population [60]. Given the percentage of disasters attributable to climate- and weather-related events in the last 30 years, people were affected by 97% by extreme weather and climate hazards.

A very recent example is the coronavirus COVID-19 pandemic, which has disrupted life globally since 2020. Considered a biological hazard, COVID-19 impacted almost every country and territory. Health systems worldwide have been overwhelmed by the number of cases, and even the wealthiest and most prepared countries have struggled. Millions of people have not had access to critical life-saving supplies in the most vulnerable countries, such as test kits, face masks, and respirators. At the end of 2021, there were over 253 million cases of coronavirus and 5.1 million deaths worldwide [200].

Over the last decades, human-induced crises emerged through wars, cyber-attacks, crime, civil disorders, protests, among others. Since 1990 up to now, there have been about 7,000

[1] https://www.ifrc.org/what-disaster

events that have affected over 3.5 million people worldwide [72]. They have harmed the general population in terms of injuries and deaths. Only in 2017, 2,934 deaths were the result of human-induced events, mainly by terrorism attacks, explosions, fires, and passenger ship sinkings [107]. Furthermore, human-induced crises have insurance losses worldwide that amounted to approximately 7.93 billion U.S. dollars in 2020 [201]. For instance, the terrorist attack on the World Trade Center in New York in September 2001 caused insured losses of almost 26 billion U.S. dollars, one-third of that resulting from Hurricane Katrina.

The community's vulnerability strongly influences the impact of a disaster on the hazard, which is the result of the whole range of economic, social, cultural, institutional, and political factors that shape people's lives [210]. Disaster risk is expressed as the interaction between the severity and frequency, the number of people exposed to it, and their vulnerability to damage. These factors can increase or decrease the disaster responses depending on the capacity of each country to deal with it [165; 211]. In this sense, the best disaster response is plan when contingencies can be anticipated, responses streamlined, and communication can be both rapid and accurate [49; 138].

Crisis communications are susceptible to public opinion, mainly because information and communication technology allow easy access to any information [92]. Failure of communication or misunderstanding may lead to loss of truth, a poor reputation, and inadequate response [191]. Recent crises (e.g., COVID-19 or The Black Lives Matter) have exhibited that -during these events- social phenomena such as polarization, misinformation, and fake news, generally emerge as a result of uncertainty [40; 55; 80; 94; 118]. In this sense, investigating the effects of disasters in a community through their communications is crucial because disasters can have long-lasting effects, re-defining cultural identities, highlighting collective needs, and changing political and institutional laws, among others [184]. Hence, mechanisms to improve the understanding of these events are crucial to explaining why they do or do not have a certain impact [13].

During emergencies, traditional media (e.g., TV and radio news) may suffer infrastructure issues, and real-time communications could be disrupted. For this reason, *Social Networking Services (SNS)* have become an important information source about real-world events. SNS has played a critical role over the past fifteen years, allowing its users in the affected locations to share real-time information, such as status updates, casualties, damages, and alerts, to the rest of the world [125; 203; 185]. Such relevant and vital information presents a valuable opportunity for authorities and public relief agencies to increase their *situational awareness*[2] about a crisis and better respond to it. For instance, the first news about the 2013 Westgate Mall Attack was a message posted on Twitter within a minute of the initial assault[3].

Twitter, a microblogging service where users post short messages (called *tweets*), is one of the most popular[4] SNS. Over 80% of Twitter users access the platform via their mobile devices, enhancing the immediacy and locality of information. Using specific keywords (called *hashtags*) to group messages into a particular topic and re-sharing messages (called *retweet*

[2] Situational awareness refers to inquiring users about perception, comprehension, and projection in situations where working activities have been interrupted [63].

[3] https://ihub.co.ke/blogs/16012/how-useful-is-a-tweet-a-review-of-the-first-tweets-from-the-westgate-mall-attack

[4] https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/

action), information rapidly propagates through the network, mainly when high-impact real-world events occur. Given the availability of data through API[5], Twitter is one of the most popular platforms used for researchers in order to analyze real-world events in online platforms [4].

Researchers have studied social media user behavior during these events to detect, summarize and classify messages to help authorities and the general public with situational awareness to provide fast and conscientious responses during crises. There is a large body of work related to the use of social media during emergencies, starting with Palen and Liu [172] who published one of the first studies on the relevance of collecting information in wikis about missing people related to the September 11, 2001 attacks on the Twin Towers in New York. However, there are undeniable challenges to effectively leveraging social media for crisis response and management. Most of these challenges are related to the unstructured and noisy nature of the data and its immense volume [104].

Works in literature have devoted considerable efforts to creating automatic approaches to extract, detect and characterize useful information published in online platforms during crises [103; 168; 179]. In general, these approaches require extensive collections of messages to train models that learn from different disaster domains, geographic locations, and languages, which makes the adaptation and analysis of a wide range of types of crises difficult. Additionally, most of the studies described in the literature are based on the analysis of messages using a predetermined set of keywords to filter crisis-related content and for a specific language [125; 101; 37; 143]. Overall, the limitations of keyword-based approaches in the context of crisis management highlight the need for more flexible and adaptive approaches that can handle a wide range of crisis types, languages, and geographic locations. In this direction, keyword-based and specific language approaches have certain shortcomings:

a) They require supervised methods to determine if the identified data actually is relevant in the context of a crisis or corresponds to a new real-time crisis event. Hence, if historical data does not exist to train models for a specific domain or language, they could have low performance and accuracy.

b) Most of these works explain phenomena just for English messages, independently whether an event did not occur in an English-speaking country. Therefore, it avoids replicating methodologies in other languages and countries where emergencies often occur (e.g., Chile, Mexico, or India).

c) Literature have focused on analyzing specific characteristics of one (or few) events instead of aggregated patterns that emerge from them by considering various hazard dimensions (e.g., temporal development and geographic spread).

d) They require specific domain knowledge of different crisis events to determine a set of keywords to filter and extract relevant information about an emerging crisis situation.

In this thesis, we propose and develop computational tools to understand general and particular patterns of social media communications, specifically in Twitter, during crises

[5] https://developer.twitter.com/en/docs/twitter-api

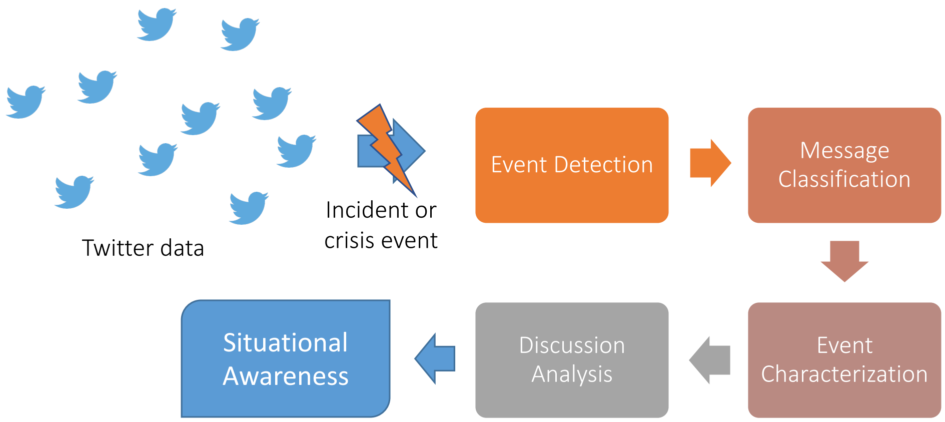


Figure 1.1: Crisis tasks addressed in this thesis to improve situational awareness during emergencies.

	Properties		**Level of Understanding**		**ML approach**		**Event Scope**	
Crisis Task	Domain-independent	Language-agnostic	Message	Event	Supervised	Unsupervised	Multiple	Case study
Detection	✓	✓		✓	✓		✓	
Classification		✓	✓		✓		✓	
Characterization	✓			✓		✓	✓	
Discussion Analysis	✓	✓	✓			✓		✓

Table 1.1: Summary of tasks, properties and scenarios developed in this thesis.

across several types of events and geographic locations. We exploit domain-independent and language-agnostic patterns for discovering and characterizing collective online activity related to emergency situations. We address this objective by studying several tasks that allow us to delve into different properties and scenarios to leverage the implicit and transversal characteristics that may exist across events. As Figure 1.1 shows, we cover the most common actions taken during crises to improve situational awareness, which involve the detection of a new unseen event using social media data, automatic filtering of relevant/irrelevant content, description of the main characteristics of the event, and the analysis of discussions generated among users.

Table 1.1 summarizes in detail the tasks and properties considered for this work. We focus on developing several computational tools that help to detect, classify, characterize and analyze useful information, which is further highlighted by the properties of being transversal across the type of event (domain-independent property) and adaptable for several languages (language-agnostic property). Additionally, we study these characteristics at different levels of understanding (by message and event), considering several machine learning approaches (unsupervised and supervised) and several crises that differ in terms of time, duration, and other hazard dimensions.

Given the amount of work in literature that focuses on extracting useful information from various collections of events (and messages), we first review, collect and consolidate crisis-related datasets from several sources that differ in geographic locations, languages, and crisis domains. By enriching and adding relevant metadata information about crises, such as hazard dimensions, affected areas, languages, and classification labels, we unified and created a consolidated dataset that allows the development of multi-dimensional analyses related to online collective reactions on Twitter.

We focus on detecting new unseen and unexpected crises on Twitter. Unlike previous studies based on a crisis-specific keyword-based approach, we leverage the idea that 1) emergencies occur in physical locations, which users often share on online platforms. 2) locations do not change over time, contributing to not specifying what type of event we need to identify. And 3) locations can only be changed among languages; however, they can be the predefined in case of analyzing multiple languages. In this direction, we introduce a novelty method that identifies these events based on location mentions on Twitter, allowing detect events in a domain-independent and language-agnostic manner. Our results show that only by tracking location signals (in one or more languages) associated with a specific country we detected up to 80% of the crises.

We further study cross-lingual and cross-domain approaches to address the problem of identifying crisis-related messages. We aim to leverage data from other languages and domains to deal with the *cold-start* problem, in which historical data do not exist to train model for filtering irrelevant content for new unseen crises. In this sense, we conduct an extensive experimental evaluation for multiple classification scenarios in several crisis domains and languages. Considering several language-agnostics representations, we note that by introducing data from English to low-resource languages (such as Spanish and Italian) and other domains, classification performance can be improved in several scenarios.

Furthermore, we conduct a quantitative and qualitative analysis to characterize crisis communications from a textual and linguistic point of view across a wide range of events. We aim to discover common patterns among crises that differ in time, location, and hazard dimensions, allowing us to study domain-independent characteristics of events published online. We address this analysis by considering several crisis dimensions, such as hazard categories (e.g., human-induced and natural disasters), fine-grained crisis subcategories (e.g., intentional and geophysical crises), and hazard types (e.g., demonstrations and earthquakes). We first discover that there are clear patterns in how people react to different extreme situations, depending on, for example, whether the event was triggered by natural causes or human action. Our results show that using only a small set of textual features, we can differentiate among types of events with 75% accuracy.

Finally, we also study long-term crises by analyzing relevant and valuable information that may help authorities and emergency agencies to comprehend - with weak human intervention - controversial online discussions that emerge during these events. In this direction, we address the study of polarization during extreme events, a trending topic that has emerged mainly with the increase of online interchanges and reaffirmed with recent uncertain high-impact events such COVID-19 [85; 94]. We propose an unsupervised approach composed of a series of steps from identifying communities and characterizing them through their discourses. Our proposed method was validated around the conversations of the 2019-2020 Chilean protests[6], a highly polarized event that disrupted the entire (Chilean) society and known as one of the most impactful crisis of the last 30 years. Although our approach was evaluated on the 2019-2020 Chilean protests, it can be applicable to a wide range of crises (and non-related emergency events) because it does not require historical data to train a supervised model or

[6] Chileans protest for a wide variety of problems ranging from inequality to the high cost of healthcare.

domain-specific knowledge about the language or event that we are analyzing.

1.1 Hypothesis and Objectives

Our hypothesis asserts that there are patterns in the self-organized activity of the Web and social media users that emerge when a crisis situation starts to unfold in the physical world. Some of these patterns arise independently of the particular type or domain of the crisis event, as well as independent of the location, language and culture of the users that participate. Considering the above statement, we propose the following research questions:

1. Can we characterize collective patterns during crisis situations independently of their language and domain?
2. Are non-textual and low-level lexical features sufficient at reducing the number of non-related emergency situations detected as crises in social media?
3. Can we leverage data from other crisis dimensions as well as languages (transfer learning tasks) to identify crisis-related messages?
4. If transfer learning shows to be effective, are there instances (languages and crisis dimensions) where the method works better?
5. Are there differences among types of emergency situations based on their hazard dimensions (e.g., categories, subcategories and geographic spread) related to social media messages posted during these events?
6. How can we automatically identify, characterize and measure relevant information posted during long-term crises events?

Our main objective is to perform a large-scale transversal study of crisis events across various types of events and geographic locations to understand general patterns of crisis communications and useful patterns. This should help us better understand the social media behavior during crises in affected locations around the world, independent of their language, domain and type of event. Accordingly, the specific objectives are described as follows:

1. Consolidate a large-scale dataset of Twitter messages from diverse crisis events enriched with relevant metadata.
2. Study domain-independent and language-agnostic representations of crisis event messages to understand communications patterns through crisis dimensions, locations and languages.
3. Develop computational tools that help to characterize, classify, detect and extract useful information shared during crises.

1.2 Methodology

In this section we present the steps that we must follow for this work. To develop each of the objectives mentioned, we present a methodology that consists of the following milestones:

- *Literature review.* To acquire a deep understanding of crises from multiple perspectives and fields, we research different investigations and reports from sociology, risk management, and social sciences, among others. Furthermore, we review the state-of-the-art in *crisis informatics* with a focus on crisis detection and characterization. In this sense, we inspect methods, resources and data that are used to address these tasks.

- *Consolidate a large-scale dataset of Twitter messages from diverse crisis events enriched with relevant metadata.* We identify, collect, and unify user-generated content datasets, with a focus on those for which data were manually labeled. Additionally, we enrich the unified dataset adding relevant information such as crisis dimensions, affected locations, merged labels, among others.

- *Study representations of crisis event messages to understand communications patterns through crisis dimensions and locations.* We identify characteristics in the social media content that allow us to differentiate among diverse types of crisis events using resources and methodologies in a particular language (e.g., English). We perform a quantitative analysis to determine differences and similarities across diverse crisis dimensions. We also develop a qualitative analysis of human-induced and natural events, contrasting our results to psychology studies on disaster victims and to other online user-generated content about disasters.

- *Study language-agnostic representations of crisis event messages to understand communications patterns through crisis dimension and languages.* We conduct several studies to address this objective. We first present a method to detect crisis events by tracking and analyzing changes on locations' mentions'. This method, by not relying on the type of event, facilitates the analysis in a domain-independent manner and is adaptable for multiples languages. We further perform an experimental analysis to identify languages and domains for which transfer learning techniques have better results in crisis event messages. We perform a quantitative analysis to determine differences and similarities across diverse crisis dimensions and languages. Finally, we research the problem of extracting relevant information from long-term crises based on online discussions among users during these events. We propose a framework that characterizes controversial discussions during these events, by requiring minimum human intervention and allowing adaptability for different events and languages.

- *Develop computational tools that help to characterize, classify, detect and extract useful information shared during crises.* For each of our proposed tasks that cover several scenarios during crises, such as characterization, information extraction, classification and detection, we develop computational tools that are available in their corresponding repositories.

1.3 Contribution of this Work

This research may help government officials, public disaster agencies, and news media, among others, understand how social media users react to crises. In addition, this should facilitate better understanding of the social media behavior during crises in affected locations around the world, independent of their language, domain and type of event. The concrete contribution of this work can be summarized in the following points:

- Datasets that include crisis-related messages shared from different locations, languages and events are available in our repositories.
- The proposal, development and evaluation of a novelty method to detect crisis events in social media that is independent of the language and domain.
- A cross-lingual and cross-domain experimental design that evaluates multiple scenarios and representations using social media data.
- A transversal large-scale study of crisis-related messages that reveals similarities and differences through crisis dimensions.
- The first study of the group polarization during the 2019 Chilean Social Unrest Movement using Twitter data.
- A low human intervention method that identifies and extracts relevant information from controversial conversations that emerge during crises, which can be adaptable for other events and languages.
- Computational methods focusing on detecting, identiying and extracting relevant information in online platforms with a domain-independent and language-agnostic approach.

In terms of publications, we provide a list of all accepted papers related to this thesis since starting this Ph.D. program:

1. Sarmiento, H., Poblete, B., & Campos, J. (2018, May). Domain-Independent detection of emergency situations based on social activity related to geolocations. In Proceedings of the 10th ACM Conference on Web Science (pp. 245-254).
2. Sarmiento, H. (2019, July). A Domain-Independent and Multilingual Approach for Crisis Event Detection and Understanding. In Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 1457-1457).
3. Sarmiento, H., & Poblete, B. (2021, March). Crisis communication: a comparative study of communication patterns across crisis events in social media. In Proceedings of the 36th Annual ACM Symposium on Applied Computing (pp. 1711-1720).
4. Sánchez Macías, C. M. (2021). Transfer learning for the multilingual and multi-domain classification of messages relating to crises. MSc. in Computer Science Thesis, Departament of Computer Science, University of Chile (this work was co-supervised by Hernán Sarmiento)
5. Sarmiento, H., Bravo-Marquez, F., Graells-Garrido, E. and Poblete. B (2022, June). Identifying and Characterizing new Expressions of Community Framing during Polarization. In Proceedings of the 16th The International AAAI Conference on Web and Social Media (ICWSM), Atlanta, Georgia, USA.

1.4 Document Structure

This thesis is organized as follows.

1. In Chapter 2, we describe several definitions and methods from Data Mining and Machine Learning fields, such as Support Vector Machine, Decision Tree, DBScan, among others. We also provide basic definitions of probability distributions and metrics for evaluating the effectiveness of machine learning models. In addition, we briefly describe several basic notions about network analysis that we consider for this work.

2. In Chapter 3, we introduce an overview of relevant literature related to this work. We explain the most relevant work in crisis informatics, event detection in social media, and message characterization and classification. Furthermore, we detail the state-of-the-art in group polarization with a focus on research using online social media data.

3. In Chapter 4, we present a domain-independent and language-agnostic approach for detecting crisis events in social media. This method was applied in a series of different types of events and languages to evaluate the algorithm's effectiveness.

4. In Chapter 5, we address the problem of classifying related and non-related messages considering a cross-domain and cross-lingual approach. We perform an analysis including several scenarios and data representations.

5. In Chapter 6, we conduct a quantitative and qualitative analysis of multiples crisis events by considering diverse hazard dimensions. We focus on studying differences and similarities in social media messages published during disasters from a linguistic point of view.

6. In Chapter 7, we focus on extracting and analyzing valuable information during long-term crises. In this direction, we study polarized topics that emerged in online conversations about the 2019-2020 Chilean unrest movement. To identify communities and understand concepts discussed by the groups, we propose an unsupervised method that is composed of a series of steps from detecting communities to discovering differences and similarities in the use of specific topics in social media messages.

7. We finalize Chapter 8 with a discussion, conclusions, and future work.

Chapter 2

Preliminaries

This chapter briefly introduces the main concepts related to disasters that help practitioners understand several dimensions of these events. Furthermore, we define the principal techniques, metrics, and algorithms that are used in the *crisis informatics* field to study extreme events incorporating Web and social media data.

2.1 Crisis Management

Crises, emergencies, and disasters are often used interchangeably. Although they describe dangerous phenomena or *hazard* events that could affect people, several differences exist among them. Crises are situations faced by an individual, group, or organization that they cannot cope with by the use of normal routine procedures and in which the sudden change creates stress [153]. Emergencies are imminent or actual events that threaten people, property or the environment and which require a co-ordinated and rapid response [195]. Furthermore, disasters are situations that overwhelm local capacity to withstand, cope and recover; necessitating external assistance and involving various stakeholders [91; 154]. Operationally for emergency agencies and government, disasters exceed the capacity of normal, workday systems to cope with them effectively [23].

Regarding how to deal with a crisis, the literature has proposed dividing these events into several stages, which describe the life cycle of an emergency. These stages comprise the following milestones [35]: Mitigation is the risk reduction according to the possible damage an event could generate. Preparedness, where the government and emergency offices educate the population on how to respond to an event. Response, which describes the immediate actions that should be taken to minimize the damage of an event. Recovery is defined as restoring the functioning of essential services and the routine life of the population.

Overall, studies in crisis management have determined that disasters differ from crises and emergencies in disrupting a system as a whole by requiring external assistance and where information is highly unpredictable [6]. In addition, crises and emergencies have several similarities where the most predominant in literature are that they are confined to a small

Category	Subcategory	Examples
Natural	Meteorological	tornado, hurricane
	Geophysical	earthquakes, volcano eruptions
	Hydrological	floods, landslides
	Climatological	wildfires, heat/cold waves
	Biological	epidemics, infestations
Human-induced	Accidental	building collapses, crashes
	Intentional	shootings, bombings

Table 2.1: Examples of hazard categories and subcategories. Table is based on the work of Olteanu et al. 2015 [168].

population and can turn into a disaster if it is neglected or mismanaged [6].

Hazards can be defined by different dimensions, which allow examining the origin of the event and temporal and geographical aspects [2; 70; 168].

- Hazard category[1]: define who triggers the events, which can be by *natural* reasons or *human-induced (or man-made)* actions. Table 2.1 shows examples of these events.
- Hazard subcategory: this dimension adds an extra level of granularity to the natural and human-induced events. Natural hazards can be divided into *meteorological*, *geophysical*, *hydrological*, *climatological* and *biological*. Human-induced events are divided into *accidental* and *intentional* hazards. Table 2.1 shows examples of these events.
- Temporal development: this dimension labels hazards as *instantaneous* if they do not allow pre-disaster mobilization of workers or pre-impact evacuation of those in danger. On the other hand, they are *progressive* hazards if not preceded by a warning period. Table 2.2 shows examples of these events.
- Geographic spread: this feature defines the area that a hazard affects. First, *focalized* hazards involve and mobilize response in a small area. Second, *diffused* hazards impact a large geographic area and/or mobilize national or international response. Table 2.2 shows examples of these events.

2.2 Data Analysis Tools

This thesis employs techniques from various data analysis fields such as Data Mining, Machine Learning and Statistical Inferences. To analyze emergency situations in social media platforms, techniques from all these areas are required, as for instance, classification, clustering and hypothesis testing.

[1] In some cases, the literature considers that human activities, such as overpopulation, pollution, and deforestation, influence natural events [66].

		Natural	Human-induced
Instantaneous	Focalized	landslides	building collapses
	Diffused	earthquakes	large-scale industrial accidents
Progressive	Focalized	infestations	demonstrations, riots
	Diffused	floods, pandemics	wars

Table 2.2: Examples of hazards divided into their temporal development, geographic spread and category.

Data Mining	**Machine Learning**
Extracting useful information from large amount of data	Introduce algorithm from data as well as from past experience
Large databases with semi or unstructured data	Existing data as well as algorithms
Used to understand the data	Teaches the computer to learn and understand from the data
Data mining is more like research using methods of machine learning	Self-learned and training system to do the intelligent task

Table 2.3: Differences between Data Mining and Machine Learning. Table is based on the work of Zhou [226] and Jordan and Mitchell [109].

In general, the terms Data Mining and Machine Learning are used interchangeably. However, there are several differences, as explained in Table 2.3. Overall, Data Mining aims to extract knowledge from an extensive large amount of data (i.e., semi-structured or unstructured). On the other hand, Machine learning aims to design and develop algorithms that allow computers to learn patterns from the data and then create predictive models.

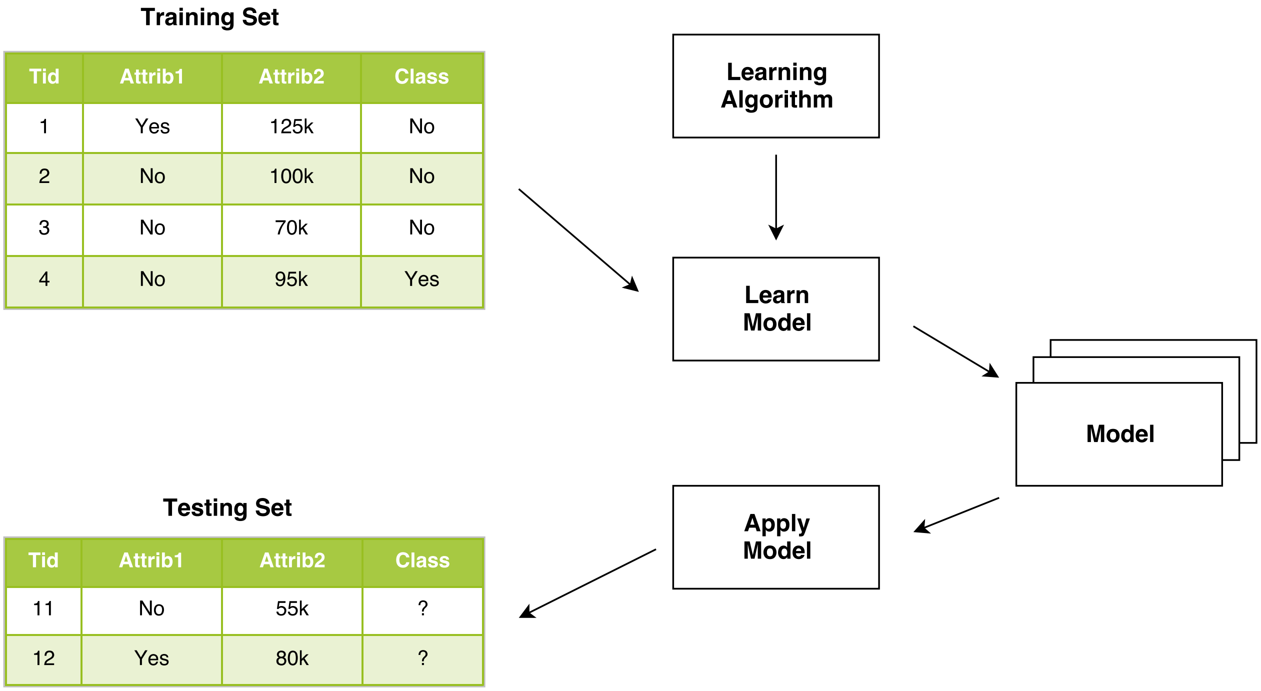
In the following sections, we detailed the most important methods used in this thesis.

2.2.1 Classification

Classification is the task of learning a target function f that maps each attribute set x to one of the predefined class labels y. The target function is also known as *classification model.* The input in this task is a collection of records, where each record is also known as an instance or example. Additionally, one instance is characterized by a tuple (x, y), where x is the attribute set and y is the class label (also known as category or target attribute)[207].

The general approach to solving a classification problem is to build classification models from an input dataset. In this way, building models with generalization capability is the main goal to predict the class labels of previously unknown records.

Figure 2.1 features a general approach for solving classification problems. First, a training set is used to build a classification model. This set consists of records whose class labels are known. Hence, a testing set is created to evaluate classification model over unknown class labels.



Training Set

Tid	Attrib1	Attrib2	Class
1	Yes	125k	No
2	No	100k	No
3	No	70k	No
4	No	95k	Yes

Testing Set

Tid	Attrib1	Attrib2	Class
11	No	55k	?
12	Yes	80k	?

Figure 2.1: General approach for building a classification model based on the work of Tan et al. [207].

.

For the goal of this work, and given the wide gamut of classification algorithms [207; 152], we select and describe the following algorithms: Support Vector Machine, Decision Tree, Random Forest, and Neural Networks. We choose these algorithms because they have been extensively used in several tasks, such as crisis-related messages [45; 68], hate speech detection [18; 169], sentiment analysis [113; 71] among others. In particular, we considered the Support Vector Machine, Decision Tree, and Random Forest as classification algorithms because they need fewer data during training, unlike deep learning algorithms. Furthermore, the former is generally used given the interpretability and their understanding of features introspection, which allows us to describe in more detail the different dimensions analyzed in this work. In contrast, we describe Neural Networks architectures because, in Section 2.4.1, We introduce the distributed word representation, a technique that considers this type of architecture to compute a word's context in a collection of documents.

2.2.2 Support Vector Machine Algorithm

The Support Vector Machine (SVM) is based on the concept of hyperplanes $(w^t \cdot x + b)$ that define decision boundaries for the binary classification problem $y_i \in -1, 1$ consisting of N training examples represented by x. The optimal hyperplane is the one that maximizes the margin between positives and negative observations in the training dataset [50]. The SVM algorithm can be formalized as the following optimization problem:

$$min_{w,b,\xi_i} \quad \frac{1}{2}||w||^2 + C\sum_{i=1}^{\mathbb{N}} \xi_i \tag{2.1}$$

$$\text{subject to } y_i(w^t x_i + b) \geq 1 - \xi_i, \xi_i \geq 0$$

Where C is a user-specified parameter that represents the penalty of misclassifying the training instance. Hence, this parameter is referred to as the soft margin regularization and controls the sensitivity to possible outliers. The SVM classifier also has a user-specificied parameter for controlling unbalanced data with respect to the number of instances for each class called class weights. There are other parameters for specific configurations of the kernels as the gamma, coefficient and degree.

The SVM formulations described above construct a linear decision boundary to separate the training examples into their respective classes. It is also possible to make SVMs find non-linear decision boundaries. A function $\phi(x)$ maps the feature space x into a high-dimensional space is used. This high-dimensional space is called Hilbert space, where the dot product $\phi(x) \cdot \phi(x')$ is known as the kernel function $K(x, x')$. So the hyperplane is calculated in the high-dimensional space $(w^t \cdot \phi(x) + b)$. Finally, we replace every dot product by a kernel function as shown in the following expression:

$$max_\alpha \quad \sum_{i=1}^{\mathbb{N}} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{\mathbb{N}} \alpha_i \alpha_j y_i y_j \cdot K(x_i, x_j) \tag{2.2}$$

$$\text{subject to } \alpha_i \geq 0, \forall i \in \{1, .., N\}, \textstyle\sum_{i=1}^{\mathbb{N}} \alpha_i y_i = 0$$

Where the parameter $\alpha_i, \forall i\{1, ..., N\}$ corresponds to the *Lagrange multipliers* of the constrained optimization problem.

Many options for kernel function exist as following:

1. Linear function: $K(x_i, x_s) = x_i^T \cdot x_s$
2. Polynomial function: $K(x_i, x_s) = (x_i \cdot x_s + 1)^d$, where $d\forall\mathbb{N}$ represents the polynomial degree.
3. Radial basis function (RBF): $K(x_i, x_s) = exp(-\frac{||x_i - x_s||^2}{2\rho^2})$, where $\rho > 0$ represents the width of the kernel.

2.2.3 Decision Tree

A decision tree is simple, yet widely used classification technique. The structure of the decision tree is like a flowchart in which each internal node represents a test on an attribute where each branch represents the outcome of the test, and each leaf node represents a class label.

Greedy strategies are used to build a decision tree by making a series of locally optimum decisions on which attribute to use for partitioning the data. One such algorithm is *Hunt's algorithm*. In *Hunt's algorithm*, *"a decision tree is grown in a recursive fashion by partitioning the training records D_t that are associated with node t and the class labels $y = \{y_1, y_2, ..., y_c\}$"* [206]. The recursive definition of *Hunt's algorithm* is to select a partition of the records using an attribute test condition into smaller subsets when D_t contains records that belong to more than one class. A child node is created for each outcome of the test condition and the records in D_t are distributed to the children based on the outcomes. The algorithm is then recursively applied to each child node. The recursion termination is applied when all the records in D_t belong to the same class y_t. Then t is a leaf node labeled as y_t.

2.2.4 Random Forest

Random Forest is a class of ensemble methods that combines the predictions made by multiple decision trees. Each tree is created based on the values of an independent set of selected random vectors [206]. Bootstrap aggregation (also known as bagging) is used in the model-building process to choose N samples, with replacement, from the original training set. About one-third of the cases are left out of the bootstrap sample and not used in the construction of the tree. In this step, the Out-of-Bag (OOB) data is used to get a running unbiased estimate of the classification error as trees are added to the forest. It is also used to get estimates of variable importance.

To perform prediction using the trained random forest, the algorithm uses the following steps:

1. Take the test features and use the rules of each randomly created decision tree to predict the outcome and store the predicted target.
2. Calculate the votes for each predicted target.
3. Consider the highest voted predicted target as the final prediction. This concept of voting is known as majority voting.

2.2.5 Neural Networks

A neural network is a series of algorithms that attempts to recognize the underlying relationships in a dataset. The neural network structure works similarly to the human brain's neural network. The basic object is called *neurons*, which are represented as a mathematical function that collects and classifies information according to a specific architecture. Figure 2.2 depicts a neuron connected with n other neurons that receives n inputs $(x_1, x_2, ...x_n)$. This structure is called a *perceptron*. The perceptron consists of weights, summation processor,

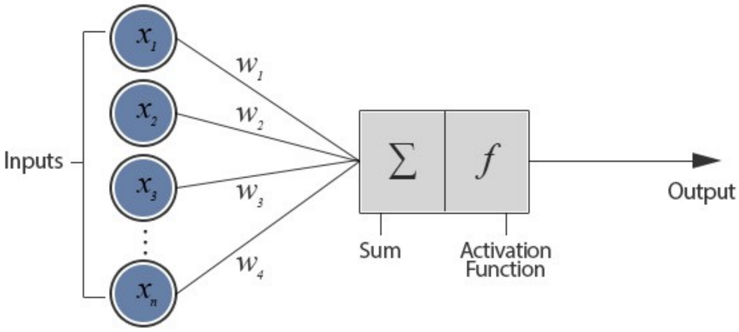


Figure 2.2: A general overview of neural networks structure.

an activation function and a threshold processor (known as bias). Hence, the mathematical notation can be represented as follows:

$$f(b + \sum_{i=1}^{n} x_i w_i) \tag{2.3}$$

where f represents the activation function, b the bias, w the weights and x the input to neuron. For more details about these and other themes derived from neural networks, please refer to the books by Goodfellow et al. [79] and Zhang et al. [225].

2.3 Clustering

A cluster is a set of similar objects based only on information found in the data that describe the objects and their relationships. The main goal of cluster analysis is for the object within a group to be similar (or related) to another and different from (or unrelated to) the objects in other groups (Figure 2.3).

One cluster can be differentiated from another using a distance measure between their attributes. Some distance measures are explained in the following:

1. Manhattan distance:
$$d_{man}(x, y) = \sum_{i=1}^{n} |(x_i - y_i)| \tag{2.4}$$

2. Euclidean distance:
$$d_{euc}(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{2.5}$$

3. Minkowski distance:
$$d_{mink}(x, y) = (\sum_{i=1}^{n} |(x_i - y_i)|^p)^{1/p} \tag{2.6}$$

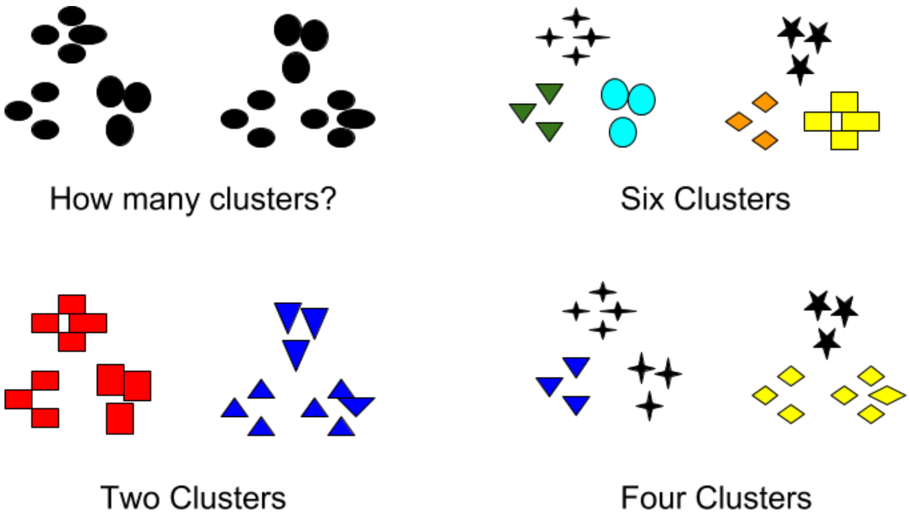


Figure 2.3: Three different ways of clustering the same set of points based on the work of Tan et al. [206].

.

4. Pearson correlation distance:

$$d_{cor}(x,y) = 1 - \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{2.7}$$

The unsupervised algorithms to be considered in this thesis are explained in the following sections. For additional information about these algorithms please refer to [206; 152].

2.3.1 Partitional Models

The main idea in this class of clustering algorithm is to create K clusters of the data, where the number K is a user-specified parameter. Each object in the data is assigned to the nearest cluster center, such that the squared distances from the clusters are minimized. Furthermore, the algorithms prefer clusters of approximately similar size, as they will always assign an object to the nearest centroid.

K-means

K-means is a *hill-climbing* algorithm, which guarantees convergence to a local optimum, but not necessarily a global optimum [133]. The main idea in this algorithm is to use the means to represent the clusters and use them as a guide to assign object to clusters.

Given an initial set of K means $m_1, ..., m_k$, the algorithm proceeds by alternating between two steps [140]:

1. Assignment step: assign each object to the cluster whose mean has the least squared Euclidean distance such that:

$$S_i^{(t)} = \big\{x_p : \big\|x_p - m_i^{(t)}\big\|^2 \le \big\|x_p - m_j^{(t)}\big\|^2 \; \forall j, 1 \le j \le k\big\} \tag{2.8}$$

   where each x_p is assigned to exactly one $S^{(t)}$.

2. Update step: compute the new means to be the centroids of the observations in the new cluster. These centroids are calculated as following:

$$m_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \in S_i^{(t)}} x_j \tag{2.9}$$

The algorithm converges when the assignments no longer change, so the convergence is satisfied. The most used criterion is the minimization of the squared error E of all the objects in the data:

$$E = \sum_{i=1}^{K} \sum_{o \in C_i} |o - \mu_i|^2 \tag{2.10}$$

where o is an object in the data that belongs to cluster C_i, μ_i is the mean of the cluster C_i and K is the number of clusters.

K-medoids

In contrast to the K-means algorithm, K-medoids chooses data points as centers known as medoids. A medoid can be defined as the object of a cluster whose average dissimilarity to all of the objects in the cluster is minimal [114]. The general procedure for the algorithm is as follows:

1. Randomly choose k objects into data as the initial medoids.
2. Each one of the remaining objects is assigned to the cluster that has the closest medoid.
3. Randomly select a nonmedoid object in the current cluster, which will be referred to as $O_{nonmedoid}$.
4. Calculate the cost of replacing the medoid with $O_{nonmedoid}$. The cost is the difference in the square error if the current medoid is replaced by $O_{nonmedoid}$.

$$E = \sum_{i=1}^{K} \sum_{o \in C_i} |o - O_{medoid(i)}|^2 \tag{2.11}$$

   If E is negative, then make $O_{nonmedoid}$ the medoid of the cluster.

Each step of the algorithm is repeated until there is no change.

2.3.2 Density Models

In this class of clustering algorithm, the main idea is to keep growing clusters as long as their density is above a certain threshold. Clusters are defined as areas and the objects in these sparse areas are usually considered to be noise and border points. In contrast to partitional clustering where algorithms detect only a cluster of a convex shape, density models detect clusters of arbitrary shape.

DBSCAN

In the DBSCAN algorithm, given a set of points in some space, it groups together points that are closely packed together. The main idea is to create clusters that have a high enough density - high enough being specified by the user. Unlike the K-means and K-medoids, the number of clusters are not specified in DBSCAN. In this algorithm, specified-user parameter are: (1) ε (known as *eps*) represents the maximum radius of the neighborhood from the core point p; (2) $minPts$, the number of points reached by the core point p; (3) distance function can be chosen by the user, and has a major impact on the results [206; 152].

The DBSCAN algorithm can be abstracted into the following steps:

1. Find the ε neighbors of every point, and identify the core points with more than $minPts$ neighbors.
2. Find the connected components of core points on the neighbor graph, ignoring all non-core points.
3. Assign each non-core point to a nearby cluster if the cluster is an ε neighbor, otherwise assign it to noise.

2.3.3 Evaluation Methods and Metrics

The following sections present some concepts, tools and techniques used in our supervised and unsupervised experiments.

Confusion Matrix

In classification tasks, the evaluation of the performance is based on the counts of test record correctly and incorrectly predicted by the model. The predicted outputs are compared with their corresponding real values from the testing dataset. Using this approach for a binary classification problem, four possible outputs can be obtained as is shown in Table 2.4, known also as Confusion Matrix.

The first outcome, *True Positive (TP)*, represents the object O, belongs to class C and is classified as such. Secondly, *True Negative (TN)* represents the object O, does not belong to class C and it is not classified as a member of class C. Unlike the TP and TN, which describe correct classification of the object O for class C, *False Positive (FP)* and *False Negative (FN)* represent objects misclassified. On the one hand, FP describes that although object O does not belong to class C, it is classified as member of class C. On the other hand, FN represents that object O belongs to class C, and is not classified as a member of class C.

Table 2.4: Classification Confusion Matrix.

	Actual Value: Positive	**Actual Value: Negative**
Prediction Outcome: Positive	True Positive (TP)	False Negative (FN)
Prediction Outcome: Negative	False Positive (FP)	True Negative (TN)

According the different outputs explained above, the following measures can be computed:

- Precision (P): the proportion of correctly classified positive observations over all the observations classified as positive.

$$Precision = \frac{TP}{TP + FP} \tag{2.12}$$

- Recall (R): the proportion of positive classified observations over all the actual positive observations.

$$Recall = \frac{TP}{TP + FN} \tag{2.13}$$

- F_1-score: the harmonic mean of precision and recall.

$$F_1\text{-}score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{2.14}$$

- False Positive Rate (FPR): the proportion of the false positives over all the negative observations.

$$\text{False Positive Rate} = \frac{FP}{FP + TN} \tag{2.15}$$

Clustering Evaluation

The Internal Criteria, also known as *unsupervised evaluation* or *internal indexes*, is the clustering evaluation that compares clustering results with the result itself and only using information present in the data set. Unsupervised measures are often divided into two classes: measures of cluster cohesion (compactness, tightness) and measures of cluster separation (isolation). For additional information about the internal criteria, please refer to [206; 95].

The External Criteria, also known as Supervised Evaluation or External Indexes, is the clustering evaluation that uses information not present in the dataset (e.g., class labels). Here, the clustering result is compared with the ground truth, and if the result is somehow similar to the reference, this final output is considered as a “good” clustering. Some measures used in this thesis are as follows:

- Purity: the purity measure focuses on the representative class, i.e., the class with the majority object, within each cluster. Purity can be computed formally with the following expression:

$$Purity(\Omega, C) = \frac{1}{N} \sum_k max_j |\omega_k \cap c_j| \tag{2.16}$$

  where $\Omega = \{\omega_1, \omega_2, ..., \omega_k\}$ is the set of clusters and $C = \{c_1, c_2, ..., c_j\}$ is the set of classes.

- Entropy: the entropy measure is the expected amount of uncertainty in a cluster. It can also be represented as a measure of disorder in the cluster. The entropy measure can be computed as follows:

$$Entropy(\Omega) = -\sum_k \frac{P(\omega_k)}{N} log \frac{P(\omega_k)}{N} \tag{2.17}$$

  where $P(\omega_k)$ is the probability of an element being in cluster ω_k and N is the number of points in the dataset.

- Normalized Mutual Information (NMI): is a measure that allows trading off quality of the clustering against the number of clusters. The NMI measure can be computed as follows:

$$NMI(\Omega, C) = \frac{I(\Omega; C)}{[H(\Omega) + H(C)]/2} \tag{2.18}$$

  and I is mutual information computed as follows:

$$I(\Omega; C) = \sum_k \sum_j P(\omega_k \cap c_j) log \frac{P(\omega_k \cap c_j)}{P(\omega_k) \cdot P(c_j)} \tag{2.19}$$

  where $H(\Omega)$ and $H(C)$ are the entropies of the Ω and C respectively, and $P(\omega_k \cap c_j)$ is the probability of a element being in the intersection of ω_k and c_j.

- Variation of Information (VI): it is highly related to the mutual information, which measures the amount of information that is lost or gained in changing from the class set to the cluster set. By a random variable view similar to the previous case, we can compute the variation of information as follows:

$$VI(C, C') = 2H(C, C') - H(C) - H(C') \tag{2.20}$$

  where $H(C)$ is the entropy associated with clustering C and $H(C')$ is the entropy associated with clustering C'.

2.4 Natural Language Processing (NLP)

Natural Language Processing is the set of methods and techniques for the purpose of learning and making human language accessible for computers [61; 99]. Nowadays, applications of NLP have become embedded in daily life. For instance, machine translation, automatic text processing and summarization, speech recognition, multilingual and cross language information retrieval, and so on.

Next, we introduce some relevant topics utilized in this thesis such as word representations and applications of sequence labeling. For more details about these topics and other themes used in NLP, please refer to the material presented by Eisenstein [61] and Hirschberg and Manning [99].

2.4.1 Word Representation

Words are usually the smallest units of speech or writing in human languages, therefore, one of the fundamental questions in NLP is how to represent each document or instance (composed of words) with the goal that models can understand, categorize, or generate text in NLP tasks.

One-hot encoded vectors

One-hot encoding is a feature extraction technique that uses a column vector of word counts. Hence, the obtained output is a vector space that represents each document in the corpus. For instance, $x = [0, 1, 1, 0, 2, 13, 0...]^T$, where x_j is the count of word j. The length of x is $|\mathcal{V}|$, where $\mathcal{V}$ is the set of possible words in the vocabulary. One-hot encoding is considered one of the easiest implementations for modeling text, but it loses the inner meaning of the word in a sentence, thus losing the context of the sentence.

The one-hot encoded vectors model has several variants, each of which extends or modifies the base, for instance, frequency vectors (count vectors) and Term Frequency/Inverse Document Frequency.

Distributed Word Representation

The distributional hypothesis states that linguistic objects with similar distributions have similar meanings [69]. Based on this hypothesis, Brown et al. [34] proposed to group words into hierarchical clusters where words in the same cluster have a similar meaning. As the authors mentioned, the created clusters can represent the similarity between words, but it is not available to words in the same group.

Unlike *distributional representations* that are computed from context statistics and represented by symbolic structures, *distributed representations* are often estimated from distributional statistics, as in latent semantic analysis and Word2Vec, and represented by numerical vectors. The idea behind distributed representations is to embed each word into a continuous real-valued vector to address the above problem. This representation is often called a *dense* representation where the term "dense" means that a concept is represented by more than one dimension of the vector, and each dimension of the vector is used to represent multiple concepts. Figure 2.4 shows a two-dimensional projection of the 1000-dimensional vectors of countries and their capital cities. As noted, this type of representation has the ability of automatically organizing concepts and implicitly learning the relationships between them, without providing any supervised information about these concepts.

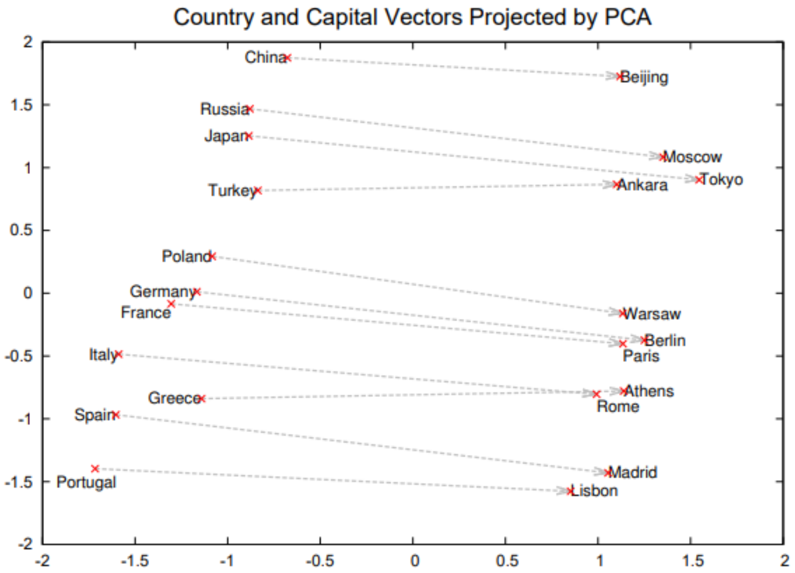


Figure 2.4: A two-dimensional projection of 1000-dimensional vectors of countries and their capital cities based on the work of Mikolov et al. [151].

.

Today, the dominant word representations are k-dimensional vectors of real numbers, known as *word embeddings*. One of the most popular software packages is *Word2Vec* [151]. In a general view, the Word2Vec can be defined as the following. Let the vector u_i represent the k-dimensional embedding for word i, and let v_j represent the K-dimensional embedding for context j. The inner product $u_i \cdot v_j$ represents the compatibility between word i and context j. By including this inner product into an approximation to the log-likelihood of a corpus, the algorithm estimates both parameters by backpropagation.

In particular, the Word2Vec algorithm includes two types of approximations. First, the *Continuous Bag-Of-Words (CBOW)* is based on the assumption that the meaning of a word can be learned from its context. This means that CBOW optimizes the embeddings to predict a target word given its context words. And second, the *Skip-gram* model learns the embeddings that can predict the context words given a target word. Figure 2.5 shows an overview of the CBOW and Skip-gram architectures based on the above explanation.

2.4.2 Sequence Labeling

The *sequence labeling* comprises a family of NLP tasks aimed to assign discrete labels to words or discrete elements in a sequence. The labels assigned usually depends on the types of the specific tasks. For instance, classical tasks include part-of-speech (POS) tagging, named entity recognition (NER), text chunking, and so on. The sequence labeling plays an

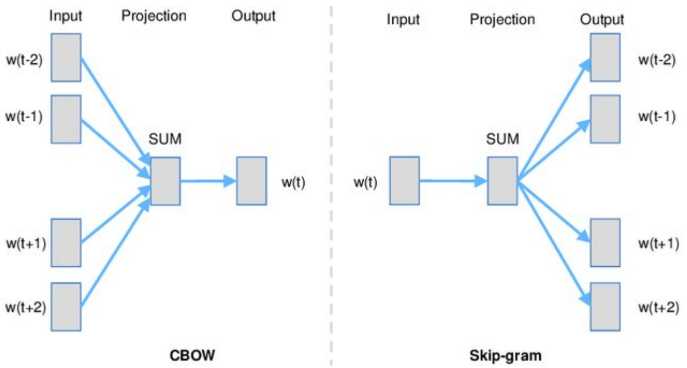


Figure 2.5: Architecture examples of Continuos Bag-of-words (CBOW) and Skip-gram training models from work of Landthaler et al. [130].

.

important role in natural language understanding because they are used in a broad range of real-world applications such as genomic research, health-informatics, anomaly detection, etc. Next, we briefly describe a few sequence labeling tasks used in this work.

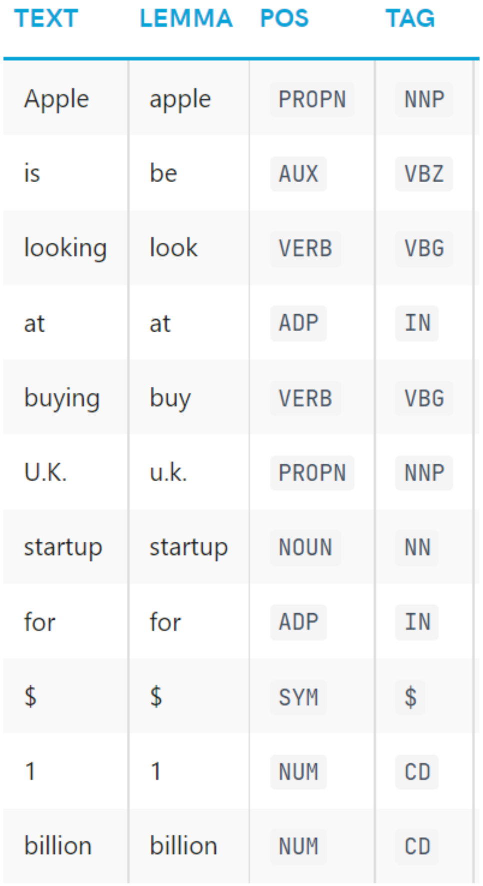
Part-of-Speech (POS)

The syntax of a language is a set of principles in which words are considered grammatically acceptable by fluent speakers. One of the basic syntactic concepts is the part-of-speech that refers to the syntactic role of each word in a sentence. This role is defined by the context in which the word appears. For instance, in the sentence *She eats like a vegetarian*, the word *like* is a preposition, and the word *vegetarian* is a noun.

Part-of-speech labels are morphosyntatic, instead of semantic, categories. They describe words in terms of how they pattern together and how they are internally constructed. In this sense, the *Universal Dependencies* project aims to create syntactically annotated corpora using a single annotation standard by designing a collection of part-of-speech tag-sets. This annotation includes *open class tags* (i.e., nouns, verbs and adjectives), *closed class tags* (i.e., adpositions, numerals and auxiliary verbs) and other tagsets (tags included from the Penn Treebank such as possessive pronouns). Figure 2.6 shows an example that extracts POS tags from the sentence, *Apple is looking at buying U.K. startup for $1 billion.*

Named Entity Recognition (NER)

One of the traditional problems in NLP is to recognize and extract mentions of named entities in text. Unlike part-of-speech that tags each token, the goal of NER is to recover spans of tokens, such as The British Army. Entities often describe people, locations and organizations. However, they have expanded to recognize other types of named entities such amounts of money, percentages, date, among others. NER is also a key task in biomedical natural language processing, with entity types including proteins, DNA, RNA, and cell lines. Similar the POS example, Figure 2.7 shows named entities extracted from the sentence *Apple is looking at buying U.K. startup for $1 billion.*



TEXT	LEMMA	POS	TAG
Apple	apple	PROPN	NNP
is	be	AUX	VBZ
looking	look	VERB	VBG
at	at	ADP	IN
buying	buy	VERB	VBG
U.K.	u.k.	PROPN	NNP
startup	startup	NOUN	NN
for	for	ADP	IN
$	$	SYM	$
1	1	NUM	CD
billion	billion	NUM	CD

Figure 2.6: Example of the POS tags. The image was extracted from the spaCy documentation

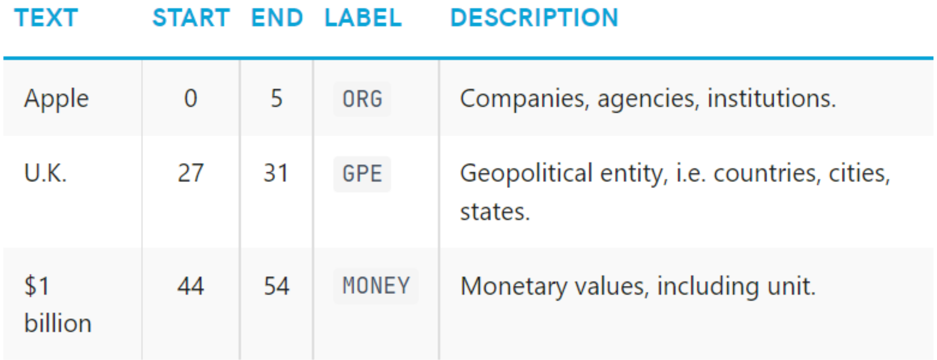
.

2.5 Network Analysis

Network analysis attracts considerable interest in the research community because of the expansion of online social interaction in various mobile and Web applications. A network can be defined as a set of relationships that contains a set of objects or actors (mathematically called *nodes*) and a description of relations between these objects.

Networks are often seen as relationships between people in a online or offline manner [187]. In that case, the connections define friendships between users or groups memberships, following/followers relationships and so on. However, networks can also represent other applications such as physical connections (e.g., a road o bridge connecting two points), biological relationships (e.g., kinship or descent), movement between places or statuses (e.g., migration or physical mobility), etc.

A basic network consists of two nodes and one relationship that links them. Depending on the direction of the relationship, two nodes can have a *undirected* relationship. For instance, two users are mutually connected in a social network (see Figure 2.8). A relationship can be also *directed* where the connection between nodes has a direction. Furthermore, directed relationships can be described according to whether the connection between two



TEXT	START	END	LABEL	DESCRIPTION
Apple	0	5	ORG	Companies, agencies, institutions.
U.K.	27	31	GPE	Geopolitical entity, i.e. countries, cities, states.
$1 billion	44	54	MONEY	Monetary values, including unit.

Figure 2.7: Example of the NER tags. The image was extracted from the spaCy documentation

.

nodes is mutual or not. When the connection is mutual, the directed relationship is called *symmetric*. For instance, similar to the undirected relationship, two users are connected, but the connection has a valence or a flow (see Figure 2.10). Contrary to the above examples, an *asymmetric* connection means that there is a relationship between two nodes, but the connection is not mutual. For instance, *user1* follows *user2*, but the last one does not follow the first (see Figure 2.9).

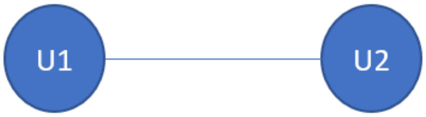


Figure 2.8: Example of an undirected graph

.

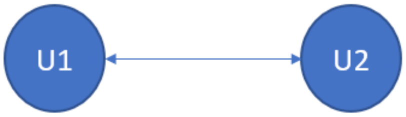


Figure 2.9: Example of a symmetric directed graph

.

We next describe several concepts and metrics used in network analysis to analyze the characteristics of the members and their relationships in a network. For more details about network analysis, please refer to the material of Kadushin [110] and [216]

2.5.1 Basic Concepts of Network Analysis

The addition of graph theory, and its formal mathematical methods, allows us to manipulate much larger and more complex networks. A graph G is an ordered triple $(V(G), E(G), \psi_G)$ consisting of a nonempty set $V(G)$ of vertices, a set $E(G)$ disjoint from $V(G)$, of edges, and an *incidence function* ψ_G that associates with each of G an unordered pair of vertices of G [32]. Figure 2.11 features an example graph. It is formally represented as follows.

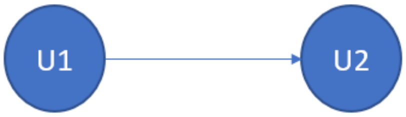


Figure 2.10: Example of an asymmetric directed graph

.

	v1	v2	v3	v4
v1	0	1	1	1
v2	1	0	1	1
v3	1	1	0	0
v4	1	1	0	0

Table 2.5: The adjacency matrix that shows the algebraic representation of the Figure 2.11.

$$G = (V(G), E(G), \psi_G)$$

where

$$V(G) = \{v_1, v_2, v_3, v_4\}$$
$$E(G) = \{e_1, e_2, e_3, e_4, e_5\}$$

and ψ_G is defined by

$$\psi_G(e_1) = v_1 v_2, \psi_G(e_2) = v_2 v_3, \psi_G(e_3) = v_1 v_3, \psi_G(e_4) = v_1 v_4, \psi_G(e_5) = v_2 v_4$$

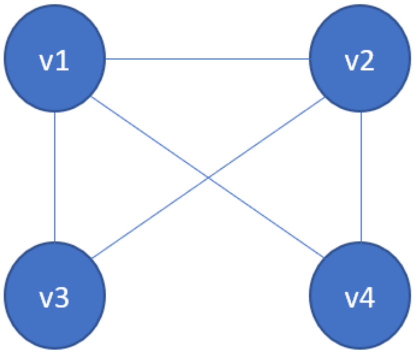


Figure 2.11: Example of a graph

.

In addition, graphs can be also expressed algebraically in order to manipulate them. Table 2.5 shows an adjacency matrix, which is a square matrix used to represent a finite graph. The elements of the matrix indicate whether pairs of nodes are adjacent or not in the network. The adjacency matrix is a $(0, 1)$-matrix with zeros on its diagonal.

2.5.2 Network and Node Descriptives

In graph theory, several measures and indices are used to analyze network's properties to express the relationship between values and the network structures. Next, we define some of these terms that allow us to understand the basic characteristics of a network.

Diameter

Let $I(i, j)$ denote the length of the shortest path (or geodesic) between node i and j. In other words, the distance between i and j. The diameter of a network is the largest distance between any two nodes in the network, which is defined as $diameter = max_{i,j} I(i, j)$.

Density

The concept of *graph density* is defined to be the ratio of the number of edges $|E|$ with respect to the maximum possible edges. Conceptually, it provides an idea of how dense a graph is in terms of edge connectivity. For undirected simple graphs, the graph density is:

$$D = \frac{|E|}{\binom{|V|}{2}} = \frac{2|E|}{|V|(|V| - 1)} \tag{2.21}$$

On the other hand, for directed simple graphs, the maximum possible edges is twice that of undirected graphs to account for the directedness. Hence, the density of directed graphs is defined as follows:

$$D = \frac{|E|}{2\binom{|V|}{2}} = \frac{|E|}{|V|(|V| - 1)} \tag{2.22}$$

Transitivity

Transitivity is the overall probability for the network to have adjacent nodes interconnected, thus revealing the existence of tightly connected communities. It is calculated by the ratio between the observed number of closed triplets and the maximum possible number of closed triplets in the graph.

$$T = \frac{3 \times \text{number of triangles in the network}}{\text{number of connected triples of nodes in the network}}. \tag{2.23}$$

Node Degrees

Node degree is one of the basic centrality measures. It is s equal to the number of node neighbors. Therefore, the more neighbors a node has the more it is central and highly connected, thus having an influence on the graph.

$$\sum_{v \in V} \deg(v) = 2|E| \tag{2.24}$$

Modularity

The modularity Q is a measure of the extent to which like is connected to like in a network. It is strictly less than 1, takes positive values if there are more edges between vertices of the same type than we would expect by chance, and negative ones if there are less. Modularity is defined as

$$Q = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \gamma \frac{k_i k_j}{2m} \right) \delta(c_i, c_j) \tag{2.25}$$

where m is the number of edges, A is the adjacency matrix of G, k_i is the degree of i, γ is the resolution parameter, and $\delta(c_i, c_j)$ is 1 if i and j are in the same community else 0.

2.5.3 Subgroups and Community Detection Algorithms

A community, with respect to graphs, can be defined as a subset of nodes that are densely connected to each other and loosely connected to the nodes in the other communities in the same graph. In this sense, one of the most important tasks in network analysis is to detect communities based on the characteristics of the nodes and edges.

In general, community detection algorithms are grouped following a simple rationale: 1) algorithms designed for static networks; 2) algorithms designed for dynamic networks. For the purpose of this work, we next briefly detail a series of community detection methods utilized for static networks. Furthermore, we focus on those algorithms that generate communities composed by nodes and where nodes belong to one and only one community.

Eigenvector

Newman's leading eigenvector method for detecting community structure is based on modularity. The algorithm creates a modularity matrix and finds the eigenvector for the largest positive eigenvalue. It then labels nodes in communities based on the sign of the elements in the eigenvector [161].

Greedy

The algorithm uses modularity to find the communities' structures. At every step of the algorithm two communities that contribute maximum positive value to global modularity are merged [44].

Infomap

Infomap is based on ideas of information theory. The algorithm uses the probability flow of random walks on a network as a proxy for information flows in the real system and it decomposes the network into modules by compressing a description of the probability flow [189].

Louvain

Louvain maximizes a modularity score for each community. The algorithm optimizes the modularity in two elementary phases: 1) local moving of nodes; 2) aggregation of the network [31]. In the local moving phase, individual nodes are moved to the community that yields the largest increase in the quality function. In the aggregation phase, an aggregate network is created based on the partition obtained in the local moving phase. Each community in this partition becomes a node in the aggregate network. The two phases are repeated until the quality function cannot be increased further.

Stochastic Block Model

The stochastic block model (SBM) is a random graph model with cluster structures. It assumes that nodes are spread into K clusters and uses a $K \times K$ matrix Π to describe the connection probabilities between pairs of nodes [186]. The cluster of each node is given by its binary membership vector Z_i sampled from a multinomial distribution:

$$Z_i \approx M(1, \alpha = (\alpha_1, ..., \alpha_k)), \sum_{k=1}^{K} \alpha_k = 1 \tag{2.26}$$

such that $Z_{ik} = 1$ if i belongs to cluster k and zero otherwise.

Chapter 3

Related Work

This chapter introduces how the massive amount of data published by users on social media platforms support crisis management by providing a near real-time understanding of the events. For this work, we review the state of the art by focusing on three specific tasks. First, we present prior work related to crisis classification and characterization to comprehend the messages' characteristics that allow filtering irrelevant information published in online platforms when crises occur. And second, we focus on crisis detection methods that provide help to traditional instruments and tools (e.g., seismometers) to identify real-world emergency events. In addition to these crisis management tasks, we also review research on group polarization to analyze this phenomenon in the context of emergencies.

3.1 Social Media Messages in Mass Emergency

A considerable amount of work has been published on using social media during emergencies over the past fifteen years. In 2007, Palen and Liu [172] published one of the first papers on the subject in which they explained the relevance of collecting information from wikis about missing people after the attacks of 9/11 in 2001. Akter and Fosso [5] conducted a systematic review indicating that the number of studies on the topic of crisis informatics in the past seven years has continualy increased (See Figure 3.1). In their systematic review they considered results from SCOPUS databases, using the following search terms and their variants: ("disaster management" OR "emergency service" OR "disaster relief operations" OR "disaster resilience" OR "emergency management") AND "big data". Another interesting finding presented by the authors was the scattered spectrum of the research fields with applications of Big Data in disaster contexts. Figure 3.2 displays results for 76 articles studied by the authors. These areas included Engineering, Computer Science, Social Science, Medicine, Environmental Science, among others.

Imran et al. [104] also conducted a survey study in which they presented an extended summary of social media usage during emergency situations. This resulted in more than 150 papers related to these topics, where they included journals, full and short papers from different conferences and workshops. In this sense, we identified top conferences where the

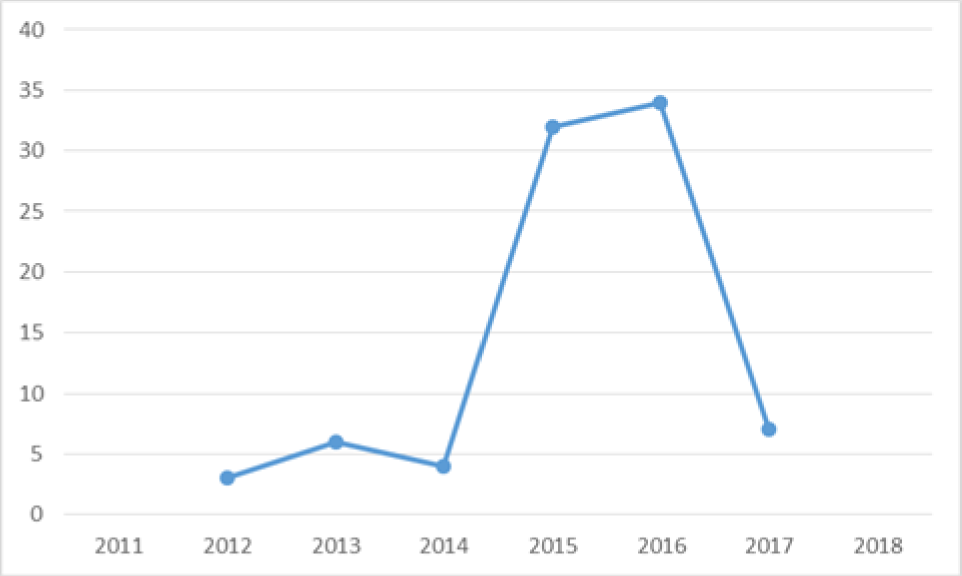


Figure 3.1: Distribution of shortlisted articles by publication year based on the work of Akter and Fosso [5]. This plot only considers works published until the beginning of 2017.

papers have been published, such as: The Web Conference (WWW), Special Interest Group on Information Retrieval (SIGIR), Conference on Information and Knowledge Management (CIKM) and Knowledge Discovery in Databases (KDD). Furthermore, other important ACM, AAAI and IEEE conferences have published several articles in this area, such as The International Conference on Weblogs and Social Media (ICWSM), Web Science Conference (WebSci), Intelligence and Security Informatics (ISI), European Conference on Information Retrieval (ECIR), Hypertext and Social Media, Visual Analytics Science and Technology (VAST), Conference on Human Information Interaction and Retrieval (CHIIR). This interest in the *crisis informatics* field has also derived in the formation of a specific venue called Information Systems for Crisis Response and Management (ISCRAM).

3.1.1 Crisis Characterization

Several insights have been discovered for specific crisis events. For example, Vieweg et al. [213] studied two natural hazard events with the purpose of identifying information that may contribute to enhancing situational awareness. They covered the Red River Flood[1] (RR Flood) and the Oklahoma Fire[2] (OK Fires) natural hazards, both occurring in 2009. This study retrieved messages from Twitter using specific terms such as `red river` and `redriver`, for pulling Red River Flood tweets, and the terms `oklahoma, okfire, grass fire` and `grassfire`, for Oklahoma Grassfire tweets. One of the most relevant results is the percentage of the on-topic (i.e., relevant to the emergency) messages with geolocation

[1] The 2009 Red River flood along the Red River of the North in North Dakota and Minnesota in the United States and Manitoba in Canada brought record flood levels to the Fargo-Moorhead area. The flood was a result of saturated and frozen ground, Spring snowmelt exacerbated by additional rain and snow storms, and virtually flat terrain.

[2] The Oklahoma Fire occurred on April 9, 2009. High winds and dry conditions fueled numerous grassfires burning through central and southern Oklahoma and parts of northern Texas.

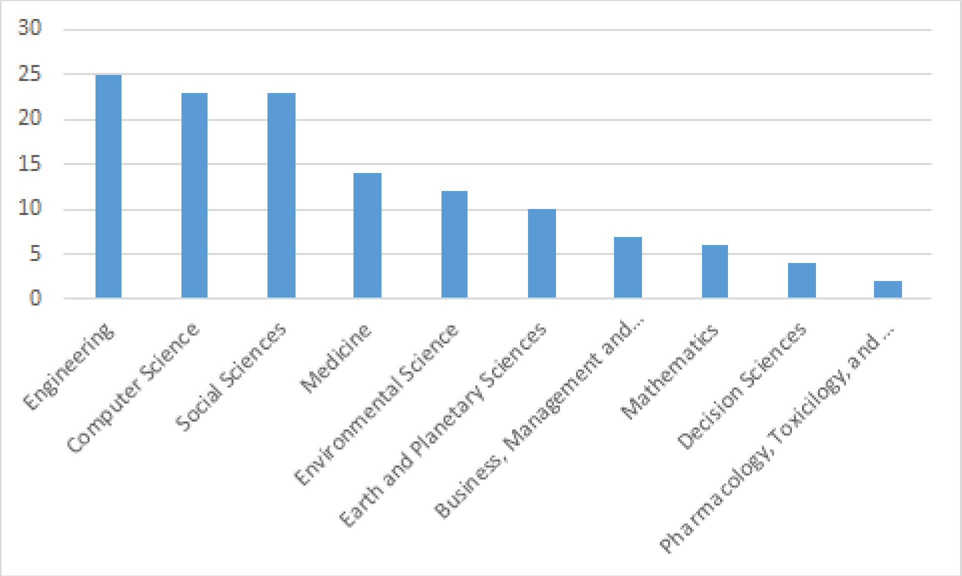


Figure 3.2: Distribution of shortlisted articles by subject areas based on the work of Akter and Fosso [5].

information. Figure 3.3 shows that the most named type of location in on-topic messages is the city hierarchy for both events (30% and 15% for “oklahoma fires” and the “red river flood” respectively). In contrast, location mentioning of country, place, and address have a lower frequency with less than 10% .

Mendoza et al. [150] presented a study related to the 2010 Earthquake in Chile[3] where they found implicit relationships between the emergency situation and the affected locations. The main objective of this work was to show how information propagated through the Twitter network, and to assess the reliability of Twitter as an information source under extreme circumstances. They studied the social phenomenon of the dissemination of the false rumors and confirmed news. The authors retrieved social media messages using the *Santiago* timezone, plus tweets which included a set of keywords. These keywords included hash-tags such as `#terremotochile` and the names of affected geographic locations.

Another work related to emergency situations and location was presented by Kryvasheyeu et al. [124]. In this paper, they found diverse relationships between the proximity of Hurricane Sandy[4] and social media activity. For example, they found several phenomena between the pass of the hurricane along cities and its impact on social media activity. New York City, a city with severe damage during the event, had high social media activity, highly related to the proximity of the hurricane. In addition, they found an inverse relationship between the number of retweets and the level of activity, because affected locations produced more original content. Finally, they observed that the popularity of content was higher in directly affected areas than in others.

Olteanu et al. [168] presented an exhaustive study of 26 crises based on manually anno-

[3] The 2010 Chile earthquake occurred off the coast of central Chile on Saturday, 27 February at 03:34 local time (06:34 UTC), having a magnitude of 8.8 Mw.

[4] Hurricane Sandy was the deadliest and most destructive hurricane of the 2012 Atlantic hurricane season.

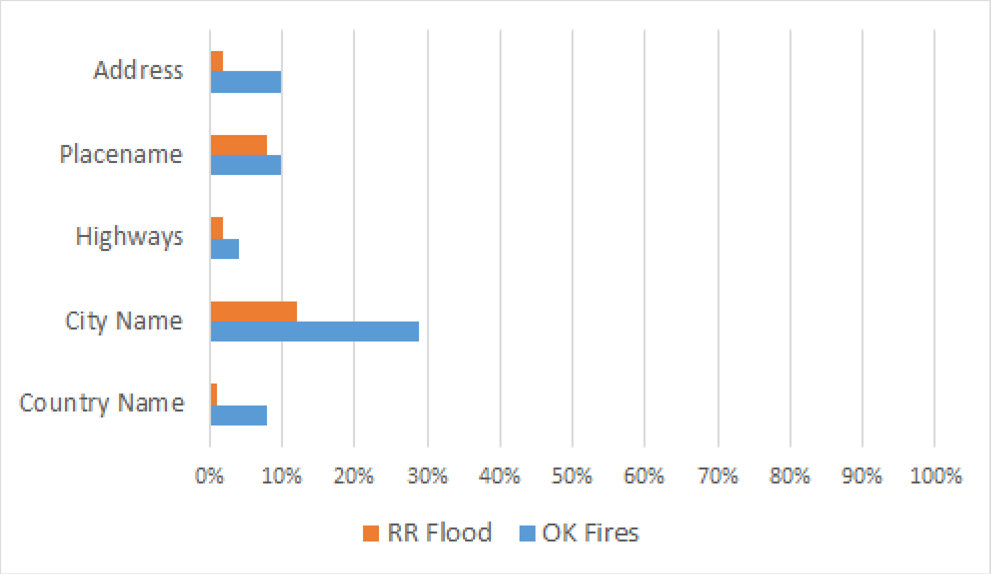


Figure 3.3: Geo-location occurrences as a percentage of on-topic messages based on the work of Vieweg et al. [213].

tated content dimensions. In this article they analyzed diverse types of events depending on the hazard category (natural or human-induced), development (instantaneous or progressive) and spread (diffused or focalized). The authors presented several findings, described next:

- With respect to the types and sources, messages from governments were often about caution and advice, such as tornado alerts. On the contrary, eyewitness tweets focused on affected individuals. Traditional news media and Internet media, on the other hand, offered a variety of information including information about affected individuals, and messages of caution and advice.
- Regarding temporal aspects, they identified differences in the total volume of messages in each information type for instantaneous and progressive events. They additionally demonstrated that in instantaneous crises, outsiders, media and NGO (non-governmental organization) messages appeared early on, though, during progressive events, eyewitness and government messages appear early, mostly to warn and advise those in the affected areas, while NGO messages appear relatively late.

Researchers have also established differences between relevant and irrelevant Twitter messages in crisis situations. For example, Graf et al. [83] presented a multidimensional study using the same social media data collected by Olteanu et al. [168]. Their findings revealed that relevant messages tend to be longer and to contain more nouns and adjectives –among other characteristics– than irrelevant messages. In addition, Ning et al. [163] introduced an analysis of six disasters where they classified informative and non-informative messages using neural networks. In this study the authors showed that these message categories contain differences in linguistic, emotional, entity and topical features. Instead of using typical textual features as previous works, Longhini et al. [135] represented messages using structural content of tweets such as number of followers, verified users, number of URLs, among others.

In addition to analyzing content shared on microblogging platforms like Twitter, researchers have also studied articles related to crises published on Wikipedia. Wikipedia's characteristics allow for users to express their ideas and perspectives from a neutral point of view without biases. Regarding the impact of traumatic events, prior work has found differences in Wikipedia articles related to human-induced events in comparison to natural events. Ferron and Massa [67] compared 55 Wikipedia articles related to human-induced and natural events. They found differences such as that articles about human-induced events contained more anxiety and anger than articles about natural disasters. In contrast, articles about natural events had more sadness-related terms than articles about human-induced events. Similarly, other studies found more negative-related terms in natural disaster articles (such as earthquakes and floods) than human-induced events articles (e.g., terrorist attacks and rail accidents) [87]. As noted, these studies suggest that even on Wikipedia the collective representation of different types of disaster events shows diverse psychological processes [67]. However, the number of Wikipedia editors is quite small in comparison to social media users. Hence, the latter may provide a more representative sample of the population [86].

3.1.2 Crisis Classification

The message classification involves multiple dimensions depending on the granularity of the task. One of the most prevalent tasks is the binary classification of messages that are *related* vs. those that are *not related* to a crisis [168; 45; 9; 209]. In the literature, the term *related* is often interchangeable with the terms *relevant* or *informative*. However, there are several differences among these terminologies regarding the degree of generalization and usefulness for emergency practitioners. First, a message is considered as *related* to a crisis if it includes implicitly or explicitly mention of a disaster event [123]. Second, a *relevant* message contains information that contributes to a better understanding of the situation on the ground [168]. And third, an *informative* message includes helpful information that could improve situational awareness for both citizens and authorities [135]. Literature has also defined fine-grained categories for describing specific requirements to support humanitarian aids. Such categorizations include identification of *personal* comments, *caution and advice*, information about *donations*, among others [102]. Table 3.1 shows a summary of the most relevant work dividing them into the task, language of the messages, and categories used. In more detail, these works will be discussed below.

Classification of relevant/irrelevant information from the data is a difficult task. In fact, social media messages include irrelevant and redundant noise that affects the effectiveness of useful information extraction using traditional techniques. Current methods, described in the state-of-the-art, are based on supervised classification, and have precision and recall of about $75\% - 85\%$, depending on the dataset and the specifics of the task. Commonly used features for identifying useful content include text-based features [101], platform specific features [102; 174], semantic abstractions [116], and word embeddings [132], among others.

Olteanu et al. [167] proposed a crisis lexicon for sampling and filtering crisis-related messages in English during several emergency events. Following this work, Olteanu et al. [168] introduced a large-scale analysis of multiple types of disasters that differ in time and domain. However, the study focused on characterizing differences among labels (e.g., information sources) instead of classifying messages. Cobo et al. [45] studied user and content based features to classify relevant tweets to an earthquake in Spanish (73.4% F1-score) using Random Forest. Alam et al. [9] proposed a deep learning framework based on semi-supervised learning to classify relevant messages in English. They used two Twitter datasets: one of the Nepal earthquake (65.11% F1) and another of the Queensland floods (93.54% F1). Li et al. [132] proposed a feature-based adaptation framework, which considers pre-trained and crisis-specific word embeddings, as well as sentence embeddings and supervised classifiers. They evaluated two classification tasks for English tweets considering three crisis datasets. Their results showed an average accuracy of 88.5% considering the best overall configuration using pre-trained GloVe word embeddings, MinMaxMean aggregation and SVM classifier. They noted that crisis-specific embeddings were more suitable for more specific crisis-related tasks (*informative* vs *non-informative*), while the pre-trained embeddings were more suitable for more general tasks (*relevant* vs *non-relevant*). Firoj et al. [68] performed a comparative study among various algorithms used to classify crisis-related messages. They reported that Support Vector Machines, Random Forest and Convolutional Neural Networks provided competitive results.

In view of the difficulty of comparing results, models and techniques in this research area, Alam et al. [12] developed a standard dataset based on existing data and provided train/dev/test partitions. The authors provided benchmark results on English messages for informative (binary) and humanitarian (multi-class) classification tasks using deep learning algorithms. Alam et al. [11] created a large-scale dataset of English-language tweets, which is composed of 19 disaster events that occurred between 2016 and 2019. The authors report the results of the classification of humanitarian information using classical and deep learning algorithms. They achieved an average weighted F1 of 78.1% with the RoBERTa model [134].

Most recent studies have also considered cross-lingual and multi-domain adaptations for crisis-related messages classification. These approaches are required because most NLP resources and labeled datasets are adapted for content published in English. Torres and Vaca [209] compared traditional and deep learning models using sparse representations and word embeddings to classify earthquake-related conversations in English and Spanish. Using a Long Short Term Memory model including multilingual stacked embeddings for cross-lingual classification, they reported a macro F1-score of 85.88% from Spanish to English and 77.49% from English to Spanish.

Lorini et al. [137] evaluated pre-trained language-agnostic and language-aligned word embeddings with Convolutional Neural Networks for classification of flood-related messages in German, English, Spanish, and French. They compared a monolingual classifier, a cross-lingual classifier with *cold start* (using no training data in the target language), and a cross-lingual classifier with *warm start* (using 300 labeled instances in the target language). In the case of the monolingual classification, results achieved an F1-score ranging from 70% to 87%. On the other hand, the cross-lingual classification reached values between 48% and 86%, depending on the target language. Furthermore, they showed that both types of word embeddings could be used to classify a new language for which few or no labels are available. However, including a small set of data from the same target language improved the cross-lingual classification. Another approach was proposed by Khare et al. [116], who considered messages in English, Italian and Spanish from 30 crisis events of different types. They proposed a statistical-semantic crisis representation, extracting semantic relationships from BabelNet and DBpedia knowledge bases. They achieved a cross-lingual classification F1-score of 59.9% on average.

Li et al. [132] proposed a feature-based adaptation framework, which considers three types of pre-trained and crisis-specific word embeddings (Word2Vec, GloVe and FastText), and four supervised classifiers (Gaussian Naive Bayes, Random Forest (RF), K Nearest Neighbors and Support Vector Machines (SVM)). They evaluated two classification tasks on English tweets for three crisis datasets, reaching an average accuracy of 88.5% considering the best overall configuration (GloVe, MinMaxMean aggregation and SVM). The authors noted that the crisis-specific embeddings are more suitable for more specific crisis-related tasks (*informative* vs *non-informative*), while the pre-trained embeddings are more suitable for more general tasks (*relevant* vs *non-relevant*). In addition, they found that SVM and RF have competitive results.

Imran et al. [106] validated the impact of adding training examples from a different domain than the target. They classified messages related to earthquakes and floods published in several languages. Their experiments showed that in scenarios where there is not enough data, increasing training examples with tweets from other languages can be useful if both are very similar (e.g., Italian and Spanish). In the case of domain adaptation, they concluded that using tweets from another domain did not appear to improve performance.

As noted in Table 3.1 and the above literature review about crisis classification, approaches differ in data, languages, targets, and methods. In this sense, existing works have mainly focused on understanding the effect of adding instances from other languages or domains to the target crisis. However, these approaches consider the existence of labeled data for the current event (or domain), which in real scenarios of crises does not exist. Our work differs because we perform a systematical study of transfer learning for crisis message classification for scenarios in which little to no data is available. We focus on the case of how to leverage labeled data from high-resource to low-resource languages, as well as from well-known crisis domains to new domains. Furthermore, we put effort into the data and experimental methodology instead of classification algorithms. Finally, we study which document representations and models work best for each specific target.

3.1.3 Crisis Detection

One main task related to emergency situations is to detect a new real-crisis event in social media. Most existing event-detection methods described in the literature are based on keywords.

TweetTracker, presented by Kumar et al. [125], consists of a case study of tweets discussing the 2010s Haiti cholera outbreak[5]. The primary mechanism for monitoring tweets were through specific keywords and hashtag filters related to Haiti. To detect a new event, emerging trends were identified based on the analysis of older tweets. The system architec-

[5] The Haitian cholera outbreak was the first modern large scale outbreak of cholera, once considered a beaten back disease thanks to the invention of modern sanitation, yet now resurgent, having spread across Haiti from October 2010 to May 2017

ture of TweetTracker consists of four major components: the Twitter Stream Reader, where they retrieved messages based on user specified keywords, hashtags and geolocations. The DataStore, where data was constantly stored. And the Visualization and Analysis Module, where tweets were analyzed and filtered. Later, a map was included to focus on tweets of interest.

EMERSE, presented by Caragea et al. [37], used a set of keywords related to the Haiti earthquake and applied a SVM algorithm to classify messages. The main goal was to translate and classify messages for different languages. Furthermore, various sources were considered such as tweets and short message service (SMS) about Haiti disaster relief.

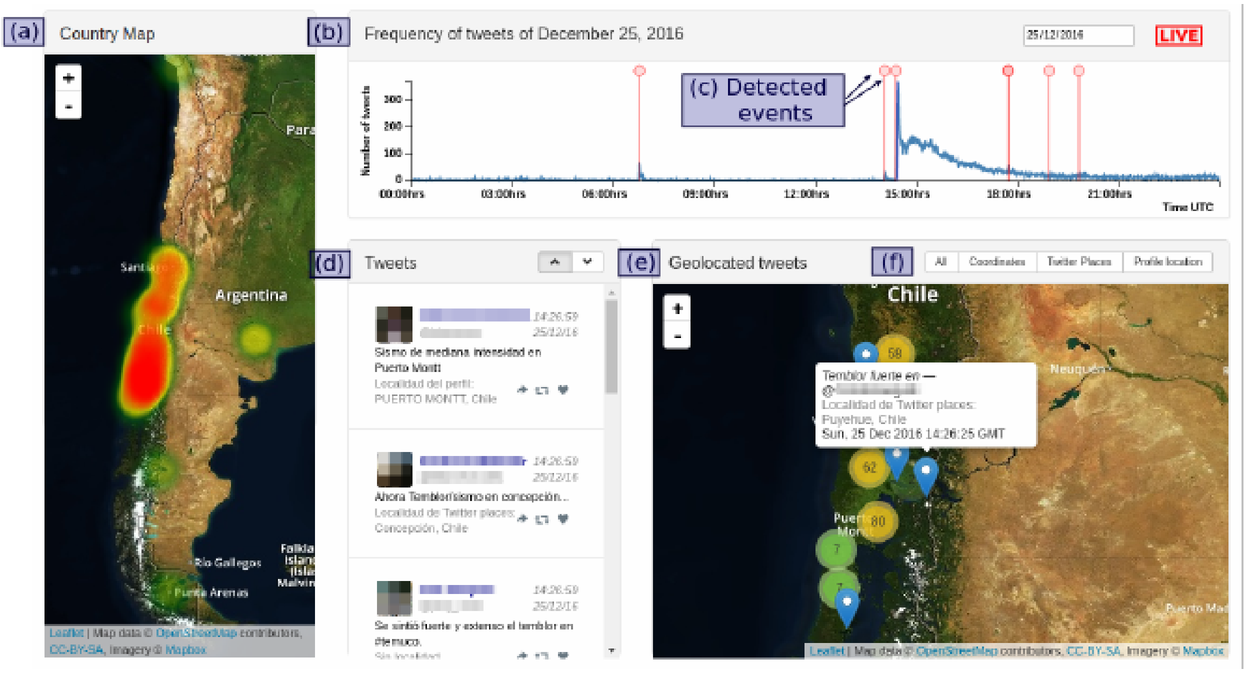
Like the *Twicalli* system [143], the authors introduced an unsupervised approach to detect earthquakes that only requires a general list of keywords. It was based on the work of Guzman and Poblete [93], where the authors detected burst activity in social media using static time-windows for determining variation of the terms using the z-score value. In Twicalli, the main idea was to retrieve messages using specific keywords for earthquakes in several languages. Messages were later filtered by their geolocation and assigned by country. They next computed z-score variations between time-windows related to earthquake terms. Finally, they visualized earthquake detections and messages in a website[6] as we can see in Figure 3.4.

Researchers at CSIRO Australia proposed *ESA* [36; 220], a system to detect disasters in Australia and New Zealand. This system was based on a probabilistic method that identifies bursty keywords, and historical data to build a language model of word occurrences. Alerts were identified if a term had a probability distribution that significantly deviates from the language model. After detecting an event, they applied clustering to get to the topics discussed for the targeted incident.

Similarly, the *Twitcident* [1] system detected incidents that rely on emergency broadcasting services, such as the police, the fire department and other public emergency services. The *Twitcident* framework translated the broadcasted message into an initial incident profile applied as a query to collect messages from Twitter, where an incident profile is a set of weighted attribute-value pairs that describe the characteristics related to the incident.

Finally, *AIDR* [103] is a platform that performs automatic classification of crisis-related microblog communications. The goal of the AIDR is to classify messages that people post during disasters into a set of user-defined categories based on the works of Imran et al. 2013a [102] and Imran et al. 2013b [101]. The authors presented an architecture where they collected crisis-related messages from Twitter, asked a crowd to label a sub-set of those messages, and trained an automatic classifier based on the label. There are two important points in this work. First, they did not use pre-existing training data because it was not a satisfactory solution, given that crises had elements in common, and also had specific aspects, which make domain adaptation difficult. The second point was that AIDR is not a system for continual event tracking. AIDR just tracks events when an instance is created.

[6] Twicalli website http://twicalli.cl/



[redacted] 14:26:59 25/12/16
Sismo de mediana intensidad en Puerto Montt
Localidad del perfil: PUERTO MONTT, Chile

[redacted] 14:26:59 25/12/16
Ahora Temblor/sismo en concepción...
Localidad de Twitter places: Concepción, Chile

[redacted] 14:26:59 25/12/16
Se sintió fuerte y extenso el temblor en #temuco.
Sin localidad

Temblor fuerte en —
@[redacted]
Localidad de Twitter places: Puyehue, Chile
Sun, 25 Dec 2016 14:26:25 GMT

Figure 3.4: A visual summary of the Twicalli website. The visual interface showing an earthquake occurred on December 25th of 2016. (a) Heat map of the complete country. (b) Signal formed by the number of published tweets every 60 seconds. (c) Marker of detected event, on click, information related with the event is displayed. (d) Last published tweets with buttons to reorder. (e) World map with clustered markers; user can see here when an event identified in the signal is occurring in other countries. (f) Buttons that filter the markers considering the source of location information, so users can choose messages in which they trust more, because some location sources are less trustworthy than others [143].

3.2 Group Polarization in Social Media

Group polarization occurs when the tendency of individual group members is enhanced by group discussion [204]. It can often trigger more radical group decisions than those generated by average individuals in the group [108]. In recent years, polarization has been widely studied within the context of online discussions. Specifically, social media has greatly increased the volume of online exchanges among users, in particular about social and political issues.

When discussing controversial issues, online users tend to be exposed to agreeable opinions [24]. One of the explanations of this exposure is homophily, where "individuals associate with similar ones" [27]. This phenomenon, among other factors related to media consumption, reinforces users' perceptions and blinds them from other sides of the issues under discussion [16]. Then, the primary input for studying polarization is to find the several stances present in the discussion and then find groups of users with the same stance. Having this categorization helps measure and mitigate the problems derived from polarization in social networks. Next, we discuss the literature used for studying user polarization in social media platforms by dividing them into the most common approaches to analyze this issue. Additionally, we summarize these techniques in Table 3.2 by dividing them into different categories, such as tasks, approaches, ML technique and datasets.

3.2.1 Content-based Polarization Analysis

Studies analyzing polarization in social media have relied on aligning users toward a set of specific topics or entities [14]. The primary assumption is that communities are preliminarily identifiable based on established target topics, use common hashtags and vocabulary for label propagation, and consider a set of seed users for constructing communities. However, the prior manual labeling can be costly in real large networks, in terms of time, distribution of stances in a dataset, quality of the inter annotators given the topic expertise, and the lack of ground-truth datasets, among other issues. Traditional features used to determine polarization are based on extracting characteristics derived from content published by users.

Authors have considered n-grams to capture the stance of the users in subjects such as abortion and gay marriage [17], and legalization of abortion and climate change [155]. Studies have also focused on the user's vocabulary. The hypothesis is that individuals with the same stance tend to use the same vocabulary choices to express their points of view [53]. The work of Klebanov et al. [119] demonstrated that people with a similar stance (e.g., against abortion) tend to use recurrent analogous vocabulary for supporting their ideas.

Using an embedding approach, Benton and Dredze [26] proposed a semi-supervised method to represent users based on their online activity. The general idea is to use the context of the users' tweets to construct author embedding and then predict the stance. Similarly, Li et al. [131] considered a joint embedding learning to determine users' stance using the Internet Argumentation Corpus. For each topic, the authors created individual embedding vectors, which represent pro and against stances. Taking a case study of the Turkey elections, Kutlu et al. [126] trained an embedding vector using fastText with a skip-gram model on related tweets. The authors relied on the work of Garg et al. [75], which demonstrated that word embeddings capture gender and racial stereotypes by comparing word vectors that are trained on different corpora to understand how a given term is defined semantically. Considering these results, the work presented by Kutlu et al. [126] used several word vector models for each politician and stance (e.g., pro and anti-Erdogan). Using a list of known-polarized adjectives and political terms, they compared the 2,000 nearest neighbors' word of each to understand the difference among low-dimensions vectors qualitatively.

Demszky et al. [55] presented a study of 21 U.S. mass shootings events to measure polarization in common frames in Twitter. Considering a predefined list of Twitter accounts of U.S. Congress members and presidential candidates, they applied a label propagation method to determine the users' political party. They trained a word embedding model to estimate frames, applied k-means clustering to discover common concepts, and manually assigned topics names to inspect the tweets. Finally, they computed a leave-out estimator to measure polarization between and within partisanships for each frame.

3.2.2 Network-based Polarization Analysis

Social media platforms yield a rich interaction structure wherein users communicate and connect with others in several ways. For example, sharing ideas from other users (a retweet action), replying to messages, mentioning others, and using similar hashtags in their content.

One of the traditional approaches that have been widely utilized to infer users' attitudes is the retweet network. Guerrero-Solé [90] analyzed the Catalan process toward independence on the 1,000 most retweeted users. To detect communities, they assigned a label to every edge to identify them by political orientation. They performed an iterative process by which a given user inherits a set of users' labels retweeted. Concerning Egyptian political polarization, Borge-Holthoefer et al. [33] presented a network approach to track polarity evolution over time. Using a label propagation algorithm for detecting communities, they identified polarized groups considering an initial list of seed users for whom the partisan leaning was clear. A study of polarization in Egypt about Secular and Islamists in several languages also considered an analysis of the retweet network [217]. Using the NodeXL and Fruchterman-Reingold community algorithms, authors showed a polarized network describing Islamist, Secularist, and Center stances. Similar to the work of Borge-Holthoefer et al. [33], the proposal needed a seed of users to obtain polarized communities.

Other works have also considered network features, but including these characteristics as attributes for supervised, semi-supervised and unsupervised approaches. For instance, Darwish et al. [52] predicted online Islamophobia over time using the 2015 Paris terrorist attack in Paris as a case study. Among other features, authors considered network features such as the accounts that a user mentioned, retweeted, and replied to. Considering three polarized events, Darwish et al. [53] presented an unsupervised framework for detecting stance on Twitter. Their approach extracted several network characteristics such as the number of unique tweets, hashtags, and retweeted accounts with computing similarity among users. They therefore, applied dimensionality reduction and clustering techniques to obtain communities. Following a similar approach, Stefanov et al. [202] identified an initial set of users' stances based on the previous methodology and then trained a classifier to determine the position of the others. Both studies reported accuracy and f1-score values over 80% obtaining two clusters on average.

Expanding the analysis using retweet networks, other works have included additional structures, such as reply and follower graphs. Coletto et al. [46] studied controversial topics in Twitter, considering a motif-based approach that enriches traditional graph features (i.e., network structure and temporal characteristics) to predict if a conversation thread is controversial or not. Garimella et al. [76] proposed a graph-based three-stage pipeline to quantifying controversy in social media, which involves the creation of a conversation graph about a topic, identifying potential sides of the controversy, and measuring the amount of controversy based on the structure of the graph. In these two mentioned articles, both works require a seed of initial keywords (or topics) to analyzed controversial themes. To understand long-term polarization effects in Twitter, Garimella and Weber [77] analyzed the increasing of US political polarization in the last eight years. Their analysis relied on re-constructing retweet, followers and shared hashtags networks among others. The authors claimed that polarization increased, depending on the measure, between 10% and 20%.

Table 3.1: This table considers only those works that collected and labeled data, most of which are publicly available. These works range from 2014 to 2019. The table is an adaptation presented by Sánchez Macías [190]

Article	**Task**	**Lang.**	**Label**	**Definition**
Olteanu et al. [167]	Filtering messages that are related to a crisis situation.	English	***On-topic***	Directly or indirectly related to a disaster.
			Off-topic	Not related to the disaster.
Olteanu et al. [168]	Identify crisis-related messages by informativeness, information type and source.	Multiple languages. However, non-English languages are omitted	***Related and informative***	Contains useful information that helps stakeholders understand the situation.
			Related but Not informative	Refers to the crisis, but does not contain useful information that helps stakeholders understand the situation.
			Not related	Not related to the crisis.
Cresci et al. [51]	Classify message by damage assessment.	Italian	***Damage***	Related to the disaster and relevant. Refers to damage on the infrastructure or the population.
			No damage	Related to the disaster but not relevant.
			Not related	Not related to the disaster.
Cobo et al. [45]	Identify relevant and irrelevant messages to a crisis situation.	Spanish	***Relevant***	Belongs to Caution and advice, Casualties and damage, People missing, found, or seen, or Information source.
			Not relevant	Not related to the situation.
Imran et al. [105]	Classify messages by information type.	English, Spanish, French	***Information types***	Categories of interest include Injured or dead people, Missing, trapped, or found people, Not related or irrelevant, among others.
Alam et al. [10]	Classify message by informativeness, humanitarian categories, and damage severity.	English	***Informative***	Tweet that is useful for humanitarian aid.
			Not informative	Tweet that is not useful for humanitarian aid.
Purohit et al. [181]	Classify and rank actionable requests for Emergency Operation Centers.	English	***Serviceable***	Must contain an explicit request or an answerable question, correctly addressed and sufficiently detailed.
			Not serviceable	Messages that are not a priority for operational response, such as complaints, gratitude, congratulations, and advertisements.
Torres and Vaca [209]	Identify messages related and non-related to a crisis situation.	English, Spanish	***Crisis related***	Belongs to Injured or dead people, Missing or found people, Displaced people and evacuations, among other categories.
			Not related	Not related or irrelevant.

Article	Task	Approach	ML technique	Dataset/event
Aldayel and Magdy [14]	Relationship between stance and sentiment	Content-based	None (statistics analysis)	SemEval stance dataset
Anand et al. [17]	Stance classification	Content-based. N-grams, LIWC and grammatical features	Supervised	Topics from Convenceme.net
Mohammad et al. [155]	Stance detection and classification	Content-based. N-grams	Supervised	SemEval stance dataset
Darwish et al. [53]	Stance detection	Both content-based and network-based approaches.	Unsupervised	Twitter political discussions
Klebanov et al. [119]	Stance classification	Content-based. Vocabulary selection	Supervised	Multiple debate corpora
Benton and Dredze [26]	Stance classification	Content-based. Word embeddings	Supervised	SemEval stance dataset
Kutlu et al. [126]	Polarization analysis	Content-based. Word embeddings	Unsupervised	2018 Turkey elections
Demszky et al. [55]	Polarization and framing analysis	Both content-based and network-based approaches.	Semi-supervised	U.S mass shootings
Guerrero-Solé [90]	Polarization analysis	Network-based. Retweet network analysis based on a set of seed users	Unsupervised	The Catalan Referendum for Independence
Borge-Holthoefer et al. [33]	Polarization analysis	Both content-based and network-based approaches.	Supervised	Egyptian political sphere
Weber et al. [217]	Polarization analysis	Network-based. Retweet network analysis based on a set of seed users	Supervised	Egyptian political sphere
Coletto et al. [46]	Polarization analysis	Network-based. Retweet network analysis based on graph patterns	Supervised	Twitter controversial pages
Garimella et al. [76]	Polarization analysis	Both content-based and network-based approaches.	Semi-supervised	Twitter controversial pages

Table 3.2: Summarization of content-based and network-based approaches in the literature to address polarization analysis and stance detection problems.

Chapter 4

A Domain-Independent Crisis Detection Approach

Existing methods for automatically detecting emergency situations using Twitter rely on features based on domain-specific keywords found in messages. These keyword-based methods usually require training on domain-specific labeled data, using multiple languages, and for several types of events (e.g., earthquakes, floods, wildfires, etc.). In addition to being costly, these approaches may fail to detect previously unexpected situations, such as uncommon catastrophes or terrorist attacks. However, collective mentions of certain keywords are not the only type of self-organizing phenomena that may arise in social media when a real-world extreme situation occurs. Just as nearby physical sensors become activated when stimulated, localized citizen sensors (i.e., users) will also react in a similar manner.

In this chapter, we present a method to use self-organized activity related to geolocations to identify emergency situations. We propose to detect such events by tracking the frequencies, and probability distributions of the interarrival time of the messages related to specific locations. In this direction, our method to identify new unseen events can be adapted to detect any type of crises and for different languages because it depends only on bursty activity related to locations. Using an off-the-shelf classifier that is independent of domain-specific features, we study and describe emergency situations based solely on location-based features in messages. Among other results, our findings indicate that anomalies in location-related social media user activity indeed provide information for automatically detecting emergency situations independent of their domain.

The work presented in this section was published as follows: *Sarmiento, H., Poblete, B., & Campos, J. (2018, May). Domain-Independent detection of emergency situations based on social activity related to geolocations. In Proceedings of the 10th ACM Conference on Web Science (pp. 245-254)..*

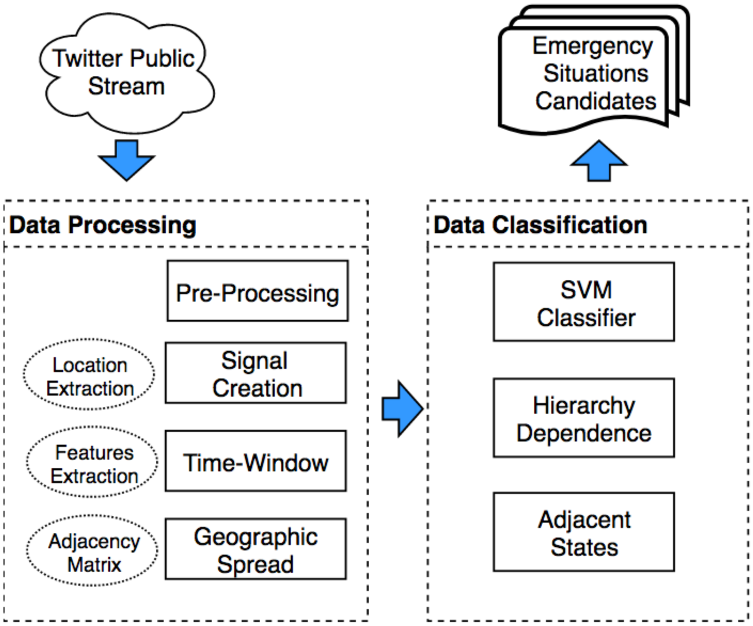


Figure 4.1: Key components of the proposed approach.

4.1 Proposed Approach

Our focus in this thesis is to detect an emergency situation based on identifying anomalies in social media activity related to locations. In this way, the main task is to extract locations from messages by reducing the noise and irrelevant information.

Figure 4.1 shows a general overview of our proposal. The *data processing* module describes the data extraction process from Twitter. Key components are divided in four tasks: (1) We pre-processed data to allow a better analysis, because social media messages are often noisy and redundant. (2) We created discrete signals based on location extraction using a geographical dictionary (also known as *gazetteer* [59]) to create a geographic hierarchy for a specific country. In addition, we created signals in various levels of the geographic hierarchy and the tweet metadata. (3) We divided signals into fixed time-windows and computed non textual features for each. (4) In order to discard those detections that occur in isolated and non-connected locations, we created a geographic spread based on the proximity between locations by using an adjacency matrix M, which an element M_{ij} represents whether a location i is directly connected with a location j.

On the other hand, the *data classification* module describes the data classification task in which emergency situations candidates are obtained. The main steps include the following phases: (1) We trained and evaluated Support Vector Machine (SVM) classifiers for each geographic hierarchy. (2) We evaluated classification predictions considering the dependence on hierarchies among signals. (3) Using a *Adjacency Matrix* to represent neighborhoods between regions/states, we considered a geographic spread analysis to reduce the amount of false positive results generated by detections on isolated locations.

4.1.1 Dataset Description

Our hypothesis is to find empirical evidence that we can identify an emergency situation without specific domain keywords over the Twitter stream. Hence, we needed to retrieve random messages about any topic and any place in the world. Additionally and strictly, a portion of the messages must contain information about an emergency situation.

Generally, several works use public datasets to improve and compare techniques. For example, the most common available catalog is *TREC (Text REtrieval Conference)*[1]. In this resource, we find topics as confusion track, query track, question answering track, microblog track and others. In the last mentioned track, the goal is to explore technologies for monitoring a stream of social media posts with respect to a user's interest profile. However, the identified interest profiles do not represent an evaluation that allows for evaluating emergency events. We therefore could not use TREC for our evaluation methodology.

In contrast, we generated our own dataset based on the messages retrieved from Twitter. For this work we collected data from Twitter Public Streaming API[2], which allows access to subsets equal to 1% of public status descriptions in real-time. With this tool we retrieved either messages using a set of keywords or messages from specific locations setting a "bounding box". In our approach, we got entire subsets of messages without using keywords or specific locations. In addition, this subset of public status descriptions represent a good sample of the full status published in Twitter for high-impact real-world events [158]. Hence, we retrieved messages related to any topic, written in any language and posted anywhere in the world in this micro-blog service.

Ground Truth

To construct our ground truth, we first identified *instantaneous-diffused* crises according to the definition presented by Carr [38]. These events are characterized by the fact that no one could do anything to prevent them and their effects were felt by an entire community. This definition is relevant because an unexpected (or instantaneous) event generates an anomaly in the frequency of the social media activity since it disrupts users' normal life. Furthermore, diffused events affect a large portion of users simultaneously, generating a collective reaction in neighboring affected locations.

Several works in literature have studied earthquakes detections in online platforms because they cannot predict, independent of the instrument or type of data [143; 37; 36]. Hence, it is possible that the lack of prior precedent could contribute to the generation of anomalies in the frequency of messages in social networks. Based on the aforementioned descriptions, we chose earthquakes (considered as instantaneous-diffused crises) as crises to study unexpected events that affect thousands or millions people at the same time.

We analyzed five earthquakes with magnitudes between $5.5Mw$ and $7.6Mw$[3], which oc-

[1] https://trec.nist.gov/

[2] https://developer.twitter.com/en/docs/tutorials/consuming-streaming-data.html

[3] Mw: the moment magnitude scale

curred in Italy and a Spanish-speaking country (such as Chile) between October 2016 and April 2017 (Table 4.1). Due to the absence of specific event-related information in the metadata retrieved, we relied on official information from the Chilean and Italian national seismological centers to establish our ground truth. Although Twitter's current tweet object model includes a context annotations field for entity recognition and extraction for topical analysis, such information was unavailable during the period of data collection for this research. Consequently, we obtained the precise date, time, and location of the event from the National Seismology Agency in Chile[4] and the National Institute of Geophysics and Volcanology in Italy[5].

In order to identify a set of messages related to a crisis event and those that were not (e.g., non-crisis related events or a normal situation), we extracted messages from 12 hours prior to and following the occurrence of an emergency situation. We assumed that anomalies in message volume were indicative of unseen and unexpected events that had affected users' daily lives. Literature on this matter has demonstrated that the first 12 hours following a crisis are critical [104; 39]. Hence, we included messages from this temporal window in order to minimize delays in crisis detection. Moreover, we incorporated messages from a period prior to the emergency to quantify the proportion of non-crisis events that were falsely detected as a crisis (i.e., false positive detection). To this end, we retrieved 20 million Twitter messages related to any topic without any keyword or location filters.

Table 4.1: List of earthquakes studied as ground truth, sorted by date.

Country	Datetime (UTC)	Magnitude (Mw)	Language
Italy	2016-10-26 17:10:36	5.5	Italian
Italy	2016-10-30 06:40:17	6.6	Italian
Chile	2016-12-25 14:22:26	7.6	Spanish
Chile	2017-04-23 02:36:06	5.9	Spanish
Chile	2017-04-24 21:38:28	6.9	Spanish

4.1.2 Data Pre-Processing

Our focus is on localized bursty user activity. In other words, we aim to detect location-based anomalies over time when an emergency occurs. The general idea behind these concepts is that users act as citizen sensors distributed in several locations, which react and are stimulated by external factors (in our case, crises) and then share their status on social media platforms.

In an ideal scenario, burst user activity could be estimated using the device coordinates (GPS) incorporated in the message's metadata. However, just a tiny portion of messages contain that metadata [84] followed by the random sampling of 1% that users can retrieve from the Twitter API. Therefore, since our focus is on users who could be at the event

[4] http://www.sismologia.cl/

[5] http://www.ingv.it/it/

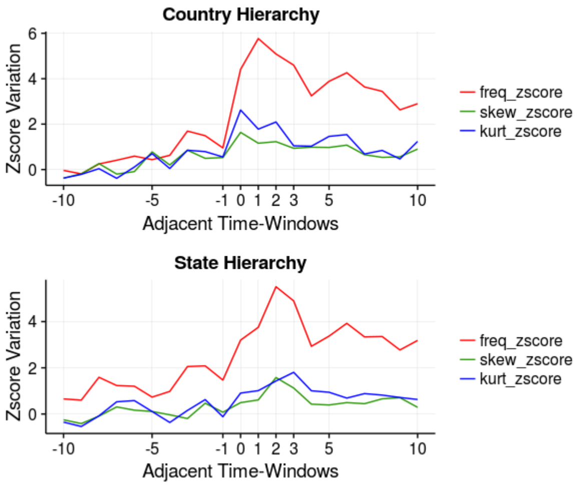


Figure 4.2: Average variation in emergency situations between time-windows. Positive and negative values in x-axis represent the following and previous time-windows from the beginning of the event respectively.

location, we initially filter messages according to the most common language used in each country (using the attribute *lang* in tweet metadata[6]). This step will allow us to create a hierarchy of gazetteers only for the primary language in which the event occurred, reducing (as much as possible) the noise and irrelevant content published from a location external to the event.

Although language does not infer the specific location or coordinates from where a message has been sent, it allows us to have an initial filter to reduce the number of messages posted from another place or country. For example, when analyzing Italy, we only consider messages written in Italian.

In addition, we remove user mentions, URLs, and special characters and apply text tokenization. We do not remove hashtags or stopwords because some locations can be included as hashtags, and some location names contain stopwords, which differentiates them from other locations or terms.

4.1.3 Signal Creation

We create a set of discrete-time signals for each location, which indicates the time that each message related to a specific location was posted. In order to explain the effect of an emergency situation in a local and national scope, we use the lowest possible geographical hierarchy level available with the aim of comparing the impact in the highest level. Furthermore, we study the anomalies at various metadata levels to understand how locations are shared in Twitter. For instance, those signals either based in the locations set by users in their profile or in the locations shared in their messages.

[6] https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/tweet-object

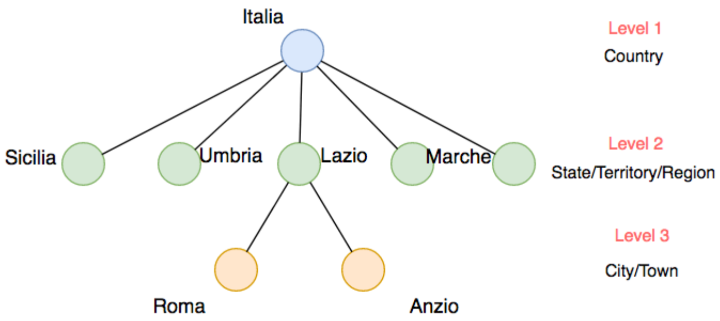


Figure 4.3: Example of gazetteer tree for Italy.

Geographical Hierarchy

We use the idea of *gazetteer as a tree* presented in Yin et al. [221] in which each place is associated with a canonical taxonomy node. We create our gazetteer tree based on Geonames[7] and Wikipedia[8]. However, in Yin et al. [221] the gazetteer hierarchy presents four levels where the lowest level represents a specific point of interest. In our approach, we use a subset of the gazetteer hierarchy with only three levels: *city*, *state* and *country*. We do so because a large number of users specify their location down to city level [96]. For example, if we have the *city:Manchester*, we associate this location with *region-state:North West* and also with *country:England*. As indicated in our data pre-processing stage, we consider only locations in the native language of the country. For instance, in the case of Italy locations, we consider *Roma* and not *Rome* (Figure 4.3).

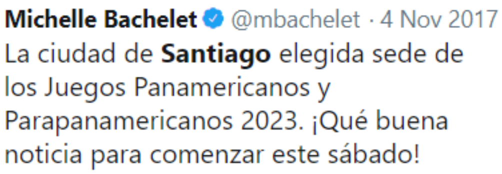
Location Extraction

The structure of the tweet metadata contains information about the message and the user. Given a small portion of users sharing their current location using GPS coordinates [84], we do not consider this level of the tweet metadata in this work.

Considering the aforementioned geographical hierarchy, we extract locations from different parts of the metadata, creating 3 signals for each location:

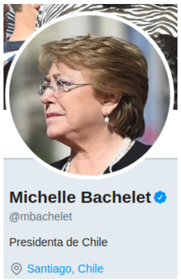
- Tweet Text: location is mentioned in the attribute `text` of tweet object, that is, on the body of the message. Figure 4.4 has an example of where the location *Santiago* is mentioned in the message.
- User Location: location is mentioned in the attribute `location` inside the `user object`, that is, the location set by the user in the profile. An example is in Figure 4.5 where the location *Santiago* is mentioned in the user profile.
- Tweet Text - User Location: location is mentioned in the attribute `text` of tweet object and also location is mentioned in the attribute `location` inside the `user object`. This

[7] http://download.geonames.org/export/dump/
[8] http://www.wikipedia.org/



Michelle Bachelet @mbachelet · 4 Nov 2017
La ciudad de **Santiago** elegida sede de los Juegos Panamericanos y Parapanamericanos 2023. ¡Qué buena noticia para comenzar este sábado!

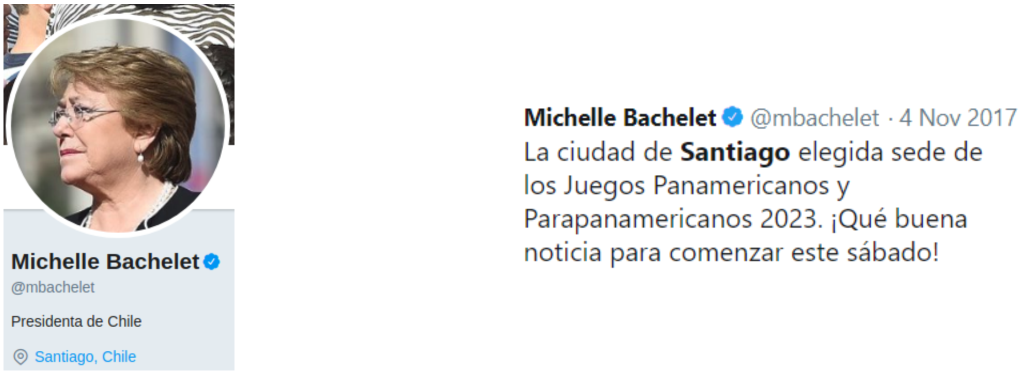
Figure 4.4: Example of location mentioned on the body of message.



Michelle Bachelet
@mbachelet
Presidenta de Chile
Santiago, Chile

Figure 4.5: Example of location mentioned in the user profile.

means that the location is mentioned in the body of the message and the user who shares the message has the same location in the profile. In this case, tweet text and user location can be different in the smallest hierarchy, but in the highest level can be the same location. Figure 4.6 shows an example where the location *Santiago* is mentioned in the message and the user who shares a message has the user location profile in *Santiago*.



Michelle Bachelet
@mbachelet
Presidenta de Chile
Santiago, Chile

Michelle Bachelet @mbachelet · 4 Nov 2017
La ciudad de **Santiago** elegida sede de los Juegos Panamericanos y Parapanamericanos 2023. ¡Qué buena noticia para comenzar este sábado!

Figure 4.6: Example of location mentioned on the body of message and the user profile.

In this way, by combining geographical hierarchy and locations in microblog metadata, we create $N \times M$ signals where N is the number of locations obtained by gazetteer tree and M is the number of metadata-levels extracted from the tweet object. For instance, we create a signal for *city:Valparaíso* and we find this hierarchy in *metadata:Tweet Text* and also in *metadata:User Location*. This means that we track the mention of *city:Valparaíso* at the level of the body of message and at the level of the user profile location individually. Furthermore, we create signals in the highest levels of the tree. Here, we create signals for *state:Valparaíso* and *country:Chile* at the level of the body of the message and the user profile respectively. An example of these signals is shown in Figure 4.7.

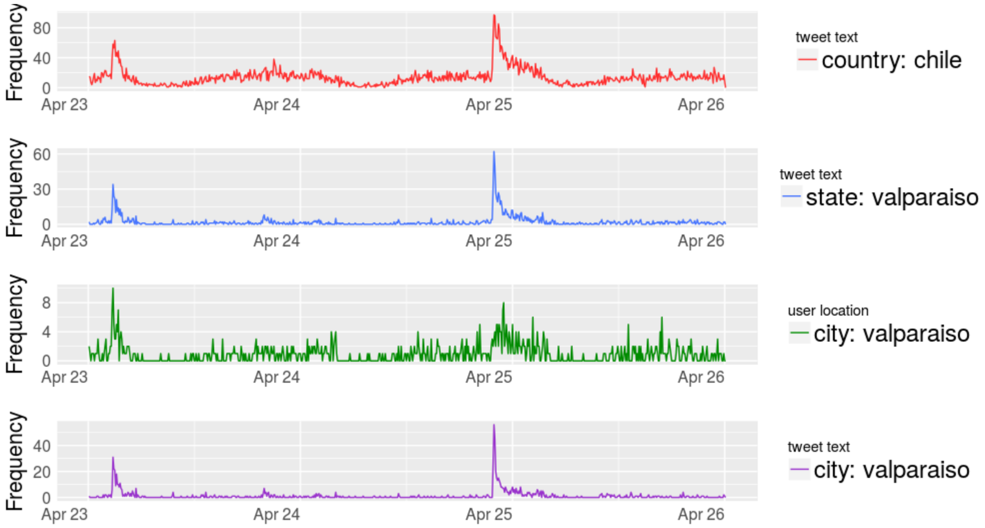


Figure 4.7: Example of signal creation using the frequency of each location and metadata level.

4.1.4 Time-Window

In this stage, we address the problem of how to divide and determine the time-window size to detect a new emergency situation. Additionally, we describe the features that are used in our methodology.

Determining Optimal Window Size

According to Guzman and Poblete [93]: "If the window size is too small, the occurrence of empty windows for a term increases, making the noise rate increase and frequency rate tend towards zero. On the other hand, if the window size is too large, the stability of the signal becomes constant and bursty keyword detection is delayed". Using this definition, we divide our signals into windows of six minutes because it divides a 24-hour day exactly, making the analysis easier to understand and to compare from different days.

Normalized Frequency

We compute the number of the messages of each time-window by signal. To normalize frequency, we compute *z-score* as following:

$$zscore = \frac{x_i - \mu_k}{\sigma_k} \tag{4.1}$$

where x_i is the frequency of the current i time-window, μ_k and σ_k are mean and standard deviation of the previous k time-windows respectively.

Interarrival Time

To characterize the urgency of the messages during a time-window, we compute the *interarrival time*, which is defined as $d_i = t_{i+1} - t_i$, where d_i denotes the difference between two consecutive social media messages i and $i+1$ that arrived in moments t_i and t_{i+1} respectively. Using this definition, which follows the work of Kalyanam et al. [111], high-activity events have a high-frequency in the first bins represented by values $d_i \approx 0$.

To quantify a high-frequency in very small values of d_i, we compute the measures *skewness* and *kurtosis*, which represent the asymmetry and the tailedness of the shape of probability distribution respectively [144]. Finally, we apply the equation 4.1 over *skewness* and *kurtosis* to calculate variation based on previous values.

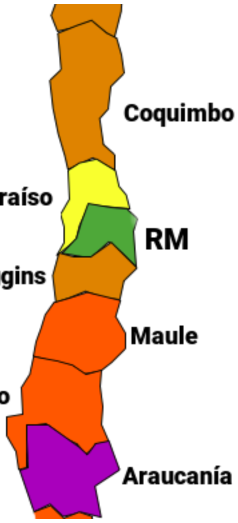
4.1.5 Geographic Spread

An emergency situation that affects and mobilizes a response in a small area is defined as *focalized*, while a disaster with a large geographic impact is defined as *diffused* [168]. Using this definition, we extend this concept to represent neighborhoods between locations obtained from section 4.1.3. For that purpose, we create an *adjacency matrix M*, where $M_{i,j} = 1$ represents if two locations are geographically connected or $M_{i,j} = 0$ if they are not connected. For instance, if an event is diffused (e.g., earthquake), the detection should be in adjacent-locations independent of metadata-level. On the other hand, if an event is focalized (e.g., terrorist attack), just one location should be detected, but in different metadata-levels simultaneously.

For example, using the administrative division (of a part) of Chile, we can construct the adjacency matrix based on the direct proximity between two locations in the country. The values of the main diagonal in the adjacency matrix are equal to 1 because the same location is connected to itself. As we explain above, $M_{i,j} = 1$ represents if two locations are geographically connected. If we look at the *Valparaíso* state on the map (Figure 4.8a), this location has three neighbors: *Coquimbo*, *RM* and *O'Higgins* state. Then, in our adjacency matrix (Figure 4.8b), we set the values $M_{valparaiso,coquimbo} = 1$, $M_{valparaiso,rm} = 1$ and $M_{valparaiso,ohiggins} = 1$ in those connected states. Otherwise, we set the value $M_{i,j} = 0$ in those non-connected states such as $M_{valparaiso,maule} = 0$, $M_{valparaiso,biobio} = 0$ and $M_{valparaiso,araucania} = 0$.

In our proposal, the geographic spread is quite relevant for removing false positive detections since an emergency situation does not have isolated locations when an event occurs.

Using Figure 4.8 that represents the Chilean administrative division, we can see that the Coquimbo and Valparaiso states are directly connected, but Araucania is not with them. In this direction, if we find detections just in these three states (Coquimbo, Valparaiso, and Araucania), we discard it as an emergency because natural or human-induced disasters generally reach a common area between neighboring locations. For instance, for events such as earthquakes or nuclear blasts - that impact several neighboring areas in a short or medium time - we expect that our method detects these crises because their locations are connected geographically. On the other hand, events such as soccer matches, music festivals, or political



(a)

	Coquimbo	Valparaíso	RM	O'Higgins	Maule	Bío Bío	Araucanía
Coquimbo	1	1	0	0	0	0	0
Valparaíso	1	1	1	1	0	0	0
RM	0	1	1	1	0	0	0
O'Higgins	0	1	1	1	1	0	0
Maule	0	0	0	1	1	1	0
Bío Bío	0	0	0	0	0	1	1
Araucanía	0	0	0	0	0	1	1

(b)

Figure 4.8: Example of a geographic spread. Left image (a) represents the administrative division for seven states of Chile. Right image (b) represents the adjacency matrix created for these states.

elections can generate detections in several states, but the scope can produce isolated or non-connected locations. However, one of the limitations may occur if two emergencies happen very close in time and are not directly connected geographically. This fact will trigger our method to discard these events as emergencies because, according to our adjacency matrix, they are isolated in space.

4.2 Methodology

According to the "data classification" module (Figure 4.1), we first trained a classifier to identify emergency events. Also, we introduced the hierarchy dependence to understand the local and national impact when an high-impact real-world event occurs. Besides, diffused and focalized events are identified with the goal of filtering false positives detections.

Our filtering task can be seen as binary classification task. The positive class (*detection label*) corresponds to messages related to instantaneous emergency situations, while the negative class (*nothing label*) corresponds to the remaining or non-related to crisis situations. Each row of our dataset is labeled as positive class when the event date occurs inside of the current time-window. However, in our work we used bags of tweets divided into time-windows with several features. We then labeled time-windows instead of tweets or messages.

We defined that a certain time-window contained an *event detection* if it had a positive variation in frequency, skewness and kurtosis with respect to the normalization of the previous values. Moreover and according to (Figure 4.2), we included the three following time-windows after the event to compensate for the imbalance between classes, given that after these number

Time-window metadata					**Attributes for classification**			
Ti	Tf	Hierarchy	Location	Metadata	Freq zscore	Skew zscore	Kurt zscore	Class
14:24:00	14:30:00	country	chile	tweet text	1.9969	0.4603	0.1252	True
14:24:00	14:30:00	country	chile	user location	1.3472	-0.06795	-0.3868	False
14:24:00	14:30:00	state	biobio	user location	0.6022	-0.3482	-0.9066	False
14:24:00	14:30:00	state	los lagos	user location	6.0913	2.7235	1.5000	True
14:24:00	14:30:00	state	metropolitana	user location	1.5681	-0.17024	-0.5626	False
14:30:00	14:36:00	country	chile	tweet text	4.1259	0.7296	0.6863	True
14:30:00	14:36:00	country	chile	user location	1.9969	0.4603	0.1252	True
14:30:00	14:36:00	country	chile	ttext-ulocation	14.1338	2.6002	3.6949	True
14:30:00	14:36:00	state	biobio	tweet text	0.4795	-0.4153	-0.6542	False
14:30:00	14:36:00	state	la araucania	user location	6.3866	8.3375	12.03857	True
14:36:00	14:42:00	country	chile	tweet text	6.3233	3.3888	6.5373	True
14:36:00	14:42:00	country	chile	user location	3.7502	-0.1372	-0.1893	False
14:36:00	14:42:00	country	chile	ttext-ulocation	9.0203	1.5187	1.2517	True
14:36:00	14:42:00	state	los rios	tweet text	14.6172	1.3849	2.7308	True
14:36:00	14:42:00	state	los lagos	user location	13.5506	1.7388	0.5650	True

Table 4.2: An example of the dataset generated by the creation of the signals and removing attributes after features selection. The table shows the time-windows metadata and the attributes for classification. Class true identifies an emergency situation and class negative does not.

of time-windows, the variation in the features decrease. Otherwise, the rows were labeled as negative class, meaning that they were not emergency events. Table 4.2 highlights an example of this dataset corresponding to the Chilean earthquake of December 2016.

To classify messages we employed traditional binary classifier Support Vector Machine *(SVM)*. As a result of the analyzed data scattering (Figure 4.9), we separated country and state in different datasets and set both kernels and classification parameters independently. On the one hand, *country* classifier uses a polynomial kernel and strict-parameters for gamma, cost and weights since a great number of messages are included in country hierarchy as an effect of the minor hierarchies. On the other hand, *state/region* classifier uses a linear kernel with soft-weights and cost.

Given that an emergency situation is not a usual event, we had a highly unbalanced data with respect to the classes after labeled ($1 \approx 2\%$ of positive class corresponding to *detection*). Therefore, we used *under-sampling* [139] over country and state datasets increasing our positive class to $15 \approx 18\%$. Additionally, to validate our model, we used *5-fold cross-validation* where one earthquake dataset is used as testing and the remaining earthquakes dataset as training.

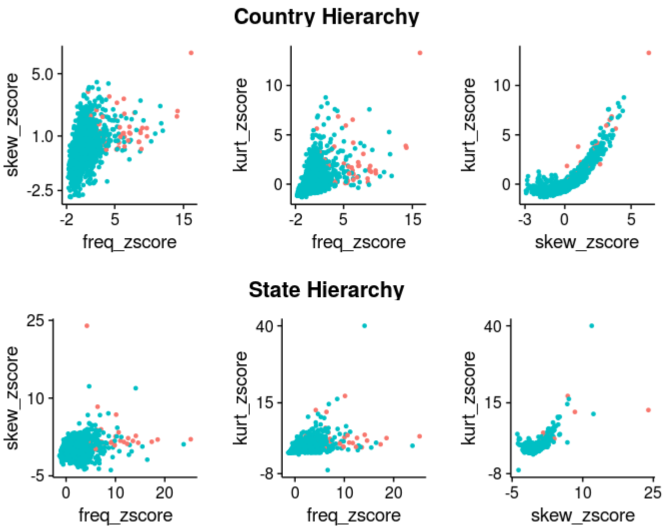


Figure 4.9: Relationship between features in country and state hierarchy. Red circles represent positive class (*detection*) and blue circles represent negative class (*nothing*).

4.2.1 Independent Analysis of Hierarchies

Our first analysis just considered the hierarchies as isolated detections. Figure 4.10 shows the results considering only the prediction over each instance in our datasets. As we noted in Table 4.2, each instance (or row) in our dataset is one specific hierarchy and metadata level with its corresponding attributes for classification.

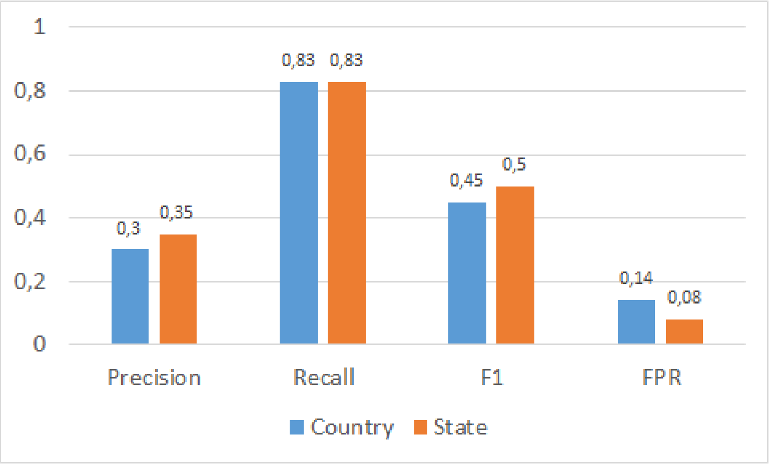


Figure 4.10: Average performance of 5-fold cross-validation by hierarchy independently just using labels.

As noted above, the assignment from the lowest level (city) to the highest (country) in the gazetteer hierarchy generated high frequency of messages, which caused multiple *bursts* in our country signal for non emergency situations. This concept can explain the values of Precision (P) and FPR in Figure 4.10.

In addition to the analysis of the number of detections by labels, we also studied the number of detections by time-window. For this analysis we aggregated the hierarchies by time-window and computed whether or not all instances were positives in the current time-window. This means that for each hierarchy in one specific time-window, we analyzed whether or not the classes were positives for each metadata-level. If all instances were positives in the time-window, the time-window was correctly assigned as positive.

According to the results shown in Figure 4.11, when we analyzed country and state independently, the values of Precision, F1 and FPR had worse values than the analysis by label. These results can be explained because we considered every location in the state hierarchy and aggregated them counting the number of positive classes. However, an emergency could not affect all locations in this hierarchy.

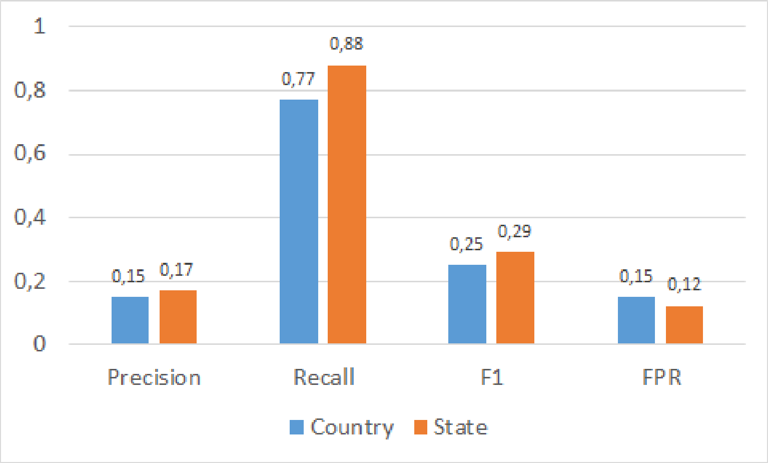


Figure 4.11: Average performance of 5-fold cross-validation by hierarchy independently just using time-windows.

4.2.2 Dependent Analysis of Hierarchies

Our second analysis considered the hierarchies as non-isolated detections. In the results explained above, we considered country and state hierarchy independently, which was not a correct analysis because an emergency situation affects states and country at the same time. For this reason, we inspected the time-windows where all metadata-level for country and state hierarchy had a correct detection simultaneously. As well as the independent analysis presented above, we aggregated the hierarchies by time-window and computed whether all instance are positives in the time-window. However, after classifying the hierarchies as a positive or negative class, in this evaluation we compared whether or not both hierarchies were positives simultaneously. We then determined whether the time-window corresponded to one detection. In our example, the state hierarchy had a false class and the country hierarchy had a true class. For this reason, the time-window was labeled as a false value.

The results are shown in Figure 4.12. In contrast to the independent analysis of country and state, we improved the Precision, F1 and FPR values as a consequence of smaller windows related to non-emergency situations that were assigned as detection. However, when we see the value obtained for FPR ($FPR = 0.03$), this rate represents an incorrect number of time-windows assigned as detection equal to 23. This means that we had 23 new emergency situations detected by our method.

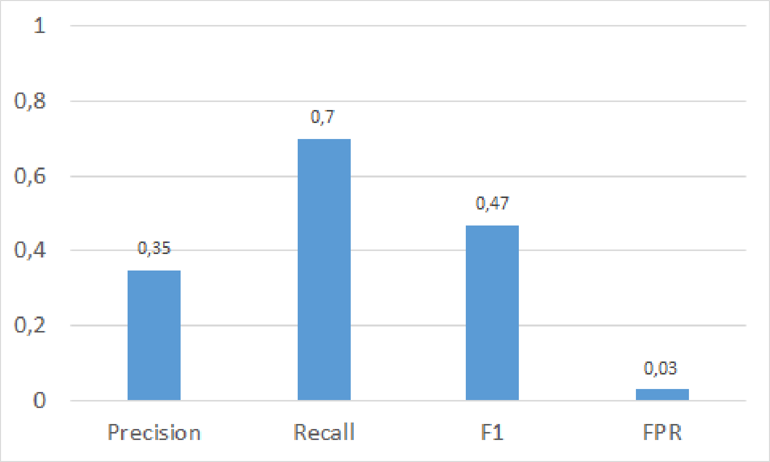


Figure 4.12: Average performance of 5-fold cross-validation by hierarchy dependently just using time-windows.

4.2.3 Geographic Spread Analysis

In addition to the results of the dependency analysis explained above, we saw that a large number of time-windows for country hierarchy ($\approx$ 82%) had more than one metadata-level when a correct detection existed. This can be explained since an emergency situation produces a collective reaction on the level of body of the message (tweet text), users sharing any messages with profile location in a specific country (*user location*) or mixing both concepts (*tweet text + user location*). For this reason, our third analysis considered the hierarchies as non-isolated detections and applies the Geographic Spread (G.S.). Using the *Adjacency Matrix* to represent neighborhoods between regions/states, we considered as a correct detection those time-windows where the state/s classified as detection were defined as *Focalized* or *Diffused* and exist dependency between hierarchies.

As well as the previous analyses, we grouped the hierarchies by time-window. However, before determining if all instances were positives, we applied two kinds of filters for state and country. For state hierarchy, we applied the geographic spread to filter non-isolated states when a detection was identified. Here, our filter determined if an event could be focalized or diffused. For country hierarchy, we considered a soft (two levels) or strict (three levels) evaluation related to the number of metadata-levels identified by time-window. The following steps were similar to previous evaluations.

Considering the geographic spread by states and the number of metadata-levels by country hierarchy, we analyzed the results shown in Figure 4.13. On the one hand, the *Country(2)+State+G.S.* represents the detection when we considered at least two metadata-levels (soft evaluation) for the country hierarchy and the geographic spread for states. In contrast to the previous analyses, we improved the values of the Precision, F1 and FPR. The last metric was very important because there were no time-windows incorrectly assigned as emergency situations.

Consequently, the Recall values decreased, which means that our method removed some time-windows classified as detection. Additionally, the percent of emergency situations detected was equal to 100% with an average delay equal to 10.4 minutes ($min = 6$, $max = 14$) from the impact of the event to the first detection.

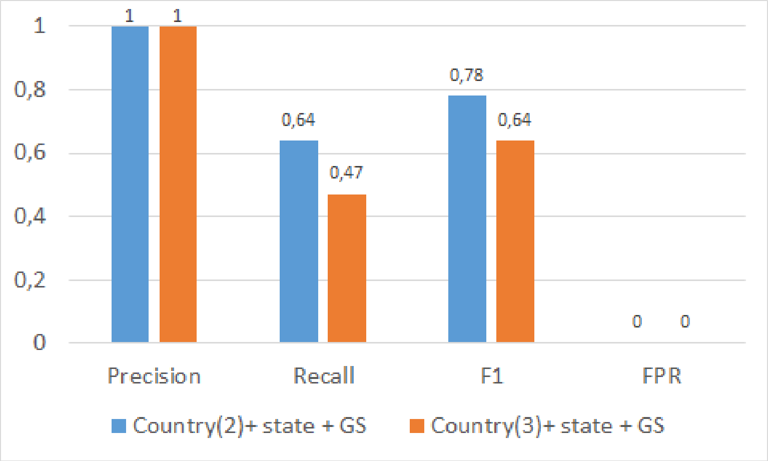


Figure 4.13: Average performance of 5-fold cross-validation by hierarchy dependently with geographic spread just using time-windows.

On the other hand, the *Country(3)+State+G.S.* (strict evaluation) represents the detection when we considered three metadata-levels for country hierarchy and the geographic spread for states. Similar to *Country(2)+State+G.S.*, we improved the values of Precision, F1 and FPR but our recall decreased from $R = 0.64$ to $R = 0.47$, detecting 80% of the emergency situations with an average delay equal to 11.5 minutes ($min = 8$, $max = 14$) from the impact of the event to the first detection.

4.2.4 Crisis Detections in the Wild

In this section, we evaluate our model in a realistic scenario using data from the Twitter Public Stream. The general idea is to determine the performance of detecting crises and filtering other high-impact real-world events that involve location references. For that purpose, we created a dataset formed by eight events that occurred in England between December 2016 and October 2017. To do this, we retrieved data considering the whole day they occurred. For instance, if Westminster terrorist attack happened on 22 March 2017, we will analyze data from 00:00 until 23:59 hours on 22 March 2017.

As can be noted in Table 4.3 and Table 4.4, we studied three terrorist attacks and five high-impact real-world events related to soccer matches, music concerts, and political elections. In the case of the terrorist attacks, we studied this type of event, since according to Carr [38], these crises are identified as *instantaneous-focalized* events, where an unexpected event affects the community but in a reduced area. In contrast to earthquakes (identified as *instantaneous-diffuse* events), we studied the ability of our classifier to detect other types of events where the number of affected people is smaller than in earthquakes, tsunamis, and other *instantaneous-focused* events. Additionally, for *Premier League Soccer Matches*, we cannot identify the

Table 4.3: Evaluation of events occurred in England by time-windows (T-W) using Country (2)+ State + G.S. method. The table shows the total number of detected time-windows, the number of detected time-windows before the beginning and after to the end of the event. The last two columns show the detection time delay with respect to the beginning of the event and the top 3 bigrams when the detection occurs.

Event	Detected T-W	T-W Before Event	T-W After Event	Delay (min)	Top 3 Bigrams
Premier League Soccer Matches	2	-	-	-	(man, utd), (new, year), (happy, new)
Westminster Terrorist Attack	13	0	13	32	(stay, safe), (terror, attack), (safe, everyone)
Manchester Terrorist Attack	12	1	11	23	(ariana, grande), (incident, arena), (grande, concert)
London Terrorist Attack	14	7	7	36	(stay, safe), (incident, bridge), (borough, market)
U.K. Elections	5	-	-	-	(theresa, may), (vote, labour), (van, dijk)
Adele Live in Wembley	9	7	2	-	(elland, road), (new, times), (phil, jackson)
England vs Slovenia Soccer Match	4	4	0	-	(simon, brodkin), (join, us), (theresa, may)
Metallica Live in London	4	4	0	-	(always, said), (chance, win), (carabao, cup)

beginning of the event since there are many soccer matches during the analyzed day. One observation about non-related-crisis events considered for this evaluation is that they can be spread over a more extended period because of news coverage, announcements, or user reactions before the event occurs (for instance, U.K. elections). Hence, although we analyzed the specific day when they happened, we are aware that they can also cover past and future days.

In our evaluation using data from the Twitter Public Stream, we trained a classifier with five earthquakes identified in our ground truth. Given that the primary language in England is English, we tested our method in a different primary language concerning the training data. This setting allows us to determine if our method has the characteristics of being language-agnostic and domain-independent. In other words, to determine if it can detect crises using data from a different domain and language of the messages.

As mentioned in Section 4.2.3, we obtained the best performance considering at least two (soft evaluation) and three (strict evaluation) metadata levels for the country hierarchy and the geographic spread for states. For this reason, our subsequent experiments will consider these settings to evaluate crisis detections in real scenarios. To know the most frequent terms when our method detects an event, we computed the Top 3 Bigrams in the detected time windows. Also, we calculated the time delay for emergency events from the beginning of the event until the first detection.

On the one hand, the first evaluation *Country(2)+State+G.S.* had full detection of the terrorist attacks with average time delay equal to 30.3 minutes. These detections are related to the event given that the bigrams represent terms associated with crisis situations. However, the *London Terrorist Attack* has 50% of the detected time-windows after the event, which means that there are seven time-windows non-related to emergency situations. Besides the crisis situations analysis, we also studied the number of detected time-windows in non-related to emergency situation events. In the same way, we had a large amount of misclassified time-windows that do not represent crisis situations, as we can see in the Top 3 Bigrams for each non-related to event.

Contrarily, the second evaluation *Country(3)+State+G.S.* decreases the number of non-

Table 4.4: Evaluation of events occurred in England by time-windows (T-W) using Country(3) + State + G.S. method. The table shows the total number of detected time-windows, the number of detected time-windows before the beginning and after the end of the event. The last two columns show the detection time delay with respect to the beginning of the event and the top 3 bigrams when the detection occurs.

Event	Detected T-W	T-W Before Event	T-W After Event	Delay (min)	Top 3 Bigrams
Premier League Soccer Matches	0	-	-	-	
Westminster Terrorist Attack	4	0	4	32	(terror, attack), (stay, safe), (terrorist, attack)
Manchester Terrorist Attack	2	0	2	23	(ariana, grande), (praying, everyone), (everyone, affected)
London Terrorist Attack	1	1	0	-	(ariana, grande), (around, world), (lady, gaga)
U.K. Elections	0	-	-	-	
Adele Live in Wembley	0	0	0	-	
England vs Slovenia Soccer Match	1	1	0	-	(per, day), (menswear, sample), (closed, roads)
Metallica Live in London	2	2	0	-	(happy, birthday), (chance, win), (always, said)

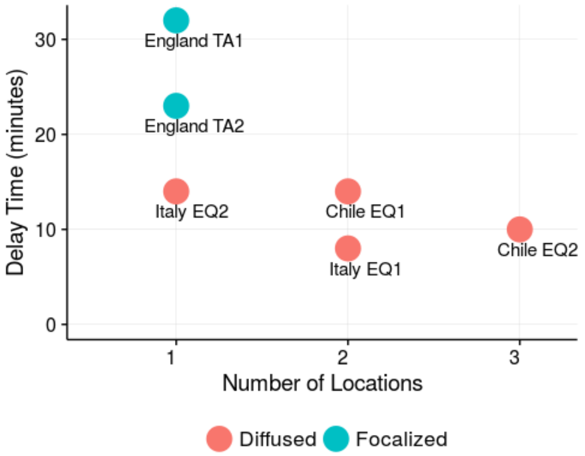


Figure 4.14: Relationship between time delay and number of locations in the first detection for diffused and focalized emergency situations.

related to emergency situations events detected as crisis situations. We can see three time-windows in two events detected as emergency situations (*England vs Slovenia*, and *Metallica Live in London*). In these cases, the time-windows are detected before the event and corresponding to non-emergency situations according to the bigrams. Furthermore, when we analyzed the number of the detected emergency situations, two-thirds (66%) of the events are detected correctly with average time delay equal to 30.3 minutes. In the case of *London Terrorist Attack*, our method detects one time-window before the event, but the bigrams describe that the detections do not correspond to crisis situations.

4.3 Discussion

Our findings suggest that there is evidence to detect an emergency situation based on anomaly frequency of messages that contain locations for a specific country. Indeed, our method based on the number of metadata-levels by country hierarchy and geographic spread by state, detects a 80% of the events related to emergency situations as we could demonstrate in our ground truth. Also, our method is independent of the textual features because we apply the model over various languages as Spanish, Italian and English. Furthermore, we test our model in various types of crisis events such as earthquakes (EQ) and terrorist attacks (TA), where these are identified in the literature as *instantaneous-diffused* and *instantaneous-focalized* events respectively. Also, we apply our methodology on different magnitudes (in the case of earthquakes) and number of affected people (e.g., *Manchester Terrorist Attack* vs *Westminster Terrorist Attack*).

However, when we apply our method in the on-line evaluation, we detect 66% of the emergency situations that affected England. This explains that the signals, and for various reasons, the number of active users in United Kingdom[9] which can affect the anomaly frequency of the messages since there exists a high daily average activity of the messages; similar locations in other countries (York ≈ New York); and soccer teams with names of cities (Manchester United, Liverpool). These issues also can affect the number of false positive detection, for which in the case of England was 30% of the non-related to emergency events.

Regarding the geographic spread where we define an emergency situation as diffused or focalized, we find some evidence that differentiates them. In the case of diffused events, the time delay of our first detection was less than 12 minutes and in focalized events was greater than 30 minutes (Figure 4.14). This explains that, in diffused events such as earthquakes, a high number of people are affected (thousands or millions) at the same time by an event, which generates a collective reaction in social media in the locations where the event impacted. In Figure 4.14, we can see that earthquakes have at least two detected locations in the first detection (except Italy EQ2). In contrast, focalized events have less eyewitnesses (hundreds or thousands) then when the users share messages in social media. The frequency does not affect the average daily message of the country in the first minutes. This is explained in Figure 4.14 where the terrorist attacks have just 1 detected location in the first detection.

Additionally, the time delay can be different for many reasons: datetime of the event (for example, during the early hours), few differences with the end of the current time-window, type of the affected locations (rural, urban cities) and the number of active users by locations.

[9] http://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/ visited on January 2018

Chapter 5

Cross-Lingual and Cross-Domain Crisis Message Classification

One of the main tasks related to the use of social media for disaster management is the automatic identification of crisis-related messages. Most of the studies in this topic have focused on the analysis of data for a particular type of event in a specific language. This limits the possibility of generalizing existing approaches because classification models cannot be used directly for new types of events or applied to other languages.

In this chapter, we study the task of automatically classifying messages related to crisis events by leveraging cross-language and cross-domain labeled data. The general objective is to address the cold-start problem by preventing it when no historical data exists to train supervised models that filter irrelevant information posted online during a specific crisis event. In this direction, we aim to use labeled data from high-resource languages to classify messages from other (low-resource) languages and/or new (previously unseen) types of crises. Therefore, we pose the research question: Can we leverage labeled data from other languages and domains to classify new, previously unseen events?

Our findings show that it is indeed possible to leverage data from crisis events in English to classify the same type of event in other languages such as Spanish and Italian (cross-lingual), with 81.5% and 80.0% F1-score respectively. In addition, we found we are able to achieve good performance for the task of classifying events from new types of crisis (cross-domain), reporting 82.5% and 77.5%. Overall, we show that it is possible to use data from high-resource languages to create effective models to classify crisis messages in other languages, preventing the *cold-start* problem. This indicates that there could exist underlying patterns in crisis situations that expand across different languages and crisis domains, which can be captured by multilingual representations.

The work presented in this section was published and co-guided as follows: *Sánchez Macías, C. M. (2021). Transfer learning for the multilingual and multi-domain classification of messages relating to crises. MSc. in Computer Science Thesis, Departament of Computer Science, University of Chile.*

5.1 Proposed Approach

We propose an experimental analysis of transfer learning approaches across several crisis domains and various languages[1]. The general idea is to evaluate the classification performance of several models by including data from other languages and domains. For our proposal and given the availability of public datasets - and their distribution across different languages - we choose those languages where the number of crisis domains (hazard types) allows us to study cross-lingual scenarios for a specific type of event. Therefore, we selected English, Spanish and Italian as languages for our analysis given the availability in several domains and number of messages for each (see Table 5.2 and Table 5.4). We detailed explain the dataset construction and their characteristics in Section 5.2.

In detail, we address the following specific objectives: 1) transferring knowledge from one or more crisis domains (e.g., *earthquakes*, *hurricanes*, etc.) to other crisis domains, and 2) transferring knowledge from crises in one language to another (low-resource) language.

Classifying microblog messages into the binary categories of *related* and *not related* to crisis events is useful, as it can help filter out irrelevant information and identify relevant messages that may be useful for crisis response efforts. In this direction, we focus on the task of classifying microblog messages about crisis events[2] into these binary categories. We adopt the definition used by Kruspe et al. [122] that considers a message as *related* to a crisis when it *refers implicitly or explicitly to the specific disaster event for which it was retrieved.* However, we generalize this definition to consider messages as *related* if they meet the above criteria for *any disaster*, not only a particular specified event. By adopting the definition used by Kruspe et al. [122] and generalizing it to consider messages as "related" if they meet the criteria for any disaster, the classification model can be applied to a wide range of crisis events (or domains) rather than being limited to a specific event. In addition, we define a *domain* as a *type of disaster (or hazard)* such as earthquakes or floods. Figure 5.1 provides an overview of the main steps followed in this work, both for dataset construction and experimentation. These steps are explained in the following sections.

5.1.1 Transfer-Learning Scenarios

We study 7 classification scenarios that span across several transfer-learning configurations. These scenarios aim to measure the impact on learning–of both domains and languages–for classifying new crisis events. We also include the scenario in which no domain nor language transfer learning is performed, which is the most traditional and simplest classification scenario. To avoid overfitting when the model is trained, we split training and testing data by ensuring that training data do not contain messages related to the event evaluated on testing. This means that messages related to a specific event are either used *training* or *for testing*, but not in both. We detail each scenario as follows:

[1] The work proposed in this section was developed in conjunction with Cinthia Sanchéz.

[2] We consider *an event as a unique occurrence in time and place*[148].

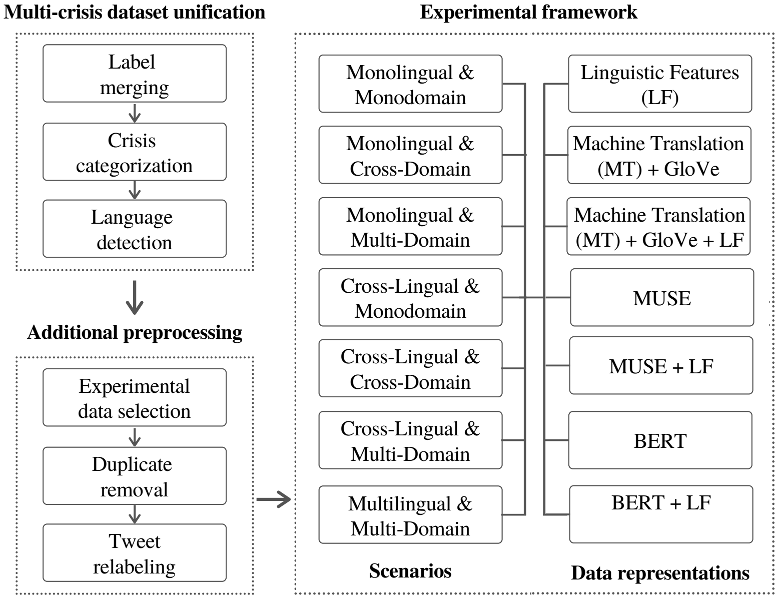

Multi-crisis dataset unification
Label merging
Crisis categorization
Language detection
Additional preprocessing
Experimental data selection
Duplicate removal
Tweet relabeling
Experimental framework
Monolingual & Monodomain
Monolingual & Cross-Domain
Monolingual & Multi-Domain
Cross-Lingual & Monodomain
Cross-Lingual & Cross-Domain
Cross-Lingual & Multi-Domain
Multilingual & Multi-Domain
Scenarios
Linguistic Features (LF)
Machine Translation (MT) + GloVe
Machine Translation (MT) + GloVe + LF
MUSE
MUSE + LF
BERT
BERT + LF
Data representations

Figure 5.1: An overview of our proposed methodology that described the unification of the multi-crisis dataset, additional data preprocessing for our experiments, and experimental framework including the transfer-learning scenarios and data representations.

- **Monolingual & Monodomain.** This scenario is the most straightforward and traditional classification task for classifying messages during crises. For this case, we train a classification model with messages from events of a specific crisis domain (e.g., earthquakes) all in the same language (e.g., English). We evaluate the model on messages from the same original domain and language, but from new events that were not used for training. To illustrate, we train a model with data corresponding to messages in English from a set of earthquake events. We then evaluate it on messages from other earthquakes in English that were not included in the original training set.

- **Monolingual & Cross-Domain.** In this scenario, we only perform domain transfer learning or domain adaptation within the same language. The objective is to evaluate the effect of training with data from past events of a specific domain and language (e.g., earthquakes in English) to classify messages from a different domain in the same language (e.g., explosions in English). This evaluation simulates the real-world scenario when a new, previously unseen type of disaster occurs and there is only labeled data from a kind of crisis and language.

- **Monolingual & Multi-Domain.** In this scenario, we perform data enrichment by augmenting the training data of a specific domain with that of other domains within the same language. This case is similar to the *monolingual & monodomain* scenario with the addition of more training data from other domains. To illustrate, we train a model with data from a set of events in English that includes all available domains, such as earthquakes, floods, and explosions. Then, we classify messages from a flood event in English that were not included in the original training set.

- **Cross-Lingual & Monodomain.** In this scenario, we perform cross-lingual adaptation within the same domain. The objective is to evaluate the effect of training a classification model with data from the target domain (e.g., floods) in a language (e.g., English) that is different from the target language (e.g., Italian). This simulates the real-world scenario where there is only labeled data from a high-resource language for a specific domain, and a new event occurs in the same domain but in another language.

- **Cross-Lingual & Cross-Domain.** In this scenario, we perform cross-lingual and domain adaptation. The objective is to evaluate the effect of training a classification model with data from a set of crisis domains (e.g., earthquakes and floods) in one language (e.g., English) to classify data of a different domain (e.g., explosions) in another language (e.g., Spanish). This simulates the case in which there is a need to classify messages of a new crisis domain in a language with no labeled data.

- **Cross-Lingual & Multi-Domain.** We perform cross-lingual adaptation and data enrichment using additional domains in this scenario. The objective is to evaluate the effect of training a model using data of the target domain and that of other (multiple) domains, all in the same language, then, to classify data of the target domain in another language. To illustrate, we train a model with data in English about floods, earthquakes, and hurricanes to classify flood-related messages in Italian. This simulates the case in which we use all the available data in a high-resource language to classify messages of a known domain but in another language.

- **Multilingual & Multi-domain.** In this scenario, we perform cross-lingual and domain data enrichment. This case is similar to the *cross-lingual & cross-domain* scenario with the addition of data from the target domain and language. Thus, providing a warm-start type of setting. To illustrate, we train a model with English and Italian data about floods, earthquakes, and hurricanes to classify flood-related messages in Italian.

5.1.2 Data Representations

We focus on message representations and state-of-the-art models that allow us to convey multilingual data in a single feature space. Our representations correspond to seven approaches described below.

- **Linguistic Features (LF).** We consider this model as the baseline. This representation models each message as a set of linguistic features. We considered 48 features[3] represented in numerical and binary form. Some of these features have been previously used for classification of crisis messages [83; 116]. These features describe traditional message characteristics such as the number of *characters*, *words*, *links*, *mentions*, *hashtags*, *question marks*, among others. Furthermore, we consider attributes that are specific to each language, including *sentiment polarity*, *Named Entities* and *Part-of-Speech (POS)* [8; 7]. We also included binary features such as *has mention* and *has location.*

- **Machine Translation (MT) + GloVe.** This approach is based on modeling messages using English as a pivot language. Hence it implies translating all messages that

[3] The full description of the features can be found in our repository.

were not written in the pivot language to English[4]. We choose English as our starting point for two reasons: 1) the high-quality of pre-trained embeddings in English, and 2) the trade-off of translating other languages to English, which is predominant in our dataset, as opposed to the other way around. Text in the pivot language is then tokenized and vectorized using a pre-trained GloVe[5] model with 100-dimensions, which was trained on tweets [176]. To represent each message, we combined single word embedding using mean aggregation. For words Out-Of-Vocabulary, we created vectors with zero.

- **Machine Translation + GloVe + Linguistic Feature (MT+GloVe+LF).** This data representation models each messages as the combination of the previous model (Machine Translation + GloVe) and LF features. The goal is to evaluate if both types of features (semantic and statistical) combined improve the performance of the models that use each type of feature separately.

- **MUSE (MUSE).** This representation is based on modeling each message as the result of its vectorization using MUSE[6]. MUSE are multilingual language-aligned word embeddings of 300-dimensions, based on fastText embeddings trained on Wikipedia [48]. As in the previous model, we combined single-word embedding using mean aggregation and created vector representations with zero for words Out-Of-Vocabulary.

- **MUSE + Linguistic Features (MUSE+LF).** This data representation models each message as the combination of MUSE and LF features.

- **BERT (BERT).** This representation is based on modeling each message as the result of its vectorization using BERT-Base Multilingual Cased model[7] with no fine-tuning. This model of 768-dimensions was trained using the top-104 languages in Wikipedia [56]. We combined the contextualized word embeddings using mean aggregation, ignoring the padding of zeros.

- **BERT + Linguistic Features (BERT+LF).** This data representation models each message as the combination of MUSE and LF features.

5.2 Unified Dataset Construction

We create a new dataset that is the result of collecting 7 crisis datasets from the literature (see Table 5.1 for details). These datasets met the criteria of having labeled data and being publicly available. The consolidated dataset, which we refer to as **Multi-Crisis Dataset**[8], contains Twitter messages from several crisis domains and events in various languages. To unify labels and to achieve an enriched, consistent dataset, we performed a process that consisted of several data curation steps. These included text preprocessing and cleaning, duplicate removal, label merging, crisis categorization, and language detection. Next, we

[4] We use the Google Translate API: `https://cloud.google.com/translate/`

[5] `https://nlp.stanford.edu/projects/glove/`

[6] `https://github.com/facebookresearch/MUSE`

[7] `https://github.com/google-research/bert`

[8] This dataset can be found at `https://github.com/cinthiasanchez/Crisis-Classification/`

Dataset	Lang.	Tweets	Events
CrisisLexT6 [167]	English	60,082	6
CrisisLexT26 [168]	Multiple	27,933	26
CrisisNLP_R1 [105]	English, Spanish, French	49,596	14
Ecuador-Earthquake [209]	English, Spanish	8,360	1
SoSItalyT4 [51]	Italian	5,642	4
ChileEarthquakeT1 [45]	Spanish	2,187	1
CrisisMMD [10]	English	11,400	7
Multi-Crisis Dataset (Ours)	**Multiple**	**164,625**	**53**

Table 5.1: Datasets used to create our Unified Multi-Crisis Dataset.

provide relevant details of this process, but more in-depth information can be found in our public repository.

5.2.1 Label Merging

The original datasets had several labels (binary and categorical), with different degrees of generalization. We decided to merge these labels into their more general form to maintain consistency. We unified labels by their relatedness to crisis events (*related* and *not related* messages), which are used in the *CrisisLexT6, ChileEarthquakeT1* and *Ecuador-Earthquake* datasets We choose these labels given that during a real crisis event, the recognition of *related* and *not related* messages is one of the first filterings by emergency practitioners using social media data [168]. In fact, recent work in the field of crisis informatics has addressed the unification of labels from various datasets because the discrepancy in the class labels used across datasets [192; 12; 116]. Following the mentioned approaches for label unification, we mapped to the *not related* category all of the labels in the original datasets that were coherent with this criteria, such as *"not related"*, *"not relevant"*, *"not related or irrelevant"*, *"off-topic"*, *"not informative"*, *"not applicable"*, and *"not physical landslide"*. In addition, we considered as *related* the following labels from the original datasets: *"related"*, *"relevant"*, *"damage"*, *"no damage"*, *"on-topic"*, *"informative"*, *"related and informative"*, *"related but not informative"*, among others.

5.2.2 Crisis Categorization

We annotated each message according to the crisis dimensions of the event to which it belongs to. We use a similar pipeline to Olteanu et al. [168]. We categorized crises by hazard types (such as *"earthquake"* or *"explosion"*), hazard categories (*"natural"* or *"human-induced"*), sub-categories (e.g., *"geophysical"*, *"hydrological"*, *"accidental"*, etc.), temporal development (*"instantaneous"* or *"progressive"*), and geographic spread (*"focalized"* or *"diffused"*). We decided to include this information because it will allow the study of communication patterns along several dimensions, for example, to study the similarities and differences in communication during natural events versus human induced events [192]. Furthermore, we aggregated information about the crisis, such as country and year.

5.2.3 Unified Dataset Description

Our dataset is mainly composed of messages in English (83.67%), followed by Spanish (7.30%) and Italian (4.25%). With regard to the domain, it is mainly composed of earthquake (25.47%), flood (19.39%) and hurricane (11.89%) domains. Tables 5.2 and 6.3 show a complete description of the number of messages per language and domain.

Table 5.2: Number of messages and percentages by language

Lang.	**Count**	%
English	137,743	83.67
Spanish	12,025	7.30
Italian	7,002	4.25
French	1,144	0.70
Portuguese	771	0.47
Tagalog	502	0.31
Russian	238	0.15
German	124	0.08
Indonesian	111	0.07
Dutch	101	0.06
Others	4,864	2.96
Total	**164,625**	**100.00**

Table 5.3: Number of messages and percentages by domain

Domain	**Count**	%
Earthquake	41,931	25.47
Flood	31,923	19.39
Hurricane	19,578	11.89
Typhoon	13,674	8.31
Explosion	12,004	7.29
Bombings	11,012	6.69
Tornado	9,992	6.07
Landslide	4,492	2.73
Wildfires	3,533	2.15
Viral disease	3,512	2.13
Others	12,974	7.88
Total	**164,625**	**100.00**

5.3 Experimental Setup

In this section we describe our evaluation subset of data and detail the experimental setup. Our setup is designed with the purpose of evaluating the proposed transfer-learning scenarios.

This includes the evaluation of the effectiveness of several data representations studied as well.

For our experiments we work with a portion of the Multi-Crisis Dataset. Specifically, we considered the languages, hazard domains and events that allowed us to have sufficient data for evaluating several scenarios. We selected messages from the top-3 most represented languages in the dataset (English, Spanish and Italian).

Following the methodology presented by Alam et al. [12], we removed duplicates and near-duplicates to avoid having very similar messages in the train and test sets, thus preventing overestimated results. In addition, we discarded messages corresponding to events that 1) contained very little data, or 2) were from hazard domains not available for more than one language.

Finally, our experimental dataset contained 67,001 tweets from various regions in 3 languages and from 3 hazard domains: earthquakes (46.9%), floods (38.4%) and explosions (14.7%). Additionally, this collection contains 80.0% of English messages, 11.3% in Spanish, and 8.7% in Italian. Regarding the label distribution, 36.0% were categorized as *not-related* to crisis while 64.0% were labeled as *related*, representing a significant imbalance between classes.

5.3.1 Model Training

We partitioned training and testing data based on the events that they belong to. This is, for every evaluation scenario, each event (and their messages) was distributed either into train or test.

To provide more representative examples for generalization in our model, we selected our training sets by prioritizing the inclusion of events with the highest and most balanced number of instances in our dataset. In the *monolingual & monodomain scenario*, for English earthquake-related messages, we trained our models with the *Nepal* crisis; and for flood-related messages, we trained using the *Alberta* flood. For messages in Spanish about earthquakes, we trained with *Chile 2010* and *Chile 2014* events. Finally, for Italian earthquakes, we trained the models with the *Emilia* earthquake. For the other scenarios, we split the training and testing data by hazard type and language, excluding the target crisis events from the training set when it is necessary.

After consolidating our training set, we noted an imbalance between the positive *(related)* and negative *(not related)* classes, where the positive was approximately 27% more represented than the negative class. Hence, we applied both random subsampling and oversampling to reach a ratio between classes of 50%−50%. Tables 5.4 and 5.5 show the number of balanced instances that were used for training classification models in each scenario, grouped by target language and domain.

Lang.	Domain	Monolingual-Monodomain	Monolingual-Cross-Domain	Monolingual-Multi-Domain
English	Earthquake	11,214	35,720	46,238
	Explosion	-	51,662	-
	Flood	10,346	34,418	46,916
Spanish	Earthquake	2,822	150	-
	Explosion	-	4,182	-
Italian	Earthquake	1,520	414	-
	Flood	-	2,114	-

Table 5.4: Number of training instances by target (language and domain) used in the monolingual scenarios, considering the balanced sets. The symbol "-" means that no experiments were performed.

Lang.	Domain	Cross-Lingual-Monodomain	Cross-Lingual-Cross-Domain	Cross-Lingual-Multi-Domain	Multilingual-Multi-Domain
Spanish	Earthquake	18,324	35,720	56,346	67,260
	Explosion	9,282	51,662	63,460	69,768
Italian	Earthquake	20,426	35,720	63,122	67,520
	Flood	26,674	34,418	63,244	67,686

Table 5.5: Number of training instances by target (language and domain) used in the cross-lingual and multilingual scenarios, considering the balanced sets.

5.3.2 Model Testing

For testing classification models, we used the messages of the remaining events for each corresponding scenario that were not used for training. For instance, to evaluate the classification of earthquakes in Spanish in the *monolingual & monodomain scenario* (five events in total), we trained our model with *Chile 2010* and *Chile 2014*, and evaluated it with the remaining events (from *Ecuador 2016, Guatemala 2012* and *Costa Rica 2012*). Furthermore, we maintained the target set for all scenarios.

Table 5.6 shows the number of messages per target detailed by class, which shows an important imbalance predominated by the positive class. In a more realistic scenario when an emergency occurs, content shared in online platforms is generally associated with unrelated, irrelevant and noisy information about the event. With this in mind, as well as making the interpretation of our results easier, we augmented the negative class to create a more balanced testing set (i.e., close to 50%-50%). As negative instances are by definition messages not related to any crises, this does not affect evaluation. In this direction, we included a random sample of negative examples, which had not been used for training, from other domains in the target language. When negative examples from the target language were insufficient to balance classes, we further augmented this data by including translated negative examples from English using Google Translate.

Lang.	Domain	Related	Not related
English	Earthquake	8,611	1,225
	Explosion	4,415	4,641
	Flood	8,272	5,747
Spanish	Earthquake	2,507	453
	Explosion	747	50
Italian	Earthquake	698	198
	Flood	1,759	138

Table 5.6: Original data available for testing for each target (before augmentation).

5.4 Results

Given the multiple combinations of scenarios, domains, and languages, we summarize the results of our evaluations explained in our experimental setup. However, our repository contains an exhaustive evaluation of all possible scenarios[9].

Next, we describe the results obtained by target language and domain. We then conclude by presenting the aggregated results by scenario for low-resource languages.

5.4.1 English Classification

We evaluated the classification of crisis related messages in English. Specifically, monolingual scenarios for the same domain, cross-domain (domain adaptation) and multiple domains (data enrichment). We did not evaluate cross lingual scenarios with English as a target language as we consider it as a high-resource language in our setup. However, we performed cross-lingual adaptation for low-resource languages using English as a source, detailed in the following sections.

Figure 5.2 shows the results obtained per scenario for each data representation and each of the three domains: earthquakes, explosions and floods. The overall results show that models achieved their best performance by using MUSE+LF and MT+GloVe+LF, while using LF and BERT did not work as well in this case (see Table 5.7 for more details). Additionally, we noted that the performance of MT+GloVe was slightly lower than including linguistic features (LF) for this representation.

The best performing model for earthquake message classification (evaluation included messages from *Chile 2014, California, Pakistan* and *Ecuador* events) was 87% F1-score in

[9] `https://github.com/cinthiasanchez/Crisis-Classification/results/balanced`

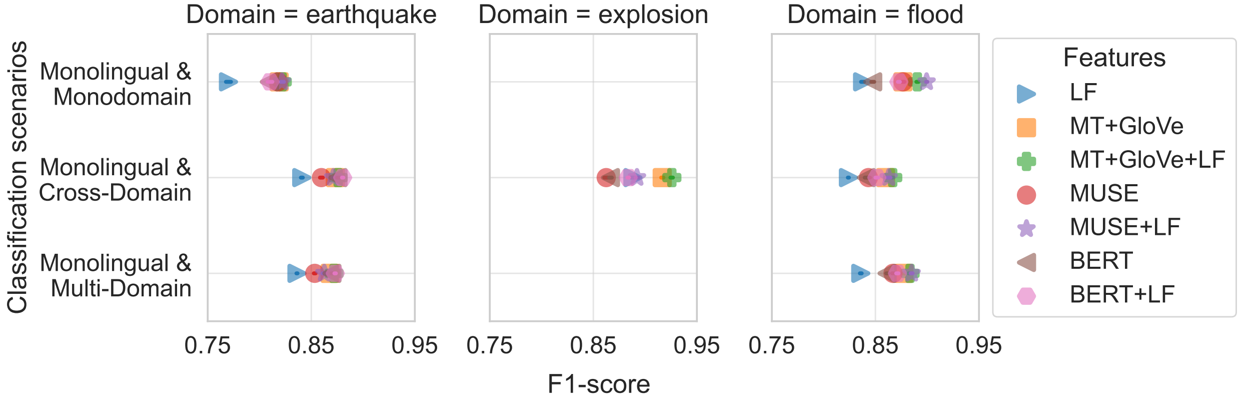


Figure 5.2: Data representation performance per scenario for English message classification.

Domain	**LF**	**MT+GloVe**	**MT+GloVe+LF**	**MUSE**	**MUSE+LF**	**BERT**	**BERT+LF**
Explosion	0.8882	0.9165	0.9258	0.8624	0.8921	0.8665	0.8836
Flood	0.8325	0.8713	0.8802	0.8623	0.8824	0.8500	0.8651
Earthquake	0.8157	0.8535	0.8566	0.8437	0.8518	0.8517	0.8546

Table 5.7: Average results per domain and data representations for English messages classification.

the cross-domain and multi-domain scenarios. In the cross-domain scenario, we developed domain adaptation from explosion and flood domains; and in the multi-domain scenario, we considered data augmentation, or enrichment, by adding multiple domains to the target domain's training data (i.e., earthquake, explosion, and flood). In particular, we observe that both of these scenarios average a 5% improvement over the simplest scenario (monolingual & monodomain).

For explosion message classification, our dataset only contained one event (*West Texas* event). Using this event as target we evaluated the only possible scenario, i.e., domain adaptation or cross-domain (training was done with earthquakes and floods), yielding an F1-score of 93% with the best data representation (MT+Glove+LF). Recall that this scenario simulates the case in which we need to classify events from a new type of unseen crisis event.

For flood message classification (evaluation included events of *Queensland, Pakistan* and *India*), the best scenario was the *monolingual & monodomain* scenario with an F1-score of 90% using MUSE+LF. Although, cross-domain (86% F1) shows good performance, similar to earthquake classification. On the other hand, multi-domain (88% F1) or domain enrichment does not improve and slightly worsens classification.

Overall, we observe that domain adaptation works well within a high-resource language. This would allow us to use past knowledge to classify new and unexpected events in the same language. Therefore, it indicates one could use a pre-trained classifier to detect new types of crises that emerge. In addition, data augmentation (multi-domain) by including data from other domains can potentially improve model performance or, in the worst case, perform similarly to the *monolingual & monodomain scenario.*

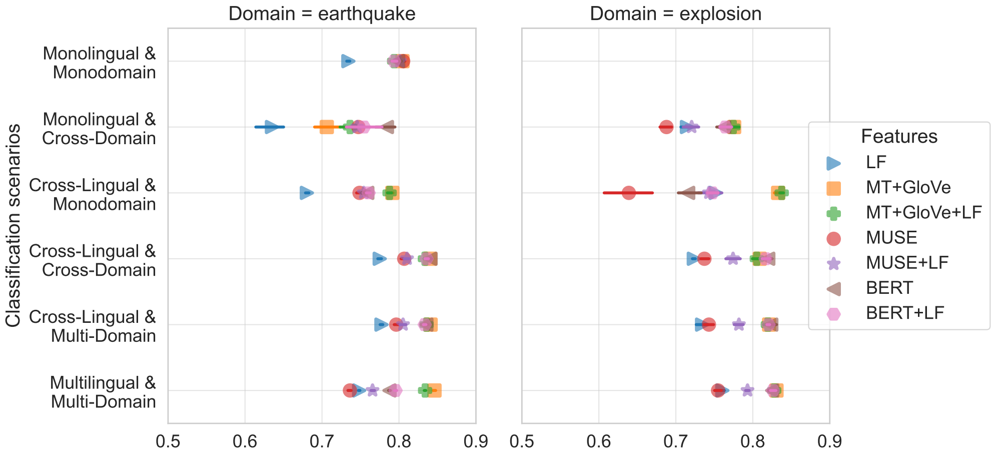


Figure 5.3: Performance per scenario for Spanish message classification.

Domain	LF	MT+GloVe	MT+GloVe+LF	MUSE	MUSE+LF	BERT	BERT+LF
Explosion	0.7368	0.8139	0.8130	0.7124	0.7626	0.7896	0.7951
Earthquake	0.7250	0.8050	0.8036	0.7735	0.7804	0.8011	0.7956

Table 5.8: Average results per domain and data representations for Spanish messages classification.

5.4.2 Spanish Classification

We consider this language as low-resource, since the amount of labeled data is significantly less than for English (see Table 5.2). Our experiments include monolingual, cross-lingual and multilingual scenarios.

Figure 5.3 presents the results for earthquake and explosion domains, broken down by the data representations. Overall, MT+GloVe obtained the best results, followed by MT+GloVe+LF and BERT. While LF and MUSE did not perform as well.

In the case of earthquake messages classification, we obtained the best performance scenarios with an 84% F1-score (on events from *Ecuador, Guatemala* and *Costa Rica*) using the MT+GloVe feature. This result overcame the *monolingual & monodomain* scenario on average by 3%. We observed this improvement for the cross-lingual and multilingual scenarios that used some sort of cross-language adaptation from English.

In particular, the *multilingual & multi-domain scenario*, which simulates classifying messages of a new type of crisis domain in a new language, already reaches the highest performance. Hence, there does not appear to be additional improvement when including data from the target domain and from the target language. In addition, the worst performance was for the *monolingual & cross-domain* scenario, which simulates the case when we attempt to classify messages from a new domain in the same language. This most likely occurs due to the small amount of training data in Spanish, which limits cross-domain learning within that language. Furthermore, we noted that the *monolingual & monodomain* scenario obtained lower performance than the other scenarios (which consider adaptation from other languages

Domain	**LF**	**MT+GloVe**	**MT+GloVe+LF**	**MUSE**	**MUSE+LF**	**BERT**	**BERT+LF**
Flood	0.7555	0.7791	0.7910	0.7463	0.7963	0.6719	0.7390
Earthquake	0.6343	0.7634	0.7533	0.7769	0.7624	0.7257	0.7300

Table 5.9: Average results per domain and data representations for Italian messages classification.

or domains). One possible reason for this result could be specific differences in Spanish languages and slang across the events that occurred in Spanish-speaking countries. This phenomenon may be explained because we trained the model on data related to earthquakes in Chile and tested it on events in Ecuador, Guatemala, and Costa Rica.

For the explosion message classification in Spanish, we only have one event worth of data (an event occurred in *Venezuela*). Therefore, we were not able to evaluate the *monolingual & monodomain* scenario. We observed that the best performance scenario, with 84% of F1-score, was the *cross-lingual & monodomain* scenario using MT+GloVe+LF as features and data in English as the source.

As with earthquakes, the worst performance scenario was training with another domain in the same language as the target (i.e., monolingual & cross-domain). The worst performance feature along all scenarios is MUSE.

In general, our results indicate that the performance of the *monolingual & monodomain scenario* can be improved by augmenting a low-resource language with high-resource language data, including multiple domains.

5.4.3 Italian Classification

With Italian, we perform a similar evaluation as we do with Spanish. Figure 5.4 presents the classification results for earthquake and flood domains for each scenario and feature. We observed that MUSE+LF representations performed better overall in Italian message classification. The other representations show behavior that varies according to the target domain. For example, LF features obtained the worst performance classifying earthquakes, and MUSE features obtained the best performance. However, such behavior is different when classifying floods.

For earthquakes the best performance (on *L'Aquila* event) was achieved for the *cross-lingual & multi-domain scenario* (80% F1) using MUSE and for floods (on *Sardinia* and *Genova* events) it was in the cross-lingual monodomain scenario (83% F1) using MUSE+LF.

In general, the cross-lingual adaptation and augmentation using English (a high-resource language) improved performance. Also, adding multiple domains from English and the target language increased this improvement.

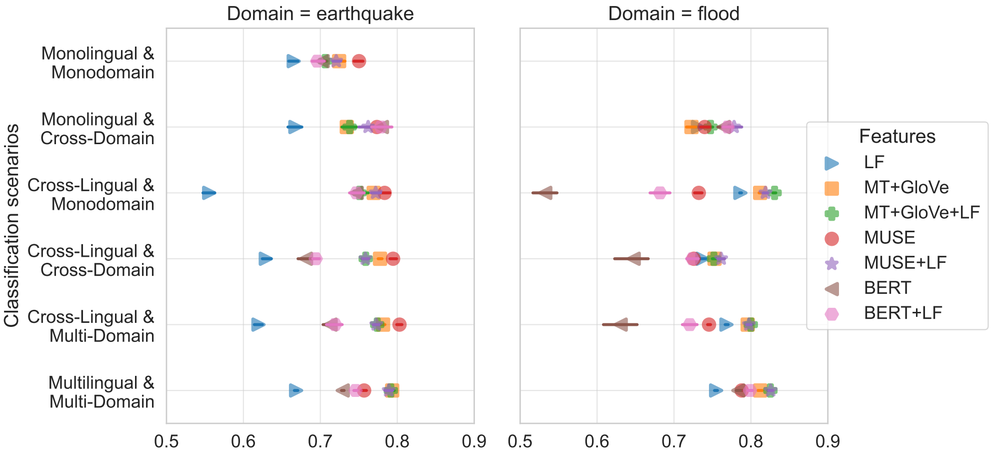


Figure 5.4: Performance per scenario for Italian message classification.

Scenarios	**LF**	**MT+GloVe**	**MT+GloVe+LF**	**MUSE**	**MUSE+LF**	**BERT**	**BERT+LF**
Monolingual & Cross-Domain	0.687	0.735	0.749	0.737	**0.773**	0.752	0.767
Cross-Lingual & Monodomain	0.693	0.801	**0.802**	0.726	0.772	0.689	0.734
Cross-Lingual & Cross-Domain	0.716	**0.795**	0.787	0.766	0.776	0.748	0.768
Cross-Lingual & Multi-Domain	0.725	**0.81**	0.807	0.772	0.789	0.751	0.773
Multilingual & Multi-Domain	0.733	**0.821**	0.82	0.759	0.793	0.781	0.792

Table 5.10: Average F1-score by classification scenario and feature for the low-resource languages Spanish and Italian, including their crisis domains. The test examples are the same in all scenarios. The best result per scenario is highlighted in bold.

5.4.4 Results by classification scenario

We present the aggregated results grouped by scenario and detailed by features. To identify the best classification approach for low-resource languages, Table 5.10 shows the average F1-score by classification scenario and feature. This result considers the crisis domains of Spanish and Italian targets such as earthquake in Spanish, explosion in Spanish, earthquake in Italian, and flood in Italian.

Regarding the classification scenarios, we achieved the best performances in the *multilingual & multi-domain* and *cross-lingual & multi-domain* scenarios. On the contrary, we obtained the lowest performance in the *monolingual & cross-domain* scenario. As for the data representations, our results show that the baseline feature (LF) presents the lowest score in most scenarios, while MT+GloVe+LF obtained the best overall performance across scenarios, followed by MT+GloVe, which is slightly lower than the former in less than 1%. Additionally, MUSE+LF obtained a competitive result, which is 1% lower than the GloVe representations mentioned above.

5.5 Discussion

Our findings indicate that *multilingual & multi-domain* adaptation is an effective way to improve the classification of low-resource languages. When no labeled data is available for a target language, a good option is to perform cross-lingual adaptation from a high-resource language using all available domain data. Most importantly, we show that it is possible to classify messages that correspond to a new, previously unseen, crisis even when they are in an unknown language. This can be very useful to identify unexpected emerging crisis situations for early response.

In addition to the availability of no labeled data for a target language, we noted that this effect was also observed when we incorporated data from other domains into the target event. In most of the experiments across the scenarios, we noticed that the performance increased by about 5% by integrating data from other domains, for instance, in the case of *cross-domain* or *multi-domain* scenarios.

A comparison of cross-domain results for Spanish and Italian reveals that transferring knowledge from one or more domains to another is useful from high-resource languages, such as English. Nevertheless, it does not help if we use as a source low-resource languages (e.g., training with Spanish messages about explosions to classify earthquake messages in the same language).

Regarding the most effective ways to represent data for knowledge transfer, translating the content into English (MT+GloVe) provides the most accurate results for Spanish messages. However, for Italian, MUSE+LF provides the best results. This can be due to variability in the quality of machine translation, for example. In practice, translating messages to English may not be cost-effective. However, this analysis show us the available alternatives. In terms of F1-score, the observed difference between MT+GloVe and MUSE+LF models is not statistically significant at a 95% confidence level.

We also observe that the LF representation was not as competitive across scenarios. However, when this feature is combined with MUSE, it improves results, specifically for Italian. Regarding BERT, we show that it provides competitive results for some classification scenarios, but this is not consistent.

For explosion in Spanish and flood in Italian, we observed a decrease in performance for the *cross-lingual & monodomain* scenario, using the MUSE and BERT compared to LF. This could be due to: 1) the dependence of specific words to the crisis domain and 2) the similarity of the representations of those words in both languages (English and the target language). The latter could be explained by the general corpus that was used to train MUSE and BERT (Wikipedia).

Finally, the similarity of studied languages can aid in classification results. For example, even though English belongs to the Germanic language family, it shares a significant number of cognate words with Spanish and Italian, which are Romance languages, amounting to around 30% to 40% [47]. This similarity is attributed to the influence of Latin, which has a shared origin with English, Spanish, and Italian [218; 141]. Therefore, selecting these languages based on the availability of data and public sources can have a positive impact on classification performance in cross-lingual and multi-lingual scenarios. However, the suitability of English as a high-resource language for other language families, such as Slavic or Indo-Iranian, is uncertain due to their significant grammatical and syntactical differences.

Chapter 6

A Comparative Study of Communication Patterns Across Crisis Events

Valuable and timely information about crisis situations such as natural disasters, can be rapidly obtained from user-generated content in social media. This has created an emergent research field that has focused mostly on the problem of filtering and classifying potentially relevant messages during emergency situations. However, we believe important insight can be gained from studying online communications during disasters at a more comprehensive level. In this sense, a high-level analysis could allow us to understand if there are collective patterns associated to certain characteristics of events.

This chapter presents a large-scale comparative analysis of 41 real-world crisis events. Our study addressed the challenge of understanding general patterns of crisis communication in a domain-independent form. In other words, instead of analyzing specific events as case studies, we comprehend differences and similarities across multiple crises, which differ in time, location, and hazard categories/subcategories.

The research presented in this chapter focuses on English messages, given the availability of public datasets for various hazard dimensions. Furthermore, our literature review presented in Chapter 3.1.1 pointed out that previous studies used pre-defined labels to differentiate hazard dimensions without extracting characteristics for the messages or finding differences between relevant and irrelevant content. Hence, we analyze relevant (Twitter) messages by extracting their textual and linguistic features to identify differences and similarities across crises automatically.

For our comparison, we considered hazard categories (i.e., human-induced and natural crises) and subcategories (i.e., intentional, accidental, and so forth). Our results show that using only a small set of textual features, we can differentiate among types of events with 75% accuracy. Indicating that there are clear patterns in how people react to different extreme situations, depending on, for example, whether the event was triggered by natural causes or human action. These findings have implications from a crisis response perspective, as they will allow experts to foresee patterns in emerging situations, even if there is no prior experience with an event of such characteristics.

The work presented in this section was published as follows: *Sarmiento, H., & Poblete, B. (2021, March). Crisis communication: a comparative study of communication patterns across crisis events in social media. In Proceedings of the 36th Annual ACM Symposium on Applied Computing (pp. 1711-1720).*

6.1 Crisis Event Dataset Creation

For this analysis we study collective social media reactions to natural and human-induced events as well as their hazard subcategories, such as intentional and accidental crises, among others. Our study focuses on linguistic aspects such as affective, cognitive and social processes. To achieve this, we retrieved, cleaned, and unified several publicly available social media crisis datasets into a large collection. These existing datasets had already been labeled according to diverse criteria by their authors, which required us to work towards consolidating these annotations. In addition, we enhanced existing labels with metadata that describes an event (e.g., geographical location, hazard category, and subcategory, temporal development, geographic spread, and language). In the remaining pages of this section we detail relevant aspects of this process.

Table 6.1: List of six datasets and their characteristics considered in our consolidated collection.

Source	Labels	Events	Size
Olteanu et al. [168] https://crisislex.org/	informativeness, information types, and information sources	26	$\approx 28,000$
Olteanu et al. [166] https://crisislex.org/	on-topic / off-topic	6	$\approx 60,000$
Imran et al. [105] https://crisisnlp.qcri.org/	information types	19	$\approx 50,000$
Alam et al. [9] https://crisisnlp.qcri.org/	relevant / non-relevant	2	$\approx 21,000$
TRECIS 2018 https://trecis.org/	high-level information type	2	$\approx 2,000$
TRECIS 2019 [177; 54] https://trecis.org/	high-level information type	3	$\approx 4,000$

6.1.1 Dataset Description

To create our dataset, we reviewed existing crisis datasets that had labeled messages and that were publicly available (See Table 6.1). Most of these collections were based on Twitter, so we focused our work on this platform. We selected 6 datasets to download from the works of Olteanu et al. [166], Olteanu et al. [168], Phillips [177] and Dataverse Scholar Portal [54] (both corresponding to TRECIS 2019 [1]), Imran et al. [105], and Alam et al. [9] that met the

[1] TREC Incidents Stream

criteria of having labeled data and being publicly available, with a total of 172, 714 messages. We should note that at download time not all messages were available (e.g., some had been deleted by their authors). We retrieved messages directly from Twitter in order to obtain their full meta-data and text.

We pre-processed data by removing duplicates (messages that had the same ID), unifying existing manually annotated labels and enrichment of data information. In more detail, these steps are explained as follows:

- Label unification: consisting of the consolidating classes into two types of messages: *Related* and *Not Related* messages. We defined a message as being *related* to a crisis event if it contained an implicit or explicit mention to the crisis situation for which it was originally collected [122]. To decide if a message was related or not to a crisis we referred to their original labels. This is, if a message was labeled as either containing *informative* content (e.g., death reports, infrastructure damages, and evacuations) or *not informative* subject (e.g., caution, advice, and emotional support).
- Data enrichment: consisting of incorporating additional qualitative crisis dimensions, such as *hazard categories, hazard subcategories, temporal development, geographic spread, language, and the country and region where the event occurred.*

The resulting dataset was composed of 97, 687 messages, corresponding to 49 disasters in 10 languages[2]. Figure 6.1 displays the percentage of messages by language, demonstrating that most of the content was compiled in English, followed by Spanish and Italian in a tiny portion of the messages compared with the former. In particular, for the goals of our current analysis, from now one, we will work with a subset of this data, consisting of messages in English (86% of total) and messages that were *related* to a crisis event. Overall, this subset consisted of 40, 097 messages corresponding to 41 crisis events. We refer to this subset as our *dataset* from now on. We selected messages in English since this was the most represented language in the dataset and has a wide variety of NLP tools available. In this sense, we prioritized not introducing variations in our analysis due to the disparity of existing tools and bias in less represented languages.

6.1.2 Annotation of Crisis Dimensions

To understand the impact of different types of content in the propagation of crisis events, we annotated their crisis dimensions using a similar pipeline to that of Olteanu et al. [168]. We categorized each event by its hazard category (*natural* or *human-induced*), hazard subcategory (e.g. *intentional*, *meteorological*), hazard type (e.g. *earthquakes*, *wildfires*), temporal development (*instantaneous* or *progressive*) and geographic spread (*focalized* or *diffused*).

In our work, we considered the traditional hazard types defined by Fischer [70] and integrated in the taxonomy listed by *The International Disaster Database*[3]. On one hand, a

[2] The full unified dataset is available on `https://github.com/hsarmiento/Multilingual_labeled_crisis_dataset`

[3] https://www.emdat.be/

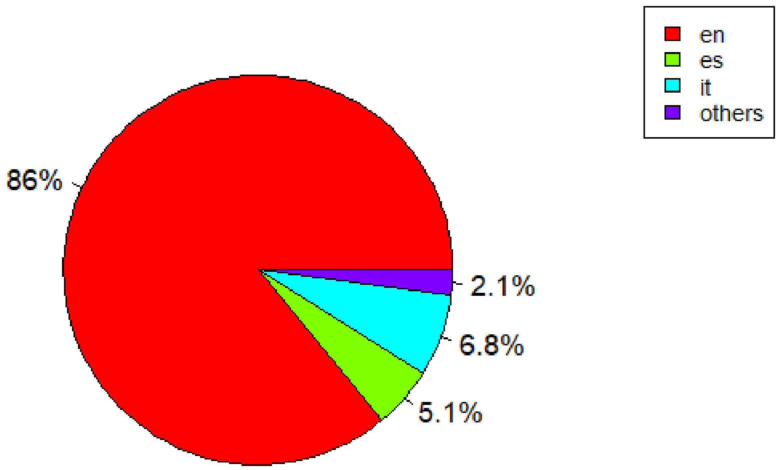


Figure 6.1: Percentage of messages by language. English messages represent more than 80% of instances in our merged dataset.

natural event considers acts of nature such as floods, tornadoes and others [20]. In opposition, a *human-induced* event is considered a consequence of technological or human hazard, for example, crashes, industrial explosions, bombings, among others. Furthermore, these categories can be divided into subcategories, for instance, *meteorological*, *geophysical* in the case of *natural* events, or *intentional* and *accidental* for *human-induced* crises. Table 6.2 shows a summary of the dataset divided into hazard categories, subcategories and types of crises.

6.2 Feature Extraction

There are several approaches to analyze the content, such as the propagation of the messages, their temporal and textual characteristics, network analysis, among others. However, due to the nature and structure of the compiled datasets, we could only investigate the content from a textual and linguistic perspective. This is because the data was labeled as isolated instances that did not consider temporal aspects or propagation aspects like conversation threads and replies among users. Literature has highlighted a challenge concerning the volume of data collected during emergencies. According to Castillo [39], crisis-related messages tend to focus on isolated messages rather than conversations, making it difficult to capture the context of human communication. This research problem aims to address this issue, as it recognizes that communication occurs within a specific context. However, data collection limitations often result in the absence of such context.

To gain insight into how social media users express relevant information during a crisis, we extracted several textual and linguistic features. As our aim is to identify domain-independent characteristics that can be observed across various types of events, we examined high-level textual features that enable generalization rather than the extraction of specific aspects for a particular event (e.g., usage of hashtags). Previous research has explored a range of features for analyzing general aspects of crisis-related messages, including classification and event detection, and across various platforms, such as Twitter and Wikipedia. For

Table 6.2: Number of English messages and events by category, subcategory and type of crisis. Symbols (*) and (**) correspond to human-induced and natural crises respectively.

Hazard Category	No. Events	No. Messages
Human-induced	12	12,439
Natural	29	27,658
Hazard Subcategory	No. Events	No. Messages
Accidental (*)	7	5,502
Climatological (**)	3	2,305
Epidemic (**)	2	1,919
Geophysical (**)	8	5,903
Hydrological (**)	7	8,699
Intentional (*)	5	6,937
Meteorological (**)	8	8,459
Others	1	373
Hazard Type	No. Events	No. Messages
Bombing (*)	2	3,025
Bombing/Shooting (*)	1	1,437
Building collapse (*)	1	411
Crash (*)	1	449
Cyclone (**)	1	543
Derailment (*)	3	975
Earthquake (**)	3	5,903
Explosion (*)	1	3,569
Fire(*)	1	98
Flood (**)	7	8,699
Hurricane (**)	3	3,639
Meteorite (**)	1	373
Shooting (*)	3	2,475
Tornado (**)	1	2,428
Typhoon (**)	3	1,849
Viral disease (**)	2	1,919
Wildfire (**)	3	2,305

instance, researchers have analyzed grammatical units in texts (e.g., part-of-speech tags), semantic categories (e.g., entities), stylistic patterns (e.g., sentiments), and platform-specific metadata (e.g., user mentions) [174; 83; 116; 67].

We computed three groups of textual characteristic following the feature selection process presented in the mentioned articles: Linguistic Features (LF), Twitter Content Features (TCF), and Entity-Based Features (EBF), obtaining an overall 54 features[4]. Each feature was represented as its relative frequency of occurrence in the event. This provides a normalized value given that each crisis had different numbers of messages. This group of features permits us to analyze events in a higher-level without incorporating themes or terms specific to each type of event or crisis (e.g., as in the case of topic analysis). Additionally, the considered

[4] The full description of features is available on `https://github.com/hsarmiento/Multilingual_labeled_crisis_dataset/tree/main/features_description`

features extracted from the text allow us a complete interpretability of messages, which is one of the main goals of this analysis. We next describe each type of features.

Linguistic Features (LF)

In this study we examined the patterns of language used in social media messages. The Linguistic Inquiry and Word Count (LIWC) tool designed by Tausczik and Pennebaker [208] allows us to compute the degree to which people use different categories of words in a text. In addition, this tool assigns terms to several linguistic and psychological dimensions of language. These dimensions are generally organized hierarchically. For instance, *anger* and *sadness* categories are included in the overarching category *affective processes*. In this work, we considered 36 linguistic features, divided into 9 overarching categories.

Twitter Content Features (TCF)

We examined the structural content of tweets using the output obtained by *TweetNLP* [171]. We estimated the occurrence of user mentions, hashtags and URLs in crisis-related messages. Instead of counting the number of these elements in a message, we indicate whether or not the message contains a user mention, a hashtag or a URL. We created three features, one for each element. Additionally, we computed the fraction of messages per event that contain at least a user mention, a hashtag or a URL.

Entity-Based Features (EBF)

We extract entity-based features with the of goal of finding general communication patterns that are not specific to a particular crisis. We used *spaCy* library[5] to extract 18 types of entities. This tool allows us to determine if a term corresponds to a specific element in the real world such as a location (on several levels), a person, a building, among others.

6.3 Crisis Analysis

In this section, we study our dataset for differences and similarities among crisis events. First, we compute the similarity among crises to quantitatively determine common textual patterns derived from social media messages across emergency events. We next divide crises into hazard categories as well subcategories to automatically detect common characteristics generated by features. We compare each of these hazard dimensions extracting features with mean significant values. We then use them to apply agglomerative clustering techniques. To evaluate the quality of clusters, we analyze if the event is correctly assigned in their corresponding hazard category (or subcategory). This evaluation allows us to estimate the accuracy of automatically grouping crises into their proper hazard characteristics.

[5] https://spacy.io/

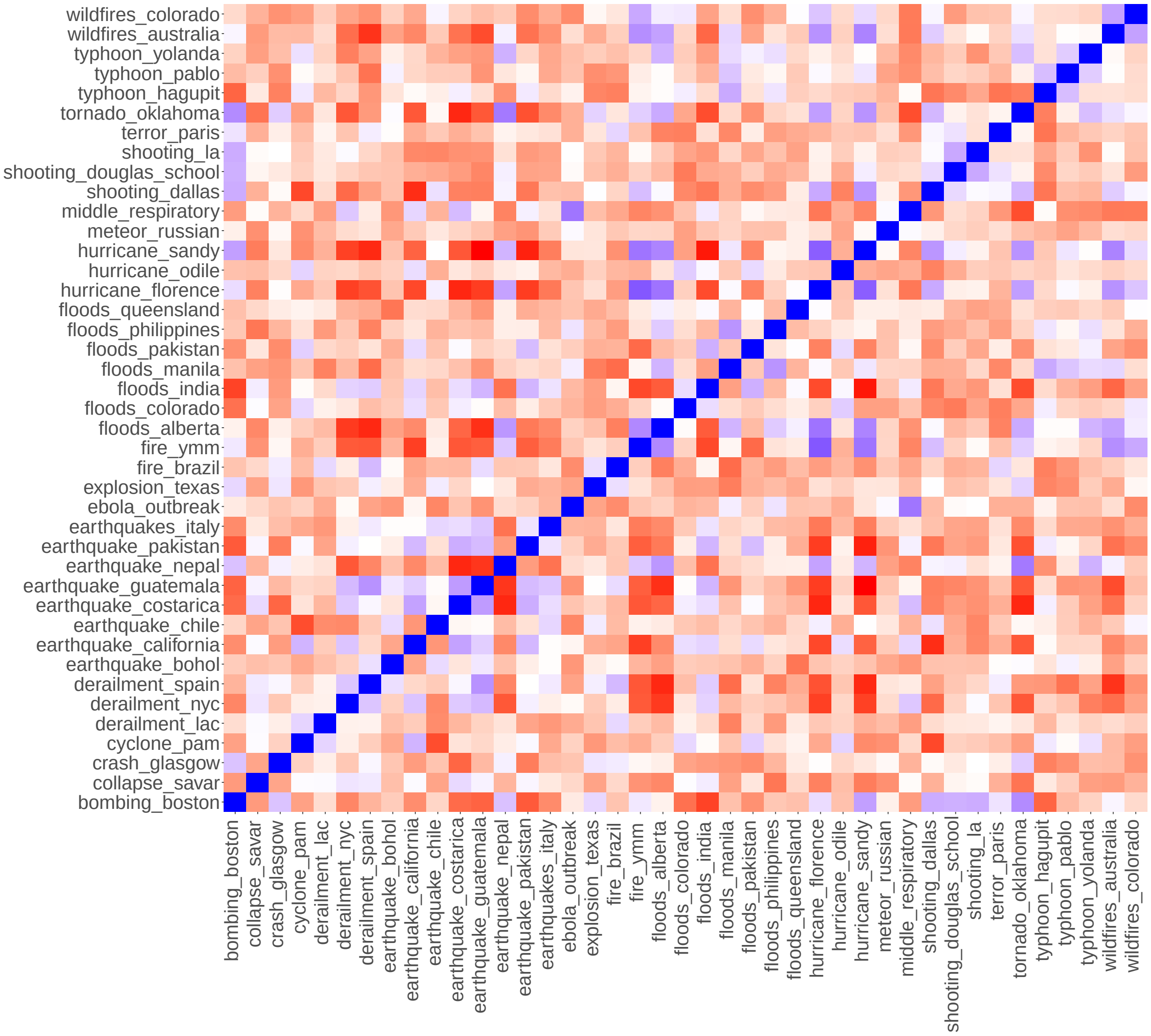


Figure 6.2: (Best viewed in color) Cosine similarity considering 54 textual features to represent crisis events. Blue and red colors mean high and low similarity values respectively.

6.3.1 Crisis Similarity

Crises are unique events that occurred within a given time period and area. Furthermore, people react in a different way depending on several factors such as culture, disaster preparation, among others. However, crisis communications can contain similar patterns based on informativeness, information types and sources [168]. In this section, we compare crises to determine whether similarities exist among common types of crises as well as differences. Instead of using labels manually classified to determine similarities presented by Olteanu et al. [168], we consider the characteristics previously mentioned in the feature extraction section. We represent a crisis as a vector feature of 54 dimensions. We use the traditional cosine similarity to quantitatively determine differences and similarities among events. Figure 6.2 shows a heatmap, which represents the results of this analysis. We observe that, in the case of events such as shootings or bombings, crises seem similar when considering all of their features (see *Bombing Boston*, *Shooting Dallas* and *Shooting Douglas School* in the heatmap). In the same way, certain earthquakes also appear to have similar characteristics among each other (see *Earthquake Guatemala* and *Earthquake Costa Rica*). Nevertheless, a

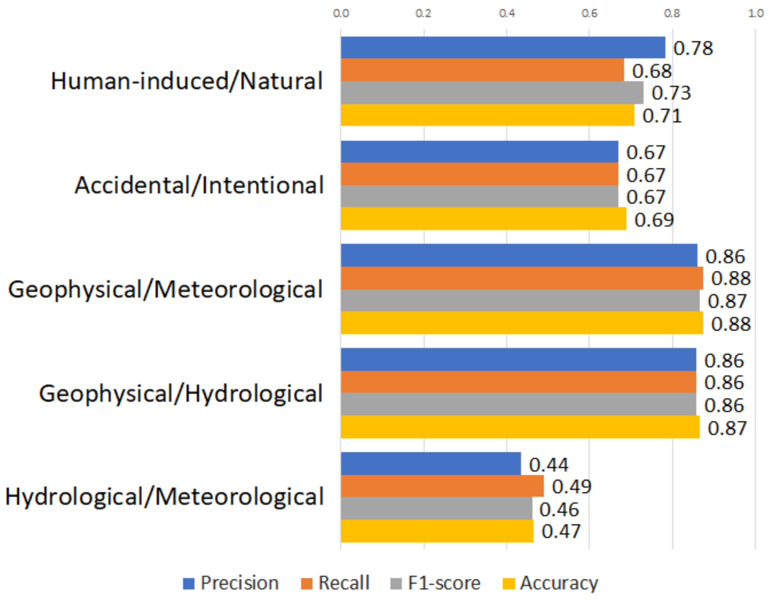


Figure 6.3: Crisis evaluation results using hierarchical agglomerative clustering techniques. Each evaluation compares two types of hazard categories or hazard subcategories.

general overview of Figure 6.2 indicates that most of crises appear to be quite different from one another.

6.3.2 Crisis Evaluation

We study events depending on their hazard categories as well as subcategories. This division allows us to determine if a set of features is significantly different from each other between two groups of crises. To do this, we compare the mean values of each feature. We first apply the arcsine transformation to evaluate a dependent variable when the raw values are proportions or percentages [170]. Based on the assumption of normality (using Shapiro-Wilk test), we apply a series of independent samples t-test to compare the values of the variables with a 95% of confidence level. In the case that the normality test rejects the null hypothesis, we use the non-parametric Wilcoxon-Mann-Whitney test. In addition to the previous tests, we also verify the homogeneity of variances using F-test and Fligner-Killeen for parametric and non-parametric assumptions respectively.

Using the features with significant differences, we apply clustering methods to automatically discover common patterns among these groups of events. To evaluate the quality of clusters, we use traditional classification measures such as accuracy, precision, recall and f1-score. For this evaluation, we consider that an event has been correctly assigned to a cluster if the cluster that contains the event also contains other events of the same hazard category (or hazard subcategory depending on the evaluation). The result of this analysis is summarized in Figure 6.3. Next, we explain in detail the analysis of the comparisons we performed for each hazard category and hazard subcategory.

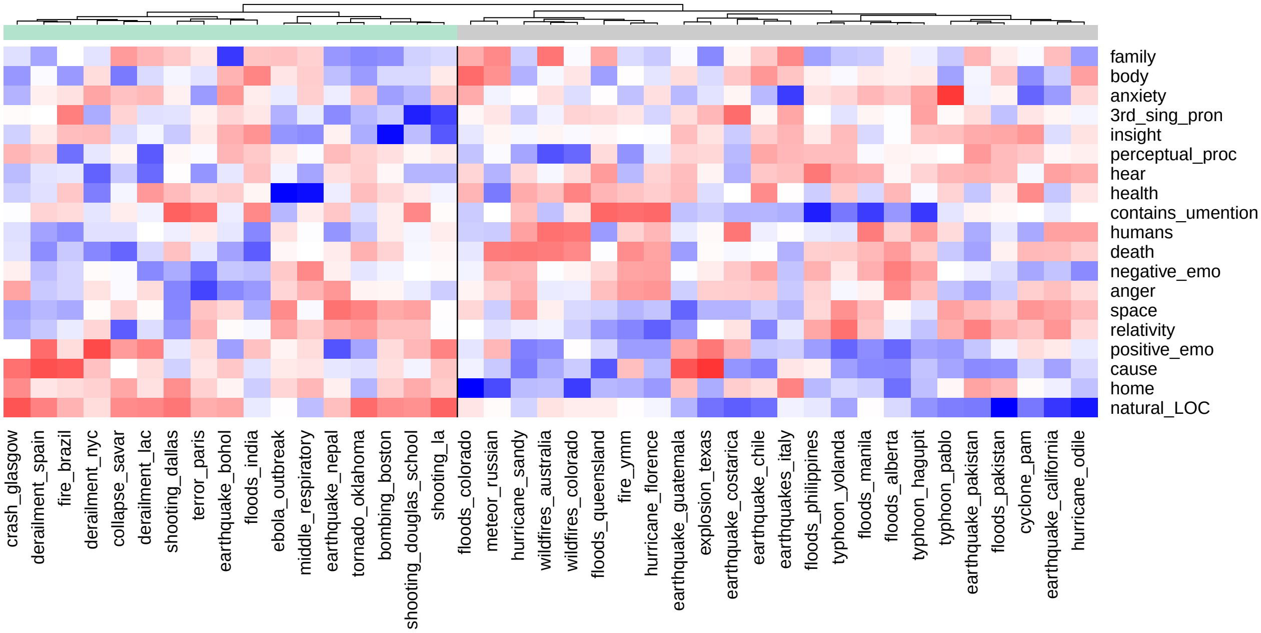


Figure 6.4: (Best viewed in color) Dendrogram obtained by hierarchical agglomerative clustering of crisis events, cut at two clusters finding a clear separation between *human-induced and natural events* (green and gray clusters, respectively). Rows represent features and columns, crisis events. Blue and red cells indicate high and low values, respectively.

Human-induced vs. Natural event messages When comparing human-induced crises against natural crisis, we found that 18 out of 54 features had significant mean value differences. Figure 6.4 shows the result of performing hierarchical clustering on the data. Most interestingly, when we cut the dendrogram at two clusters (the division is shown by a black vertical line) we obtain a very clean division between human-induced crises (left, green cluster) and natural crises (right, gray cluster). This indicates a clear separation feature-wise between both types of crisis situations.

In detail, we found that the natural event cluster was very pure (right, gray cluster in Figure 6.4), containing only natural events. These crises were mostly characterized by a high occurrence of messages that included natural places (*natural_LOC* feature), *home*-related terms, *cause*-related terms and *positive emotions*. We also note that some flood and typhoon events displayed high levels of user mentions in messages (*contains_umention feature*).

The human-induced cluster (left, green cluster Figure 6.4) was less pure, containing four natural events (i.e., a tornado, a flood and two earthquakes). Despite this misclassification, 73% of the events in these categories were correctly grouped, summarized in Figure 6.3.

Accidental vs. Intentional event messages For this analysis, we obtained 17 features with significant mean value differences. Figure 6.5 shows the result of applying agglomerative clustering method. Similarly to the previous analysis, we cut the dendrogram at two clusters (represented by a black vertical line), obtaining a clear division between intentional crisis events (right, green cluster) and accidental crises (left, gray cluster).

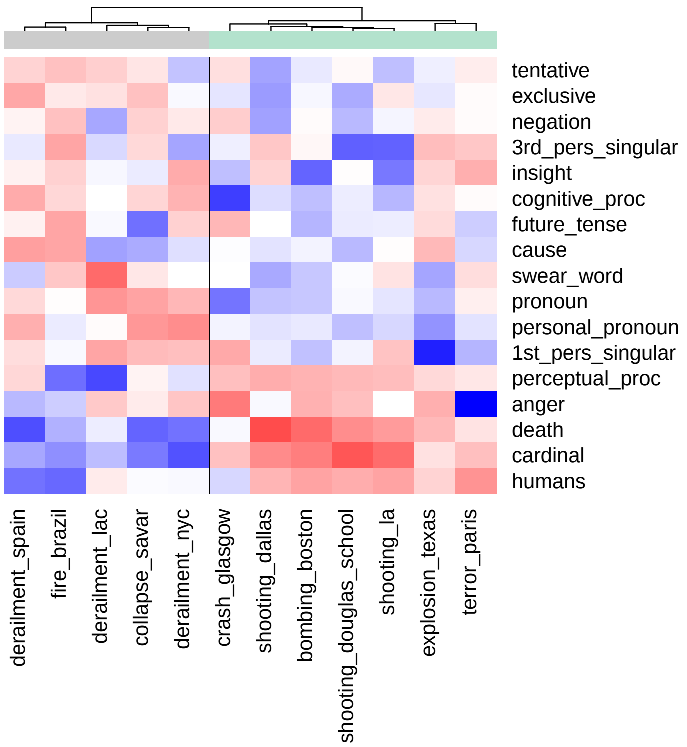


Figure 6.5: (Best viewed in color) Dendrogram obtained by hierarchical agglomerative clustering of crisis events, cut at two clusters finding a clear separation between *accidental and intentional human-induced events* (gray and green clusters, respectively). Rows represent features and columns, crisis events.

Our analysis shows that the accidental event cluster was very pure, only consisting of unintentional crises such as train derailments, a building fire, and a building collapse. Additionally, results display that features such as *personal pronouns*, *swear words*, *future tense* and *insight*-related terms, are commonly used in messages about intentional events. In contrast, messages about accidental events mostly contain terms about *humans*, *death* and *anger*. Our results indicate that we are able to group up to 67% of correct events. (see Figure 6.3).

Geophysical vs. Hydrological vs Meteorological event messages First, we compare geophysical and hydrological events. By running tests for differences between means, we obtained 6 features for which we found dissimilarities. Figure 6.6 shows the results, where left and right clusters represent hydrological and geophysical events respectively. Results display that the geophysical event cluster contains a very pure group of geophysical events (in this case, earthquakes). In contrast, the hydrological event cluster had only one misclassified event (*Earthquake Nepal*). We evaluate the quality of the groups, where we correctly group up to 86% of events (see Figure 6.3). In addition, our findings show that messages related to geophysical events had a prominent use of terms that were related to *anxiety*, *religion* and *cardinal* numbers.

Second, we compare geophysical and meteorological events. Figure 6.7 shows that we found 19 characteristics with significant mean value differences. Similarly to the previous analysis, we obtained a very clean geophysical event cluster, containing only earthquakes.

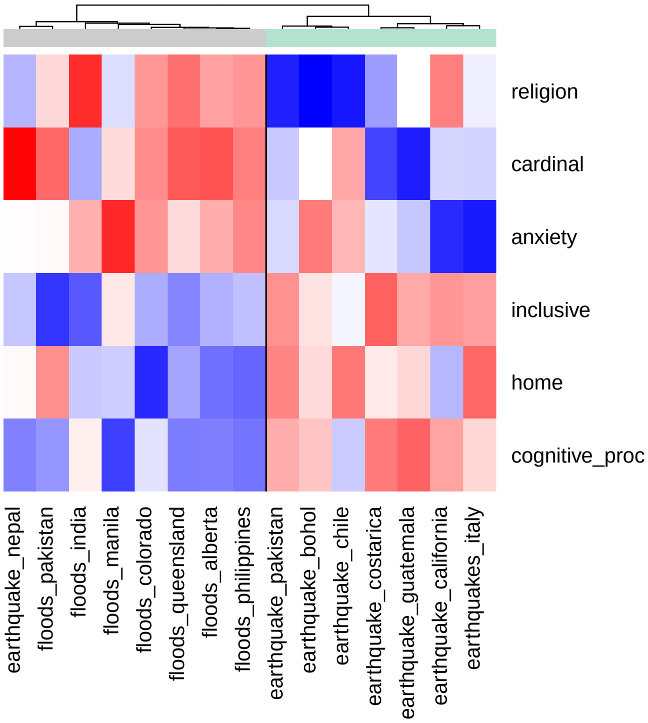


Figure 6.6: (Best viewed in color) Dendrogram obtained by hierarchical agglomerative clustering of crisis events, cut at two clusters finding a clear separation between *geophysical and hydrological natural events* (green and gray clusters, respectively). Rows represent features and columns, crisis events.

Furthermore, the meteorological event cluster had only one misclassified event (*Earthquake Nepal*) Based on clustering measures, we correctly group up to 87% of the events (see Figure 6.3). Further, our results show that messages about geophysical events commonly contain terms related to *death*, mentions of countries, cities and states (*GPE* feature) and mentions of buildings, airports, among others (*FAC* feature).

Third, we compare hydrological and meteorological events. Figure 6.8 shows agglomerative clustering results. For this analysis, we just found that 4 characteristics with significant mean value differences. However, our clustering evaluation measures show a poor cluster quality, which we correctly group up to 46% of the events (see Figure 6.3). In contrast to other crisis evaluations where just a group contain misclassified events, we note that both hydrological and meteorological clusters have incorrectly assigned events. Based on this result, we consider that these features are not sufficient to characterize these two hazard subcategories.

6.4 Discussion

Throughout our analyses we were able to automatically find cohesive groups of events using clustering techniques. On average 73% of the events were clustered into groups with other events in their same hazard category, as well subcategories. This indicates that coherent

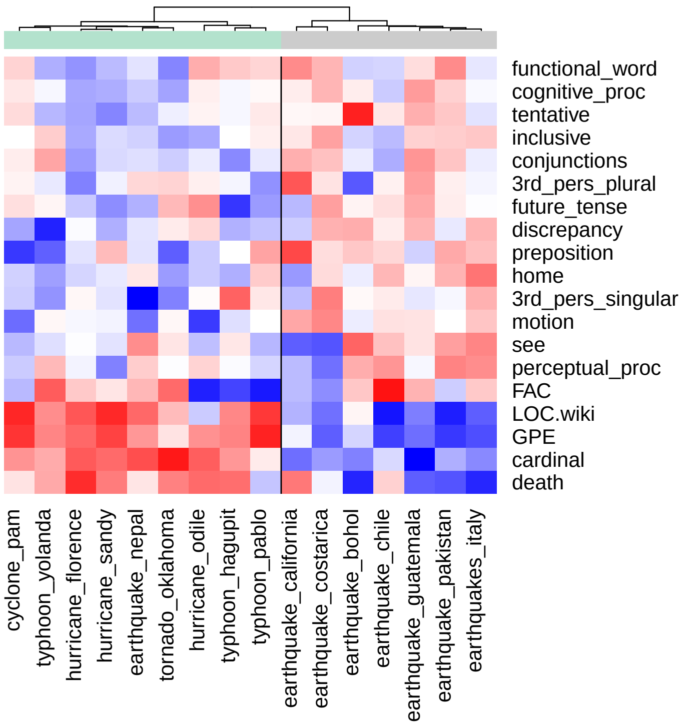


Figure 6.7: (Best viewed in color) Dendrogram obtained by hierarchical agglomerative clustering of crisis events, cut at two clusters finding a clear separation between *meteorological and geophysical natural events* (green and gray clusters, respectively). Rows represent features and columns, crisis events.

communication patterns emerge within different types of disasters. This was most clear when we compared geophysical and meteorological events as well as for geophysical and hydrological crises. In the first case, geophysical and meteorological disasters displayed cohesive collective behavior, allowing us to correctly cluster them into their category corresponding hazard category (only one earthquake was misclassified). For geophysical and hydrological events, although we only had one type of event per category (i.e., we only had earthquakes and floods), we observe that just a few features are necessary for clustering these events.

Analyzing events that were not correctly assigned to their hazard category we found that the *Nepal earthquake* was incorrectly assigned in its subcategory. We believe that results are generated given that this earthquake triggered a huge avalanche (considered as a hydrological event). Hence, people shared content about the earthquake as well as the avalanche.

When we compared hydrological and meteorological events, results revealed that there exists a high similarity among these two groups. In particular, we found only four features that had significant differences, and performance metrics showed that misclassification was very likely. Hence, our analysis suggests that either textual features are insufficient to identify differences between these two hazard subcategories, or that these events are extremely similar from a linguistic point of view.

In the case of accidental and intentional analysis, we identified two events that were misclassified: the *Glasgow helicopter crash* and the *Texas fertilizer company explosion*. According to Figure 6.5, these events were completely different from other accidental crises. For

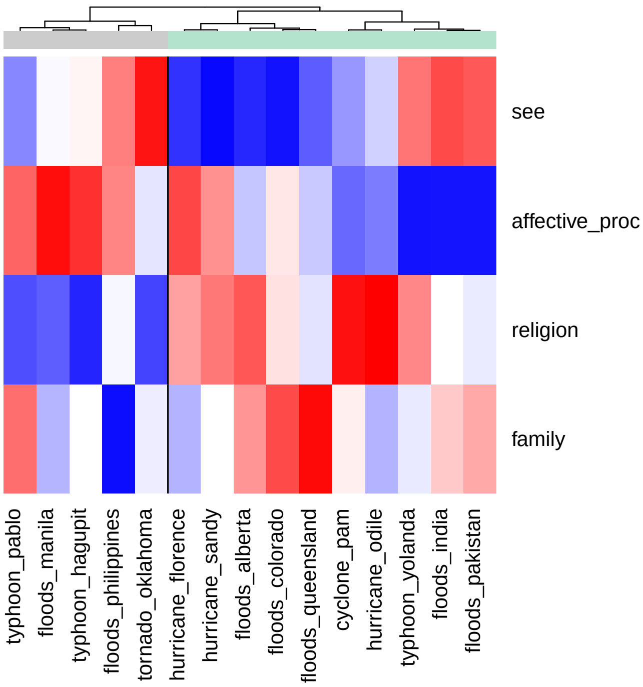


Figure 6.8: (Best viewed in color) Dendrogram obtained by hierarchical agglomerative clustering of crisis events, cut at two clusters finding a clear separation between *hydrological and meteorological natural events* (green and gray clusters, respectively). Rows represent features and columns, crisis events.

example, they had low values of *anger*-related and *death*-related terms.

For the human-induced and natural events analysis, we found that human-induced events generate more negative reactions on social media. This is interesting because according to Olteanu et al. [168], just a small portion of messages posted in Twitter are generated by eyewitness (9% on average). Hence, despite that social media users were not directly affected by an event, they had similar emotional reactions to people that were directly affected. Furthermore, our results showed that messages about natural events mainly contain *positive emotions*, *home*-related terms and mentions of natural places (*natural_LOC* feature). Additionally, we did not find differences between these types of events in features such as the use of *sad*-related terms or *swear* words. For this analysis, we obtained that six natural events were misclassified as human-induced events: a tornado, a flood, two viral disease events and two earthquakes. Like other human-induced events in the cluster, these six events contained low values of *home*-related, mentions of natural places (*natural_LOC feature*) and/or *cause*-related terms. Hence, we found that despite some events not having a direct (human) intentionality to trigger an event, they seem similar in linguistic terms. Finally, in Table 6.3 we summarize of our findings for each pair of hazard categories as well hazard subcategories. Overall, regardless of the nature of an event, and the variety of cultural barriers, geographical differences, and country vulnerabilities, we found commonalities, as well as dissimilarities across crises, exist.

To complement our results for human-induced and natural disasters, we analyze the social effects of these events studied in psychology. Baum et al. [25] were some of the first authors that addressed the problem of understanding the effects on people of human-induced and natural events. Authors identified several characteristics that differ between these hazard categories such as suddenness, power, destructiveness, predictability, perception of control and low point of the event.

The first observation in our results is about the *emotional reaction.* We found a distinctive presence of *negative* emotions and words related to *anger* and *anxiety* in human-induced events. This may be explained given that this type of event generates feelings of loss of control, which may provoke anger, anxiety and a negative mental state. According to Baum et al. [25], people may also evoke higher levels of negative effects given that the event may be controlled and triggered by others. This is interesting because according to Olteanu et al. [168], just a small portion of messages posted in Twitter are generated by eyewitness (9% on average). Hence, despite that social media users were not directly affected by an event, they had similar emotional reactions to people that were directly affected.

On the contrary, natural events present more *positive* emotions. This feeling may emerge after the low point given that people placed their attention on recovery, sharing more sympathy-related messages. In general, this positive reaction is provided by common sense and history, which allows people what to know they should do and how to confront it [112]. Furthermore, our results showed that messages about human-induced events had high values in the use of terms related to *insight*, *perceptual processes* and *space* categories. This effect may be explained given that people seem to be more perturbed about human-induced catastrophes with respect to event perception [224]. This usually happens because the loss of control is unexpected, which may require more efforts to understand and find explanations when a person triggers an event. According to Kaniasty and Norris [112], questions such as *who did it and why?, will it happen again?, when will it be safe?*, are frequently thought by people affected by human-induced events. With regard to *casualty and damage* explained in psychology studies, people seem more perturbed about damage and destruction of homes, business, disruption of power, among others. In our analysis we found that users share more content about *home*-related terms during natural events than human-induced.

In addition to psychology studies on the effects of human-induced and natural events, other studies have analyzed Wikipedia articles related to these hazard categories. For example, Greving et al. [87] revealed that there are more *anger*-related and *anxiety*-related terms articles related to human-induced events (e.g., terrorist attacks) than those of natural crises (e.g., earthquakes). Similarly to the above findings presented by the authors, we also found that Twitter users include negative content in their messages when they are exposed to human-induced crises. Ferron and Massa [67] further studied other linguistic dimensions in Wikipedia articles that included psychological categories such as *cognitive* and *social* processes. Their findings showed that articles about human-induced events focused more attention on *cognitive* processes than natural events. For example, they found higher values in *insight*, *inhibition* and *exclusive* categories. In our study, we only found that the *insight* feature was similar to the mentioned study. Furthermore, results in Wikipedia articles showed higher values in human-induced events for the overarching *social* processes category. However, we only found similar outcomes in the *family* category.

Table 6.3: List of differences found for hazard categories and subcategories. Each row represents a comparison between two hazard categories or subcategories, represented as columns in the table.

Human-induced event tweets contain	**Natural event tweets contain**
1. more *family*-related and *humans*-related terms. 2. more *body*-related and *health*-related terms. 3. more terms about *perceptual* processes. 4. more *negative* emotions, *anger*-related, *anxiety*-related and *death*-related terms terms.	1. more mentions of natural places (*natural_LOC* feature) 2. more *home*-related terms. 3. more *positive emotions.* 4. more *cause*-related terms.
Accidental event tweets contain	**Intentional event tweets contain**
1. more *humans*-related terms. 2. more *anger*-related terms. 3. more *death*-related terms. 4. more terms about *perceptual* processes.	1. more *pronouns* and *future tenses* and *swear words.* 2. more *insight*-related, *negation*-related and *cause*-related terms.
Geophysical event tweets contain	**Hydrological event tweets contain**
1. more *anxiety*-related terms. 2. more *cardinal* numbers.	1. more terms about *cognitive* processes and *inclusive*-related terms. 2. more *home*-related terms.
Geophysical event tweets contain	**Meteorological event tweets contain**
1. more *death*-related terms. 2. more terms about locations (*GPE* and *FAC* features). 3. more *cardinal* numbers.	1. more *pronouns*, *future tenses*, *prepositions* and *functional words.* 2. more *home*-related terms. 3. more terms about *cognitive* processes, *inclusive*-related and *tentative*-related terms.
Hydrological event tweets	**Meteorological event tweets**
1. Clustering metrics do not support significant differences.	1. Clustering metrics do not support significant differences.

Chapter 7

Discussion Analysis during Crises

Failures in communication could create a misconception of reality in scenarios of uncertainty by triggering - in social media platforms, for instance - the increase of mis/disinformation propagated across the network and controversial discussions that emerge during the duration of the event, among other issues. Motivated by these challenges (and additional issues) confronted by online communications during crises, we focus in this chapter on extracting and analysing valuable information shared in online platforms during emergencies in order to understand one of these threats.

Given the various dimensions that could be analyzed to understand the different problems of crisis communication, we focus on extracting valuable information generated in polarized discussions in social networks during uncertain events. Although polarization is often studied in the political sphere, recent research has shown that controversial discussions among online groups also occur during crises, for example, conversations about COVID-19 [85; 94] and mass shootings [55]. In practical usage, this study can be helpful for emergency agencies and authorities as they aim to quickly and effectively respond to crises. It can also help to understand the controversial aspects of these events.

To address the domain-independent study of crises, we propose a weak-human intervention approach for identifying and characterizing community framing (i.e., discovering and understanding polarized and controversial concepts) in online platforms. This approach allows us to discover and quantify differences between communities from a semantic point of view by helping contextualize controversial issues that emerge mainly in scenarios of uncertainty, such as a crisis event.

Our approach is based on the sequential application of community detection, topic modeling, and word embedding methods. Additionally, our framework facilitates the performance of scalable and objective framing analyses with minimal human intervention, as it does not require prior domain or network knowledge.

We choose a long-term crisis as a case study, that affected the entire population of the Chilean society. This *social uprising*, as it was called, radically affected the nation's status quo. Chile experienced a series of important protests, fueled by the country's significant social inequity, between October and December 2019. These manifestations were characterized by higher levels of violence and human rights violations executed by the Carabineros de Chile and the Chilean Armed Forces. Among other issues, a large portion of the population demanded a new constitution and changes to the current government, whereas another part of the population rejected these social demands and institutional reforms. This created a highly polarized scenario that was evidenced in online social media interactions.

Regarding our results, an apparently similar conversation topic across communities can have completely different meanings for each group. We noted, for instance, that while an online community linked the term *gente (people)* with communism and terrorism, the other associated it with police and military aggression to citizenship. Analyzing controversial issues that emerge naturally from conversations in online communities offers a deeper and great-scale understanding of today's political and societal discussions, especially during uncertain scenarios triggered by crises. In this direction, our findings have implications for contextualizing real-world social issues on online platforms, describing how users discuss similar concepts with opposing views. In addition, although communities with opposing views discussed similar concepts, our results also provided clues that conversations could converge to common themes, bridging the gap in polarized discussions.

The work presented in this section was published as follows: *Sarmiento, H., Bravo-Marquez, F., Graells-Garrido, E. and Poblete. B (2022, June). Identifying and Characterizing new Expressions of Community Framing during Polarization. In Proceedings of the 16th The International AAAI Conference on Web and Social Media (ICWSM), Atlanta, Georgia, USA..*

7.1 The 2019 Chilean Social Unrest Movement

During October 2019, a series of demonstrations were initiated in Santiago, Chile's capital and largest city. Initially, the reason was the adjustment of fares for Santiago's public transport system reaching 830 Chilean pesos (US$1.20). On October 18th, secondary school students coordinated a fare evasion campaign, leading spontaneous takeovers of Santiago's main subway stations. This triggered open confrontations with the Carabineros de Chile (the police). The situation worsened when groups of people destroyed the city's infrastructure, specifically in retail and several stations of the Santiago Metro Network, generating extensive damage across the city. On that same day, the president of Chile, Sebastián Piñera, declared a state of emergency in the most populated regions across the country. The next day, he proclaimed a curfew in Santiago to enforce order and prevent the destruction of public property.

Protests took place in several cities with demands for a new constitution, as well as Sebastián Piñera's resignation. These are considered the worst civil unrest in Chile since the end of the military dictatorship. Hundreds of human rights violations were documented during demonstrations executed by the Carabineros de Chile and the Chilean Armed Forces. Researchers analyzing this event have concluded that brutal police repression and governmental mishandling intensified the crisis as it erupted [199; 159].

Demonstrators across the country differed in terms of age, gender, and social status [58].

Keywords
chiledesperto, toquedequedaya, renunciapiñeraculiao, estadofallido, estadoemergencia, evasion, evade, estadodeexepcion, evasionmasivatodoeldia, evasionmasiva, estadoemergencia, chile, toquedequeda, piñerarenuncia, chilesecanso, toquedequedachile, milico, dictadura, estadoemergencia, chileresiste, chileenmarcha, yonoestoyenguerra, piñeradictador, estopasaenchile, fuerzachile

Table 7.1: List of keywords used to collect data from Twitter

The most iconic and popular place of manifestation was Plaza Italia square, an emblematic location in Santiago that connects several city areas with different socio-economic levels. These activities were mainly driven by young people from middle and lower income areas. They primarily demanded a fairer society based on social rights recognized and validated in a new constitution. However, a portion of the population claimed that, instead, some reforms to the existing supreme law could support social demands [98]. Journalists and social scientists identified a polarized organization of society during the event, characterized by being *against* and *in favor* of social movements [58; 182; 65].

7.2 Dataset

We collected Twitter data covering October 19 through November 30, 2019. We created our initial data collection considering a set of keywords (see Table 7.1) related to the event (e.g,. *#chiledesperto*, *#piñera*). Given that multiple sub-events occurred in the country during the social movement, identifying trending topics was challenging in collecting data in terms of the bias and diversity of the content. We complemented our data with two applications that crawl Twitter data to deal with these issues. On the one hand, we used data from Galean [175]. This platform compiles news events on Twitter and extracts conversations about them. We extracted messages that describe news events referring to Chile in their headlines for our purpose. On the other hand, we considered data from Twicalli, a tool that constantly retrieves tweets posted from Chile utilizing a list of coordinates [179].

We considered only messages published in Spanish, since Spanish is the main language in Chile. Additionally, our study analyzes well-established user accounts created before the event. Hence, we removed over 64,000 users whose creation date was after October 18, 2019. This means that almost 291,000 messages published by removed users were also filtered.

After merging the three sources and removing duplicates by tweet id, our unified collection contains almost 30 million messages shared by 2.1 million unique users, where over 7 million represents tweets and over 22.5 million retweets[1]. Figure 7.1 shows the normalized frequency by day for tweets and retweets. As noted, the frequency was much higher on the first days of the event and increased slightly between November 11 and 16. This growth was because the idea of voting on a referendum for a new constitution was discussed in those days.

[1] Our dataset, which is available for non-profit research purposes at `https://github.com/hsarmiento/`

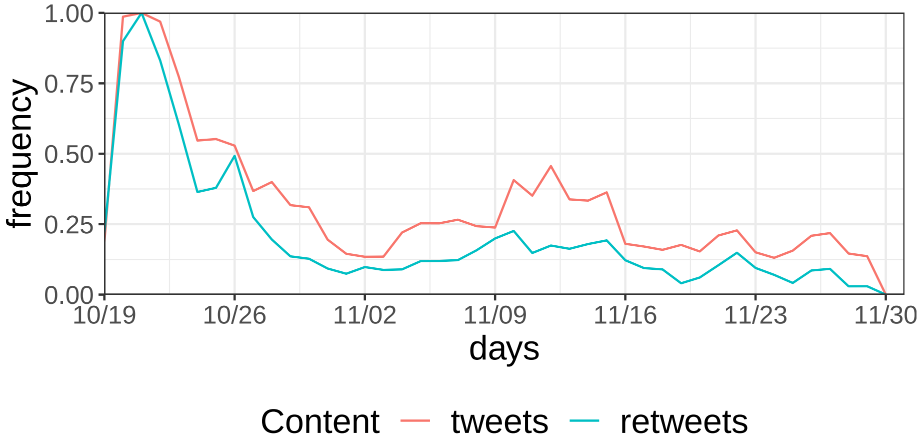


Figure 7.1: Max-min normalization of daily tweet and retweet frequencies in the period considered for our study.

7.3 Proposed Approach

In this paper, we interpret "frames" as concepts discussed by different communities around a common object (or event), which in our case is the 2019 Chilean Social Unrest Movement. We hypothesize that in these events, where communities tend to polarize towards these frames in social media, an appropriate combination of lexical and network analysis tools should allow us to analyze this framing automatically. In this sense, we consider that common topics emerging from each community conversation can be used as a proxy for "frames". Previous studies have explored this approach of identifying topics (and their most salient concepts) automatically for framing analysis [173; 222; 223].

Figure 7.2 shows a general overview of our proposed methodology. We first identify groups of users who are likely to perceive the same frames differently. For this purpose, we considered community detection methods that are evaluated in the retweet network. We then determine topics discussed by these communities using topic modeling techniques. Mapping these topics between the different communities allows us to automatically determine our frames (i.e., the concepts relevant to all other communities). Next, we extract the most salient words of these common topics to establish the target words that represent the framing. Finally, we train a joint word embedding model to project the meaning of these frames (represented by our target words) into a unified semantic space. It therefore enables us to quantify the framing by calculating vector operations between the same concept for different communities.

7.3.1 Detecting communities

We rely on the concepts of echo chambers, which states that opinions or beliefs stay in communities created by like-minded people who reinforce and endorse views of each other. We inferred communities considering that the retweet network with a clustered structure could represent different opinions and points of view. Initially, our retweet network comprised

chilean_unrest_dataset

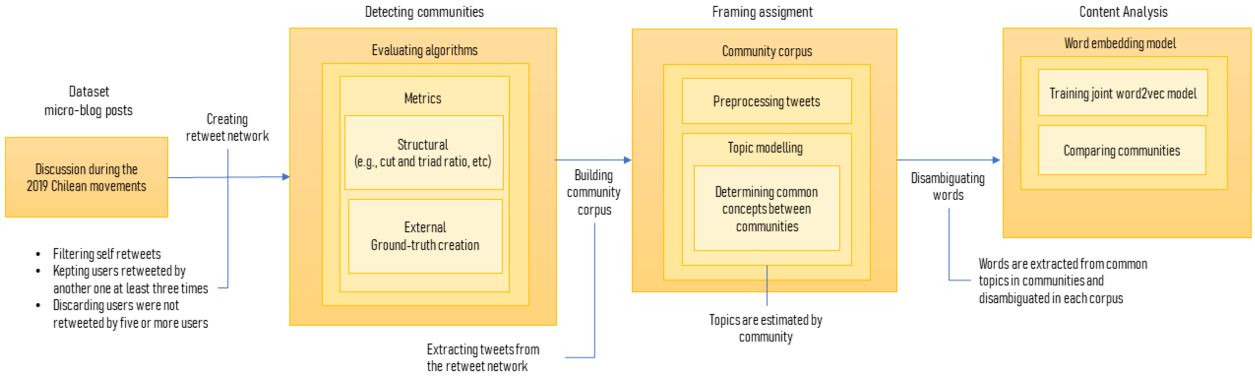

Dataset
micro-blog posts
Discussion during the 2019 Chilean movements
Filtering self retweets
Kepting users retweeted by another one at least three times
Discarding users were not retweeted by five or more users
Creating retweet network
Detecting communities
Evaluating algorithms
Metrics
Structural (e.g., cut and triad ratio, etc)
External Ground-truth creation
Extracting tweets from the retweet network
Building community corpus
Framing assigment
Community corpus
Preprocessing tweets
Topic modelling
Determining common concepts between communities
Topics are estimated by community
Disambiguating words
Content Analysis
Word embedding model
Training joint word2vec model
Comparing communities
Words are extracted from common topics in communities and disambiguated in each corpus

Figure 7.2: An overview of the proposed methodology to identify the communities and characterize the common topics they discuss.

over 2.1 million users (nodes) and 22.5 million interactions (directed edges). We filtered the network based on three conditions to reduce the noise by small communities or weak connections among users. We removed self retweets (a user that retweets his/her tweet). We kept users retweeted by another one at least three times. We discarded users that were not retweeted by five or more users. After filtering, our network contained 220,118 users and almost 900,000 connections.

To detect communities, we considered five commonly used community detection algorithms implemented on the *cdlib* library [188]: Eigenvector, Greedy, Infomap, Louvain, and Stochastic Block Model. For each algorithm, we used their default parameters (except for Louvain because we increased the resolution) according to the library documentation[2]. This approach allows the algorithms to automatically determine the number of clusters based on the data rather than imposing a specific number on the algorithms. This can be a useful approach when the number of clusters is unknown or varies in the data. Hence, we did not set the number of expected communities or clusters a priori.

To evaluate the best model, we expected to get the lowest number of communities possible and obtain the best performance by computing four classes of scoring functions described in the work of Yang and Leskovec [219]. These metrics include conductance, cut ratio, triad ratio, flake odf, erdos modularity and fraction over median degree.

Our first observation about the results was the number of communities obtained by the algorithms. We noted that most of them generate more than two clusters (in some cases, tens or hundreds of groups) except for the Stochastic Block Model (SBM), in which - in several executions - the algorithm grouped the users into two communities. Second, we show the performance of each algorithm based on the metrics mentioned above, plotted in Figure 7.3. For the first two metrics (conductance and cut ratio), we aim to minimize them. These values indicate more significant interaction with members outside their community than the community they belong to when the value is near one. For the rest of the metrics, we expect to maximize them. Our results showed that the Stochastic Block Model algorithm obtained two clusters and a better performance in almost all metrics in our experiments, except in

[2] cdlib documentation for node clustering algorithms `https://cdlib.readthedocs.io/en/latest/reference/cd_algorithms/node_clustering.html`

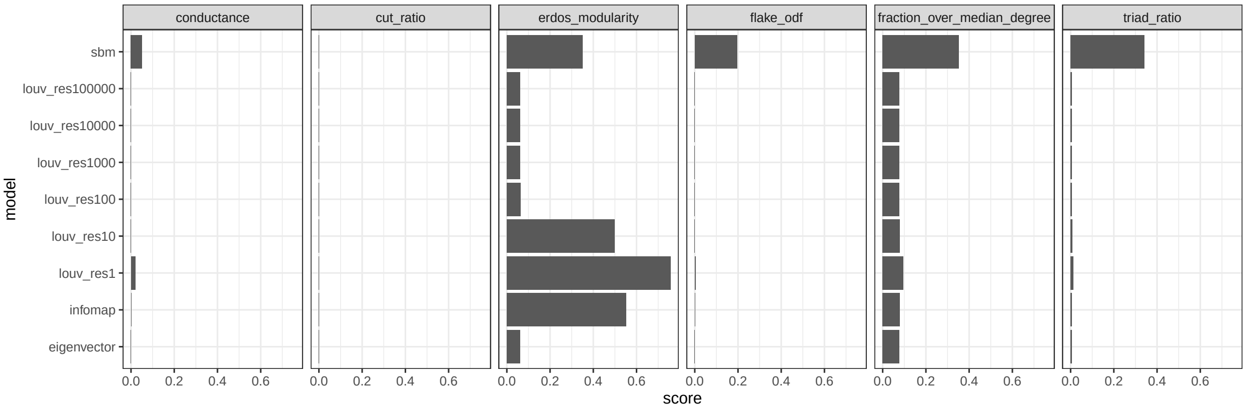


Figure 7.3: Structural metrics evaluation for different community detection algorithms. Empty scores represent minimal values in a model. For the Louvain algorithm (louv), we tested with several resolution parameters. Our results show that the Stochastic Block Model (sbm) obtained better performance for most metrics.

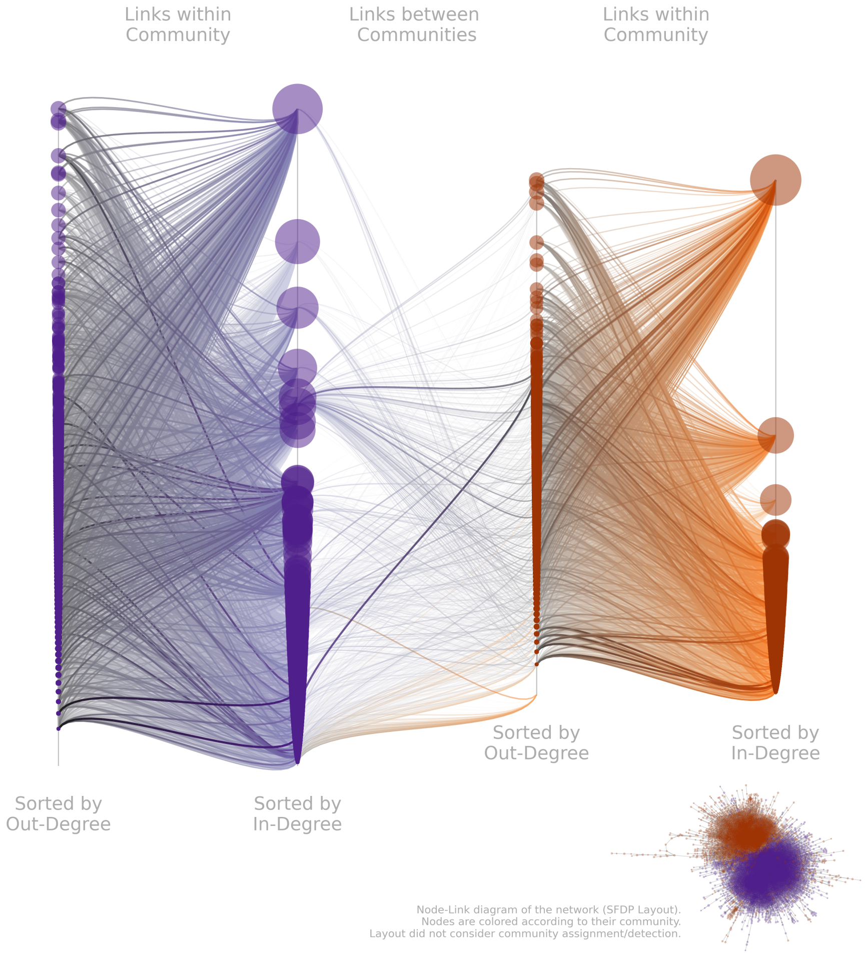
the modularity value. This can be explained because the other methods are based on the concept of maximizing modularity.

Figure 7.4 shows a hive diagram of the retweet network, where nodes (users) are colored according to their community. We sorted users by in-degree and out-degree values in each community to represent those who re-shared content from others (retweet action) and those who posted a message (tweet action). Our results revealed that the retweet interaction mainly occurred among users of the same community, where 92% of the user connections were made by users within the same group.

We evaluated the quality of the network community based on a ground-truth sample. For this task, we deemed into two types of users described as *in favor* and *against* the social movement. We randomly labeled 2,100 accounts that represent 1,110 and 990 in favor and against users, respectively. Our results display that the Stochastic Block Model obtained a f1-score, precision, and recall of 0.777, 0.7986, and 0.7574, respectively.

7.3.2 Framing assignment

Domain knowledge is an essential aspect of analyzing group polarization on social media. It allows studying different points of view by choosing specific themes or concepts to compare two or more communities. The process that people develop a particular conceptualization of an issue or reorient their opinions about a matter is described as *framing*. In general, this requires expertise and familiarity with the matter, challenging and demanding, especially for unexpected and dynamic events. To deal with this challenge, we aim to identify topics discussed in the tweets of both communities, which can be used as a proxy for framing analysis. Creating two corpora consisting of the tweets shared by the groups, we extract topics independently by the community and compare their similarities based on the terms they compose. These terms or *target words* will be later used for comparing conversations between users' groups. Our target words correspond to the most representative terms for the



Node-Link diagram of the network (SFDP Layout).
Nodes are colored according to their community.
Layout did not consider community assignment/detection.

Figure 7.4: Hive diagram of the detected communities using the Stochastic Block Model results. Orange and purple (right and left) represent against and in favor communities, respectively. The figure shows users sorted by in-degree and out-degree values in each community.

topics that are jointly discussed by all groups.

We estimated the topics using two well-known topic modeling algorithms: the Latent Dirichlet Allocation (LDA) and Hierarchical Dirichlet Process (HDP). We considered the messages of those 23,864 users in our retweet network that publish at least a tweet. On the one hand, the *against* community comprised 10,485 users and 611,976 messages. On the other hand, the *in favor* community contained 13,379 users and 945,832 messages. We applied commonly used pre-processing steps to the text, such as remove accents, URLs, Spanish stopwords, hashtags, user mentions, punctuation, and numbers, and convert to lowercase. Nonetheless, we kept numbers included in the text (e.g., p1ñ3r4) and transformed emoticons and emojis[3] to plain text to avoid encoding problems. Additionally, we applied a phrase detection model [4] to automatically extract multi-word expressions instead of using traditional n-grams. Finally, given that our interest is in users, we group messages by the user and concatenate them in one string. Hence, we trained our topic models considering each user as

[3] We used the emoji library available at https://github.com/carpedm20/emoji

[4] We used the gensim implementation available at https://radimrehurek.com/gensim/models/phrases.html

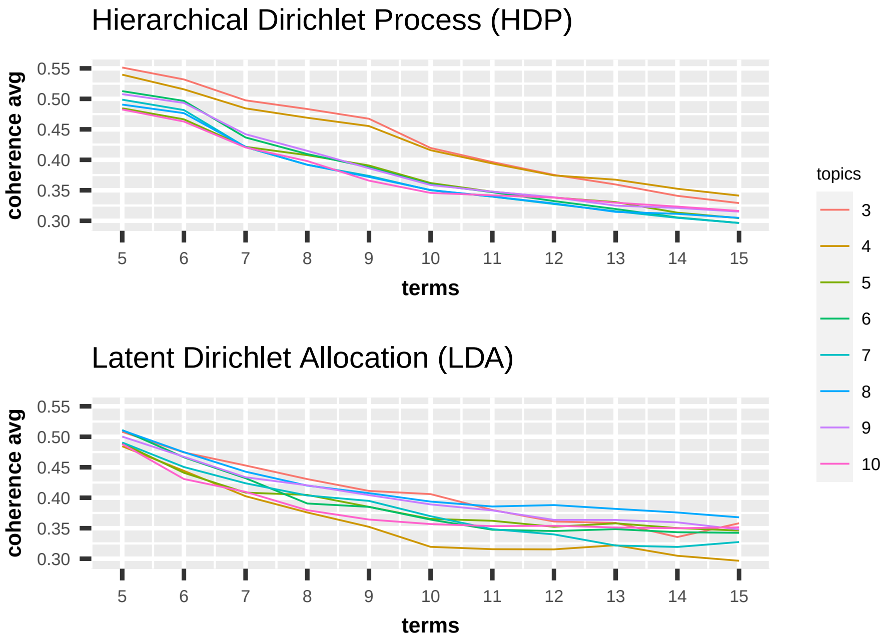


Figure 7.5: Average coherence for the HDP and LDA algorithms. Results show better average coherence between community topics for HDP.

a document.

We ran both algorithms with default parameters and extracted the topmost probable words in each topic in a range of 5 and 15 terms. In the case of the LDA method, we trained the model using the number of topics parameter between 3 and 10. For the HDP model, the algorithm did not require to set the number of topics to train the model. However, when extracting the topics, we considered the same number as the LDA method for a fair evaluation of both models. We used the coherence metric to evaluate the quality of the topic models, computed on each community's resulting topics. Figure 7.5 shows the performance of each model where we note that the HDP model shows a higher coherence value than the LDA model. In our experiments, we chose the HDP model with $topics = 3$ and $terms = 5$ which represents the best performance.

After extracting topics in each community corpus using the HDP method, we identified what are the similar themes discussed by both communities and then extract the most salient concepts that characterize group conversations. Using the overlapping similarity with a threshold of 0.8, we obtained that 6 terms described common themes between groups' conversations. These *target words* obtained from the common topics are the following: *chile, dictadura (dictatorship), gobierno (government), gente (people), piñera* and *venezuela.*

7.3.3 Content analysis among communities

To understand community framing in polarized discussions, we studied how different are the *target words* in groups. We treated the same target words from different communities as

	Cosine similarity		Euclidean distance	
term	original	null model	original	null model
gente	0.5624	0.7223 ± 0.033	2.5073	2.2915 ± 0.186
gobierno	0.5867	0.7669 ± 0.009	2.9752	2.3310 ± 0.085
venezuela	0.6023	0.7959 ± 0.013	3.4270	2.4905 ± 0.111
dictadura	0.6949	0.7799 ± 0.020	2.7852	2.4231 ± 0.118
piñera	0.7677	0.8428 ± 0.011	2.8218	2.0483 ± 0.087
chile	0.8043	0.9050 ± 0.008	1.2672	0.8948 ± 0.045

Table 7.2: Cosine similarity and Euclidean distance for target words between communities in our word embedding model. Rows are sorted by cosine similarity.

different lexical units, but forced them to reside in the same semantic space. To do this, we created a joint word vector model in which the *target words* are disambiguated according to the community they appear in. This means that the target word $word_i$ will be renamed in each corpus with a prefix *cj* to identify in which j community the word appears. For instance, the target word *piñera* was renamed as *c1_piñera* and *c2_piñera* in $corpus1$ and $corpus2$ respectively. Thus, we can apply vector operations on these words (e.g., similarity, neighborhood) to measure community framing and polarization.

Concatenating both community corpora, we trained a word2vec model using the skip-gram negative sampling method [151] and the following parameters: $window\_size = 7$, $epoch = 15$, $vector\_size = 100$ and $min\_freq = 5$. Considering our word embedding model, we first computed the Euclidean distance and the cosine similarity of a pair of disambiguated target word vectors. Using as example the target word *chile*, we estimated both metrics as the $sim\_cosine(c1\_chile, c2\_chile)$ and $eucl\_dist(c1\_chile, c2\_chile)$, where both terms were represented as word vectors in our model.

Table 7.2 shows the results of computing these metrics. Our results exhibit that the terms *gente (people)* and *gobierno (government)* obtained the lowest cosine similarity values between 0.56 and 0.58. In contrast, the words *piñera* and *chile* had the highest similarities reaching values close to 0.8. Inspecting the Euclidean distances of them, an exciting observation highlights that *chile* achieved the lowest value. At the same time, the rest of the three mentioned terms obtained more than double the first one. These initial findings suggest that: communities could be aligned to a much similar meaning between groups regarding the themes derived from the target word *chile*. Furthermore, communities could be more polarized around the frames *gente (people)* and *gobierno (government)*, this is, diverge to different subjects.

We validated our results using a null model by considering that the communities' content was randomly assigned in each group. We trained this model by swapping randomly half of the content between both communities corpora and then disambiguating the target words in each. Next, we merged both corpora, trained a joint vector model with the same parameters as the original model, and computed the cosine similarity and Euclidean distance using this null model. Table 7.2 includes the average results of the null model by swapping the corpus and training a new word vector model 100 times. We noted that both cosine similarities and Euclidean distances were utterly different from the original. Hence, our results suggest that despite the vocabulary in both original and null models being the same, our method relies on correctly disambiguating target words in each corpus.

We estimated the top nearest terms for each target word to gain more insights into the mentioned differences in the target words. We then visualized them, applying a dimensionality reduction using the T-SNE (T-distributed Stochastic Neighbor Embedding) algorithm. Figure 7.6 shows two examples of the top-25 nearest terms for the target words *gente (people)* and *chile*, which had the lowest and highest cosine similarities respectively [5]. Our results show that, for both target words, nearest words have a similar distribution in a 2-dimension space, where communities can be visually identifiable. However, we noted differences in the number of overlapping words that both communities had depending on the target word analyzed. The word *gente (people)* did not share any term between communities, while the *chile* had three common terms.

Expanding the analysis to a higher number of top nearest terms for these target words, we noticed a similar pattern in which the number of common words increases depending on how close the cosine similarity was. Figure 7.7 shows the Jaccard index for different number of k-nearest terms, revealing that the target words *gente (people)* and *gobierno (government)*, which had the lowest cosine similarities, slightly suffer a change in the number of common nearest terms when the neighbors' words increase. In contrast, the other target words have more than four times the common nearest words compared to the two previous ones.

To understand what themes or concepts surround the target words by the community, we quantitatively inspect the top nearest terms to find differences or similarities in the group discussions. Like the previous analyses, we focused on those target words where we found the highest and lowest cosine similarity values. We identified for the target word *gente (people)* that most of the terms found in the against community were linked to communism attacks associated with crime, street vandalism and terrorism (See Table 7.3). Several statements of this community declared that, in the chance of yielding to the demands of the people, Chile would turn us into Venezuela and Cuba [73]. And that the whole social outburst is not the fault of Sebastian Piñera but Cuban-Venezuelan secret agents. Hence, most of the conversations may be in opposition to communist governments. In contrast, the in favor community highlighted words that allude to violence and repression by the police and the military against the country's citizens. During the social unrest, both institutions were accused of human rights violations during curfew hours [162]. These actions were reported in social networks with multiple messages and images. Additionally, different expressions showed that Carabineros and Chilean Armed forces were mentioned colloquial and derogatory (paco and milico) followed by an insult or curse word. Finally, several terms also mentioned criticism of press coverage and unease with the government's actions (See Table 7.3).

For the target word *government (gobierno)*, we noted that the nearest terms for the against community referred to protecting the police and normalizing the situation because of the demonstrations. Oppositely, the in favor group linked their nearest terms with criticizing the president and the deployment of the military in the streets. This last situation alluded

[5] The complete visualization of all target words is on the Appendix

Target word	Against	In favor
gente (people)	comunistas_atacando (communism_attacking), comunistas_protestas (communism_protests), romper_quemar (smash_burn)	milicos_disparando (military_shooting), golpeando_disparando (smashing_shooting), pacos_pegando (carabineros_smashing), noticieros_culiaos (fucking_news), piñera_payaso (piñera_clown)
gobierno (government)	respeto_instituciones (respect_institutions), proteger_carabineros (protect_carabineros), normalizar_situacion	piñera_cobardia (piñera_cowardice), autocritica_gobierno (self-criticism_government), ejercito_salir (army_takesstreet)
chile	aparece_manifestaciones (appears_manifestations), violencia_marxista (marxist_violence), pdte_piñera (president_piñera), chile_mierdaaa, querida_patria (dear_country)	chile_levantamiento (chile_ uprising), tanques_calles (tanks_streets), pdte_piñera (president_piñera), chile_mierdaaa, querida_patria (dear_country)

Table 7.3: Examples of nearest terms for target words divided into against and in favor communities.

to those who suffered in Chile 30 years ago, where the military took to the streets before the 1973 Chilean coup d'état[6].

Regarding the target word *chile* (having the highest cosine similarity), we found fewer differences than the previous words. For instance, the against community mentioned Marxism violence while the in favor group discussed torture and dictatorship. However, as Figure 7.7 shows, we found several common nearest words between communities related to the president of Chile and expressions of support related to the country.

Finally, we consider a low dimension analysis to quantify target word differences in communities. We followed a similar proposal presented by Sweeney and Najafian [205] that measures fairness in word embeddings via the relative negative sentiment. The general idea is that words can be projected from the embedding space into a sentiment probability by training a logistic regression on some pre-labeled words (a set of 80 positive and negative emojis in our case [215]). Then, the probability of negative sentiment for some sensitive words (the target words of each community in our case) is normalized to a probability distribution. Subsequently, the negative sentiment distribution of each community is compared to a uniform distribution using the Kullback-Leibler (KL) divergence as a measure of bias. This allows us to quantify the magnitude of polarization for the target words of each community under the assumption that a neutral community should be closer to a uniform distribution.

Figure 7.8 shows the estimated sentiment probability of the target words for each community. Our results suggest that, in general, the same target word exhibits a different sentiment probability between groups. We also noted that the term *piñera* was unique in the sense

[6] https://es.wikipedia.org/wiki/Golpe_de_Estado_en_Chile_de_1973

that it obtained a similar polarity in both communities. Observing the polarities in detail, we obtained an unexpected result for the target word *dictadura (dictatorship)*, where the in favor community showed a higher positive level than the against the group. This result can be attributed to the inability of word embeddings to discriminate between the different meanings of a word. Thus, "dictatorship", referring to the Chilean (right-wing) dictatorship of Pinochet, and the Venezuelan (left-wing) dictatorship of Maduro, could be being conflated.

Regarding the KL scores, we compared the negative distributions of each group with a uniform distribution using the Kullback-Leibler divergence (KL) (see the bottom image in Figure 7.8). We obtained values of $KL_{against} = 0.0215$ and $KL_{infavor} = 0.0498$ for the against and in favor communities respectively. As noted, the value of $KL_{against}$ doubles the value of $KL_{infavor}$, representing a significant difference in the amount of information necessary to encode and transmit from one distribution to another. Hence, our results suggest that target words exhibit different polarities in both communities and that the in favor community shows more intense sentiment states.

7.4 Discussion

We have presented a polarization analysis using as a case study the 2019 Chilean social unrest movement. We aimed to identify users with particular stances and to understand how different they were based on social media data. Compared to previous works that relied on supervised methods and domain-specific knowledge to identify communities and check similarities and differences in content, our unsupervised methodology requires only minimal human intervention to address these challenges. Additionally, our procedure is adaptable to other events and languages as it extracts topics of interest directly from the text and trains the embedding model using the communities' corpora.

Our results showed that we correctly assigned users' stances with an f1-score of 0.77. Although our results were slightly worse than those in the literature (f1-score of 0.80 on average), our method compensated for that difference by providing disentangled meaning around specific topics. This suggests that user interaction and content are closer than expected, bringing additional insight into the understanding of polarized users.

Our method considered a framing assignment process to automatically find commonly discussed concepts among the communities. The benefit of this automatic selection was that it did not require prior domain knowledge expertise about the themes under study, especially when events seem powerfully dynamic and unexpected. For Twitter-based research, this could complement the use of hashtags for polarization studies, particularly when these cannot be identified for two or more communities, as well as suggesting common keywords that can be compared across groups.

Our approach to estimate topics that are jointly discussed by several groups can contribute to real-time message collection for this type of event. Currently, data retrieval is performed by considering a set of initial keywords related to the target event. For dynamic analysis, this requires updating keywords based on trending topics and hashtags, which are usually estimated by global frequencies. However, this approach has difficulties in representing topics related to minority communities. In this sense, our method overcomes this drawback because the salient topics are estimated at the community level.

Our research could benefit political and social scientists in understanding how dynamic conversations on Twitter show insight into high-impact events in the real world. We observed that our studied frames, and the semantics around them, are still present in the ongoing discussion. For instance, the country's perception (Chile), institutional violence and human rights, the high adherence to manifestations, and the reactions to government actions. In addition, our study could exhibit future social connectedness and political behavior for forthcoming events. Chileans voted whether a new constitution should be drafted a year after the movements. Results showed that 78.28% favored a new constitution, while 21.72% rejected the change. The option "reject a new constitution" won mainly in the three wealthiest municipalities, historically associated with right-wing electoral strongholds. This electoral analysis references the political and economic elite contrary to the reforms promoted after the 2019 social movements [212]. Although our method did not determine social and economic status for users, our results provide insights that communities in Twitter can reveal existing polarized groups about specific topics. Measuring and contextualizing polarization helps researchers complement traditional methods such as surveys or opinion polls, displaying the general overview of the population during social discussions around themes in a low-cost and real-time manner.

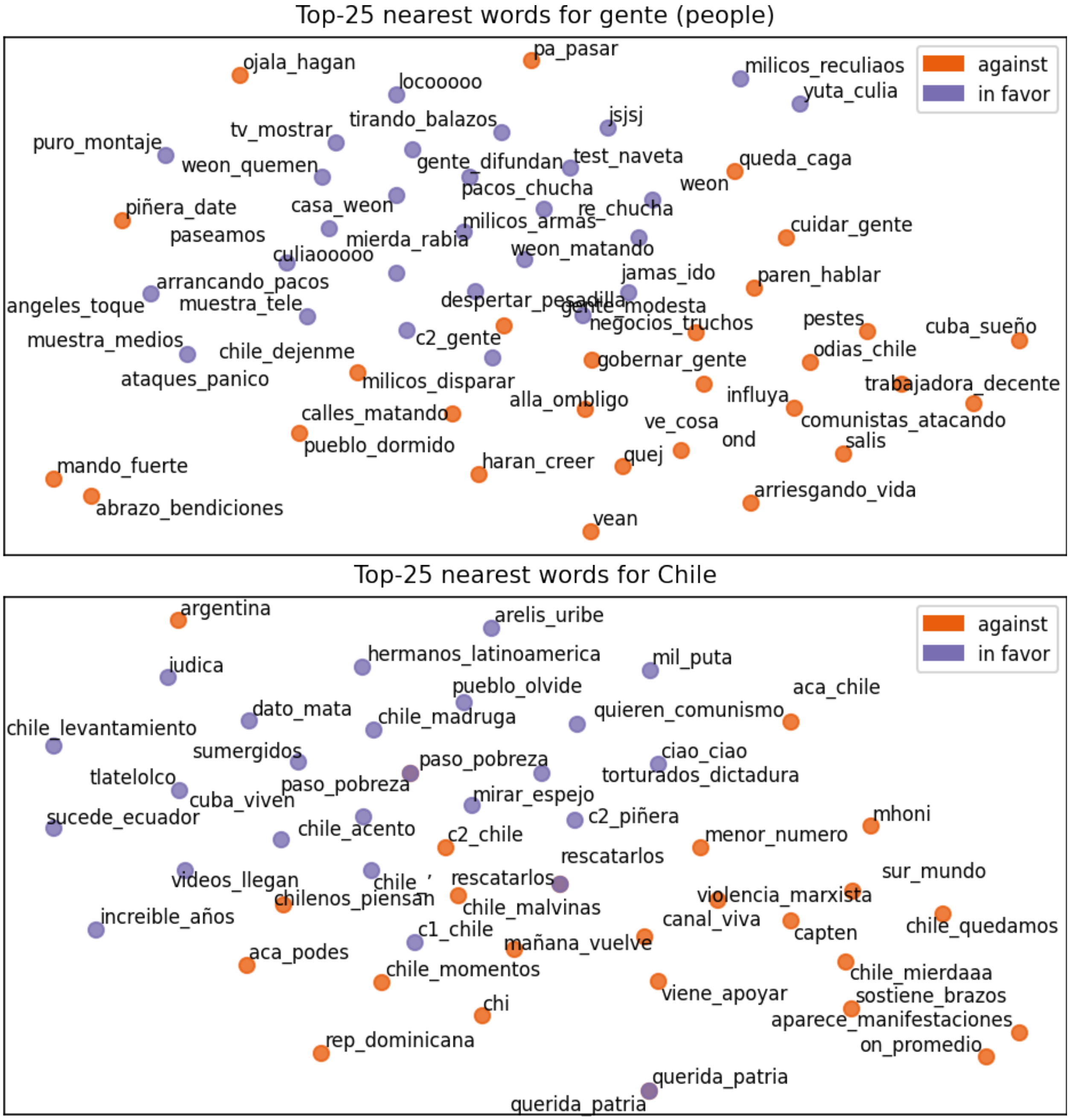


Figure 7.6: Two dimensional word vector representations of top-30 nearest words for the target words *gente (people)* and *chile*. Orange and purple points represent the nearest words against and in favor communities, respectively.

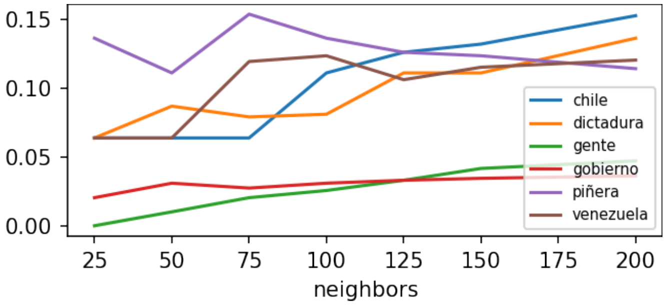


Figure 7.7: Jaccard index for the k-nearest terms found between communities for each target word. Results show that when the number of neighbors words increased, terms such as *gente* and *gobierno* had non-variable Jaccard index.

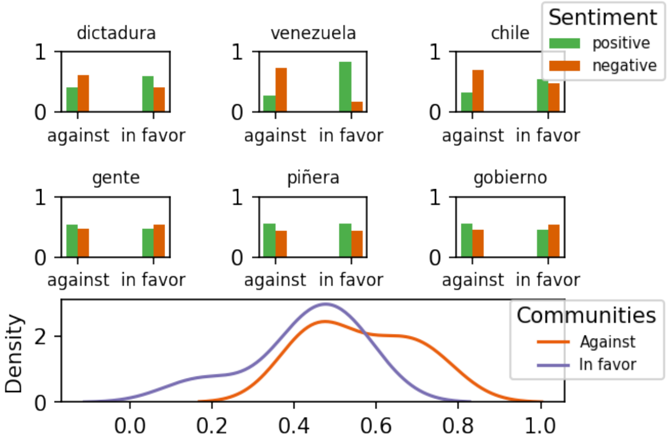


Figure 7.8: Sentiment probability of the target words. The top image shows the community's positive and negative probabilities of each target word. The bottom picture displays the probability density distribution of the negative sentiment.

Chapter 8

Conclusions and Future Directions

In this chapter, we present the main conclusions of the presented thesis. We first discuss and evaluate the hypothesis, objectives, and research questions described at the beginning of this work. We then outline the future projections and challenges derived from this work.

8.1 Conclusions

In Chapter 4 we presented a methodology for detecting an emergency situation based on the locations of a specific country. This approach is independent of the textual features and can be used in different types of events and languages. We show that the users act as self-organized in the affected locations like citizen sensors when an emergency occurs. We furthermore have presented an analysis of geographic spread for different types of events that can be categorized. However, our experiment considers just a small portion of emergencies, which is not representative of all types of crises according to either the hazard type (natural or human-induced), temporal development (instantaneous or progressive), or geographic spread (diffused or focalized).

In Chapter 5, we assessed the impact of transfer learning in the cross-domain cross-lingual classification of messages relating to crises. In particular, data from a high-resource language such as English can contribute to the classification of messages from low-resource languages such as Spanish and Italian. Furthermore, adding messages from the target language also helps in some cases. Our findings indicate that there exist patterns in crisis communications that expand across crisis domains and languages. As a result we can increase our ability to classify data in languages and domains for which we have little to no labeled data. However, the most efficient data representations may vary depending on the target language.

In Chapter 6, we addressed an analysis of messages related to a large collection of crises. Using traditional natural language processing techniques, we found differences and similarities among diverse hazard categories and subcategories as well. We showed that with a compact textual representation, crises can be described and grouped into events with similar characteristics based on information published on micro-blog platforms. Furthermore, we showed that it is possible to group events in their corresponding categories with an average accuracy of 75%. We also found that social media users display similar reactions to those of people directly affected by disasters in the physical world. These findings suggest that it could be possible to focus on online user-generated content platforms to study the initial effects on people after a crisis.

Unlike psychology studies that are based on interviews of a few hundred people, we exploited the large adoption of online social networks, which provide data on millions people in near real-time. Furthermore, this content is contributed spontaneously, which makes a huge difference with psychometric methods where personal interviews are required. This fact can contribute to gaining fast and low-cost situational awareness in times of uncertainty during crises. For example, emergency practitioners, government, news media and the general public, can provide rapid and conscious responses when they know what type of reactions are generated by crises in affected people.

In Chapter 7, we studied discussion analysis in social media during extreme events, where we considered a case study of the 2019 Chilean unrest movements. The study is mainly motivated by the characteristics of the event, such as the messages' language, event location, duration, the impact on Chilean society, and (polarized) discussions that emerge during highly uncertain events. Our results exemplified how controversial topics were perceived in extreme events, creating parallel realities regarding the same issue in online conversations. The implication of this analysis can accelerate situational awareness during crises, especially when people's reactions are often measured by peer review questionnaires to a limited portion of citizens.

Finally, our work also released several resources. We published two datasets composed of a large collection of crisis-related messages posted from different public sources, hazard domains (i.e., earthquakes and bombings), languages (i.e., English, Italian and Spanish), and locations. In addition, we released a novelty dataset that comprises social media messages related to the 2019 Chilean unrest movements. This collection is composed of millions of Spanish messages used to study polarization during the main stages of the event. Furthermore, we enriched these three datasets with useful information such as the unification of labels for consistency, adding metadata in relation to the event (e.g., geographical location, hazard category and subcategory, temporal development, geographic spread and language) and so forth. Subsequently, we uploaded all our experimental analyses and proposed methodologies used across the different parts of this thesis to reproduce our experiments.

8.2 Evaluation of the Thesis Objectives

Our hypothesis presented in Chapter 1 states that "there are patterns in the self-organized activity of the Web and social media users that emerge when a crisis situation starts to unfold in the physical world. Some of these patterns arise independently of the particular type or domain of the crisis event, as well as independent of the location, language and culture of the users that participate." Having completed the work described in this thesis, we are able to validate our hypothesis.

Our method for detecting crises empirically demonstrated that independent of the location and language from where social media messages were shared, high-impact real-world events can be detected using anomalies related to locations mentioned on social media platforms. Furthermore, these events can be differentiated depending on their type (i.e., earthquakes or terrorist attacks) considering propagation properties on social media platforms. Hence, online users react in a different way considering the type of event, but in a similar manner for similar crises independent of the language and location of the messages.

Consequently, our cross-domain and cross-lingual evaluation of crisis-related messages, across different classification scenarios and data representations, showed that cold-start issues can be partially addressed introducing data from different domains and languages. This demonstrated that characteristics of other events and languages can be incorporated to reduce the amount of noise and non-related messages published during emergencies.

We also conducted an analysis to find differences and similarities of a large collections of crisis-related messages. Given the availability of a great number of messages for different events and diverse languages, we focused our study on English messages. We discovered transversal patterns in social media messages based on a series of linguistic features extracted automatically from the posts. These findings showed that events can be automatically grouped into common crisis dimensions (i.e., hazard categories or subcategories) using compact representations.

Finally, we analyzed online discussions in order to extract relevant information that may improve situational awareness during emergencies. We addressed this challenge by proposing a weak human intervention method that helps to identify and characterize polarized conversations, one of the threats affecting crisis communications in the digital era. Although our methodology was evaluated only in Spanish messages about the 2019 Chilean unrest movements, our method can be adapted for any language and domain because it does not rely on supervised methods for detecting communities or pre-trained textual models to understand the content shared by users during crises.

In the following, we review the objectives of the work and assess the level of their achievements.

1. **Consolidate a large-scale dataset of Twitter messages from diverse crisis events enriched with relevant metadata.** A comprehensive review of research and public sources was presented in Chapters 5 and 6 . Based on the literature, we consolidated and enriched multilingual and multi-domain datasets released in our repositories, which were enriched with relevant metadata such as event location, languages, crisis dimensions, etc. In Chapter 7, we introduced a novelty dataset composed of messages related to the 2019 Chilean unrest movement. This collection differs from the state-of-the-art in language, location, and characteristics of the event.
2. **Study domain-independent and language-agnostic representations of crisis event messages to understand communications patterns through crisis dimensions, locations and languages** We proposed different approaches to understanding crisis communications. In Chapter 4 we presented a method for emergency event detection that identifies extreme events in social media in a domain-independent

and language-agnostic perspective. Our experiment was conducted considering different events which occurred in different locations and evaluating it for messages in several languages.

In Chapter 5 we presented an extensive evaluation of different scenarios and data representations to study cross-domain and cross-lingual message classification. Like the analysis presented in Chapter 6, we considered multiple events from different locations, times, and several languages. We showed experimentally that it is possible to use data from high-resource languages and from multiple domains to classify messages of new (previously unseen) event domains and languages. Furthermore, we addressed the cold-start deployment scenarios, dealing with a major problem of labeled data scarcity in low-resource languages.

In Chapter 6, we characterized crisis-related messages published in English. Our findings showed that by using only a small set of text-based features, we are able to differentiate among different types of crisis events. Implying that different events create very specific reactions that can be identified within microblog communications.

In Chapter 7 we proposed an unsupervised approach to automatically extract and characterize valuable information posted during crises. Given the diverse challenge that threaten crisis communications, we focused on analyzing polarization, specifically polarized concepts that are used as a proxy for framing analysis. We applied our method to the conversations around the 2019 Chilean social unrest to quantify the differences and similarities in the messages. Our findings showed that communities were polarized regarding users' interaction and the common terms derived from their messages. Results demonstrated that these terms displayed different quantitative and qualitative patterns between groups.

3. **Develop computational tools that help characterize, classify, detect and extract useful information shared during crises.** Across the different proposed studies conducted in this thesis, we make available all the codes and implementations in our repositories for future research, facilitating the reproduction of our methods for researchers.

8.3 Future Directions

In the presented work, we aim to perform a large-scale transversal study of crisis events across different types of events and geographic locations to understand general and useful patterns in social media messages. To achieve this goal, we proposed several approaches for characterizing, extracting, classifying and detecting crisis-related messages from multiples events, languages and locations. As the thesis work was being concluded, we identified several challenges for future research and improvements that we enumerate as follows.

There are many things that can improve our results for the event detection. We can add Point of Interest to our gazetteer tree to increase the frequency by time-windows in each hierarchy. Furthermore, we may add more non-textual features as number of retweets and tweets, unique locations detected and special locations. We also plan to study the relevance of the different metadata-levels and assign weights for each. Finally, we can create a web application to visualize events in real-time.

For the crisis-related message classification task, we want to explore in-depth different deep learning classifiers and techniques for cross-lingual and domain adaptation. Furthermore, we would like to further improve our dataset by manually relabeling certain mislabeled messages that are effectively related to crises according to our task definition. This will allow us to have a data collection that is oriented towards crises in general, rather than event oriented. Moreover, we will explore this knowledge transfer approach in a more fine-grained task, such as categorizing actionable humanitarian information.

In the case of crisis characterization presented in Chapter 6, we are aware of the limitations given that our work was not exhaustive in terms of the number of events per category, messages and imbalanced data. We understood that by the nature of the events, people are usually more affected by natural crises than human-induced disasters. To deal with these issues, our future work will include other public data sources, our own data retrieved from Twitter as well as datasets made available by other authors. We will also compute other textual features such as topic modeling and word embeddings. Additionally, our analysis will explore whether or not these patterns can be replicated in other languages like Spanish and Italian.

Finally, our unsupervised approach for identifying and understanding polarization can be also applied in other datasets that differ in locations and languages. In addition, our methodology could track frame evolution over time and measure their lifetime, especially when sub-events determine the agenda of the social unrest. Furthermore, we can study more embedding models and representations, for example, by introducing contextualized sentence embeddings to model tweets instead of terms.

Bibliography

[1] Fabian Abel, Claudia Hauff, Geert-Jan Houben, Richard Stronkman, and Ke Tao. Twitcident: fighting fire with information from social web streams. In *Proceedings of the 21st International Conference on World Wide Web*, pages 305–308. ACM, 2012.

[2] David S Adams. Policies, programs, and problems of the local red cross disaster relief in the 1960s. 1970.

[3] Saifuddin Ahmed, Jaeho Cho, and Kokil Jaidka. Framing social conflicts in news coverage and social media: A multicountry comparative study. *International Communication Gazette*, 81(4):346–371, 2019.

[4] Wasim Ahmed. Using twitter as a data source an overview of social media research tools (2021). *Impact of Social Sciences Blog*, 2021.

[5] Shahriar Akter and Samuel Fosso Wamba. Big data and disaster management: a systematic review and agenda for future research. *Annals of Operations Research*, pages 1–21, 2017.

[6] Hajer Al-Dahash, Menaha Thayaparan, and Udayangani Kulatunga. Understanding the terminologies: Disaster, crisis and emergency. In *Proceedings of the 32nd annual ARCOM conference, ARCOM 2016*, pages 1191–1200, 2016.

[7] Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. Polyglot: Distributed word representations for multilingual nlp. *arXiv preprint arXiv:1307.1662*, 2013.

[8] Rami Al-Rfou, Vivek Kulkarni, Bryan Perozzi, and Steven Skiena. Polyglot-ner: Massive multilingual named entity recognition. In *Proceedings of the 2015 SIAM International Conference on Data Mining*, pages 586–594. SIAM, 2015.

[9] Firoj Alam, Shafiq Joty, and Muhammad Imran. Domain adaptation with adversarial training and graph embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1077–1087, 2018.

[10] Firoj Alam, Ferda Ofli, and Muhammad Imran. Crisismmd: Multimodal twitter datasets from natural disasters. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12, 2018.

[11] Firoj Alam, Umair Qazi, Muhammad Imran, and Ferda Ofli. Humaid: Human-annotated disaster incidents data from twitter with deep learning benchmarks. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 933–942, 2021.

[12] Firoj Alam, Hassan Sajjad, Muhammad Imran, and Ferda Ofli. Crisisbench: Benchmarking crisis-related social media datasets for humanitarian information processing. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 923–932, 2021.

[13] Frederike Albrecht. *The social and political impact of natural disasters: Investigating attitudes and media coverage in the wake of disasters*. PhD thesis, Acta Universitatis Upsaliensis, 2017.

[14] Abeer Aldayel and Walid Magdy. Assessing sentiment of the expressed stance on social media. In *International Conference on Social Informatics*, pages 277–286. Springer, 2019.

[15] Abeer Aldayel and Walid Magdy. Your stance is exposed! analysing possible factors for stance detection on social media. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–20, 2019.

[16] Abeer AlDayel and Walid Magdy. Stance detection on social media: State of the art and trends. *Information Processing & Management*, 58(4):102597, 2021.

[17] Pranav Anand, Marilyn Walker, Rob Abbott, Jean E Fox Tree, Robeson Bowmani, and Michael Minor. Cats rule and dogs drool!: Classifying stance in online debate. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2.011)*, pages 1–9, 2011.

[18] Aymé Arango, Jorge Pérez, and Barbara Poblete. Hate speech detection is not as easy as you may think: A closer look at model validation (extended version). *Information Systems*, page 101584, 2020.

[19] Kathya Araujo, Patricia Poblete, M Alejandra Norambuena Montiglio, and Gabriel Valdés Echenique. Hilos tensados. para leer el octubre chileno. *Santiago: Editorial USACH*, 2019.

[20] Mohammad Assar. Guide to sanitation in natural disasters. 1971.

[21] Icek Azjen. Understanding attitudes and predicting social behavior. *Englewood Cliffs*, 1980.

[22] Christopher A Bail. The cultural environment: Measuring culture with big data. *Theory and Society*, 43(3-4):465–482, 2014.

[23] Daniel Baker and Karen Refsgaard. Institutional development and scale matching in disaster response management. *Ecological Economics*, 63(2-3):331–343, 2007.

[24] Pablo Barberá, John T Jost, Jonathan Nagler, Joshua A Tucker, and Richard Bonneau. Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological science*, 26(10):1531–1542, 2015.

[25] Andrew Baum, Raymond Fleming, and Laura M Davidson. Natural disaster and technological catastrophe. *Environment and Behavior*, 15(3):333–354, 1983.

[26] Adrian Benton and Mark Dredze. Using author embeddings to improve tweet stance classification. In *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text*, pages 184–194, 2018.

[27] Alessandro Bessi, Fabio Petroni, Michela Del Vicario, Fabiana Zollo, Aris Anagnostopoulos, Antonio Scala, Guido Caldarelli, and Walter Quattrociocchi. Homophily and polarization in the age of misinformation. *The European Physical Journal Special Topics*, 225(10):2047–2059, 2016.

[28] Alessandro Bessi, Fabiana Zollo, Michela Del Vicario, Michelangelo Puliga, Antonio Scala, Guido Caldarelli, Brian Uzzi, and Walter Quattrociocchi. Users polarization on facebook and youtube. *PloS one*, 11(8):e0159641, 2016.

[29] Marco Bitschnau, Leslie Ader, Didier Ruedin, and Gianni D'Amato. Politicising immigration in times of crisis: empirical evidence from switzerland. *Journal of Ethnic and Migration Studies*, 47(17):3864–3890, 2021.

[30] Marco Bitschnau, Dennis Lichtenstein, and Birte Fähnrich. The "refugee crisis" as an opportunity structure for right-wing populist social movements: The case of pegida. *Studies in Communication Sciences*, pages 1–13, 2021.

[31] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

[32] John Adrian Bondy, Uppaluri Siva Ramachandra Murty, et al. *Graph theory with applications*, volume 290. Macmillan London, 1976.

[33] Javier Borge-Holthoefer, Walid Magdy, Kareem Darwish, and Ingmar Weber. Content and network dynamics behind egyptian political polarization on twitter. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 700–711, 2015.

[34] Peter F Brown, Vincent J Della Pietra, Peter V Desouza, Jennifer C Lai, and Robert L Mercer. Class-based n-gram models of natural language. *Computational linguistics*, 18 (4):467–480, 1992.

[35] Jane Bullock, George Haddow, and Damon P Coppola. *Introduction to emergency management*. Butterworth-Heinemann, 2017.

[36] Mark A Cameron, Robert Power, Bella Robinson, and Jie Yin. Emergency situation awareness from twitter for crisis management. In *Proceedings of the 21st International Conference on World Wide Web*, pages 695–698. ACM, 2012.

[37] Cornelia Caragea, Nathan McNeese, Anuj Jaiswal, Greg Traylor, Hyun-Woo Kim, Prasenjit Mitra, Dinghao Wu, Andrea H Tapia, Lee Giles, Bernard J Jansen, et al. Classifying text messages for the haiti earthquake. In *Proceedings of the 8th international conference on information systems for crisis response and management (ISCRAM2011)*. Citeseer, 2011.

[38] Lowell Juilliard Carr. Disaster and the sequence-pattern concept of social change. *American Journal of Sociology*, 38(2):207–218, 1932.

[39] Carlos Castillo. *Big crisis data: social media in disasters and time-critical situations*. Cambridge University Press, 2016.

[40] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 675–684, 2011.

[41] Erica Chenoweth. *Civil Resistance: What Everyone Needs to Know®*. Oxford University Press, 2021.

[42] Dennis Chong and James N Druckman. Framing theory. *Annu. Rev. Polit. Sci.*, 10: 103–126, 2007.

[43] Pulso Ciudadano. Pulso ciudadano: Crisis en chile, 2019. URL `https://chile.activasite.com/wp-content/uploads/2019/10/Pulso-Ciudadano-Crisis-en-Chile.pdf`.

[44] Aaron Clauset, Mark EJ Newman, and Cristopher Moore. Finding community structure in very large networks. *Physical review E*, 70(6):066111, 2004.

[45] Alfredo Cobo, Denis Parra, and Jaime Navón. Identifying relevant messages in a twitter-based citizen channel for natural disaster situations. In *Proceedings of the 24th International Conference on World Wide Web*, pages 1189–1194, 2015.

[46] Mauro Coletto, Kiran Garimella, Aristides Gionis, and Claudio Lucchese. A motif-based approach for identifying controversy. In *Eleventh International AAAI Conference on Web and Social Media*, 2017.

[47] Bernard Comrie et al. *The world's major languages*. Routledge London, UK, 1987.

[48] Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*, 2017.

[49] W Timothy Coombs. Protecting organization reputations during a crisis: The development and application of situational crisis communication theory. *Corporate reputation review*, 10(3):163–176, 2007.

[50] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20 (3):273–297, 1995.

[51] Stefano Cresci, Maurizio Tesconi, Andrea Cimino, and Felice Dell'Orletta. A linguistically-driven approach to cross-event damage assessment of natural disasters from social media messages. In *Proceedings of the 24th International Conference on World Wide Web*, pages 1195–1200, 2015.

[52] Kareem Darwish, Walid Magdy, Afshin Rahimi, Timothy Baldwin, and Norah Abokhodair. Predicting online islamophopic behavior after# parisattacks. *The Journal of Web Science*, 4, 2018.

[53] Kareem Darwish, Peter Stefanov, Michaël Aupetit, and Preslav Nakov. Unsupervised user stance detection on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 141–152, 2020.

[54] Web Archive Dataverse Scholar Portal. 2016 Fort McMurray Wildfire. `https://dataverse.scholarsportal.info/dataset.xhtml?persistentId=hdl:10864/12033`, 2019. [accessed May 29, 2020].

[55] Dorottya Demszky, Nikhil Garg, Rob Voigt, James Zou, Jesse Shapiro, Matthew Gentzkow, and Dan Jurafsky. Analyzing polarization in social media: Method and application to tweets on 21 mass shootings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2970–3005, 2019.

[56] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[57] Paul DiMaggio, Manish Nag, and David Blei. Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of us government arts funding. *Poetics*, 41(6):570–606, 2013.

[58] Mario Fernando Garcés Durán. *Estallido social y una nueva Constitución para Chile.* Lom Ediciones, 2020.

[59] Laurence Echard. *The Gazetteer's: Or Newsman's Interpreter: Being a Geographical Index. Of All the Considerable Cities, Patriachships,... in Europe.... The Seventh Edition, Corrected and Very Much Enlarged with the Addition of a Table of Births, Marriages, &c. of All Kings,... of Europe. By Lawrence Eachard,...* John Nicholson, and Samuel Ballard, 1704.

[60] The Economic, Social Commission for Asia, and the Pacific. Asia-pacific disaster report 2021, 2021. URL `https://www.unescap.org/sites/default/d8files/knowledge-products/Asia-Pacific%20Disaster%20Report%202021-Full%20report.pdf`.

[61] Jacob Eisenstein. Natural language processing, 2018.

[62] Heba Elfardy and Mona Diab. Cu-gwu perspective at semeval-2016 task 6: Ideological stance detection in informal text. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 434–439, 2016.

[63] Mica R Endsley and Daniel J Garland. *Situation awareness analysis and measurement.* CRC press, 2000.

[64] Robert M Entman. Framing: Towards clarification of a fractured paradigm. *McQuail's reader in mass communication theory*, pages 390–397, 1993.

[65] Jorge Fábrega. ¿despertó chile? no todavía, 2020. URL `https://www.latercera.com/la-tercera-domingo/noticia/que-tan-polarizados-estamos/EKIQPDVZTZAQFK2VTYK5GZBUIU/`.

[66] Niall Ferguson. *Doom: The politics of catastrophe*. Penguin UK, 2021.

[67] Michela Ferron and Paolo Massa. Psychological processes underlying wikipedia representations of natural and manmade disasters. In *Proceedings of the eighth annual international symposium on wikis and open collaboration*, pages 1–10, 2012.

[68] Alam Firoj, Muhammad Imran, and Ferda Ofli. CrisisDPS: Crisis data processing services. In *Proceedings of the 16th International Conference on Information Systems for Crisis Response and Management*, pages 719–733. ISCRAM Association, 2019.

[69] John R Firth. A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*, 1957.

[70] Henry W Fischer. *Response to disaster: Fact versus fiction & its perpetuation: The sociology of disaster*. University press of America, 1998.

[71] Veny Amilia Fitri, Rachmadita Andreswari, and Muhammad Azani Hasibuan. Sentiment analysis of social media twitter with case of anti-lgbt campaign in indonesia using naïve bayes, decision tree, and random forest algorithm. *Procedia Computer Science*, 161:765–772, 2019.

[72] Centre for Research on the Epidemiology of Disasters. The international disaster database. URL `https://www.emdat.be/`.

[73] Bruno Fuentes. El tridente del fascismo se asoma sobre chile, 2020. URL `https://www.revistadefrente.cl/el-tridente-del-fascismo-se-asoma-sobre-chile/`.

[74] Mario Garcés. October 2019: Social uprising in neoliberal chile. *Journal of Latin American Cultural Studies*, 28(3):483–491, 2019.

[75] Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644, 2018.

[76] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. Quantifying controversy on social media. *ACM Transactions on Social Computing*, 1(1):1–27, 2018.

[77] Venkata Rama Kiran Garimella and Ingmar Weber. A long-term analysis of polarization on twitter. In *Eleventh international AAAI conference on web and social media*, 2017.

[78] Paolo Gerbaudo. The indignant citizen: anti-austerity movements in southern europe and the anti-oligarchic reclaiming of citizenship. *Social movement studies*, 16(1):36–50, 2017.

[79] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[80] Michael Gottlieb and Sean Dyer. Information and disinformation: social media in the covid-19 crisis. *Academic emergency medicine*, 27(7):640, 2020.

[81] Eduardo Graells-Garrido, Ricardo Baeza-Yates, and Mounia Lalmas. Every colour you are: Stance prediction and turnaround in controversial issues. In *12th ACM Conference on Web Science*, pages 174–183, 2020.

[82] Eduardo Graells-Garrido, Ricardo Baeza-Yates, and Mounia Lalmas. Representativeness of abortion legislation debate on twitter: A case study in argentina and chile. In *Companion Proceedings of the Web Conference 2020*, pages 765–774, 2020.

[83] David Graf, Werner Retschitzegger, Wieland Schwinger, Birgit Pröll, and Elisabeth Kapsammer. Cross-domain informativeness classification for disaster situations. In *Proceedings of the 10th International Conference on Management of Digital EcoSystems*, pages 183–190. ACM, 2018.

[84] Mark Graham, Scott A Hale, and Devin Gaffney. Where in the world are you? geolocation and language identification in twitter. *The Professional Geographer*, 66(4): 568–578, 2014.

[85] Jon Green, Jared Edgerton, Daniel Naftel, Kelsey Shoub, and Skyler J Cranmer. Elusive consensus: Polarization in elite communication on the covid-19 pandemic. *Science advances*, 6(28):eabc2717, 2020.

[86] Shane Greenstein and Feng Zhu. Is wikipedia biased? *American Economic Review*, 102(3):343–48, 2012.

[87] Hannah Greving, Aileen Oeberst, Joachim Kimmerle, and Ulrike Cress. Emotional content in wikipedia articles on negative man-made and nature-made events. *Journal of Language and Social Psychology*, 37(3):267–287, 2018.

[88] Hannah Greving, Ulrike Cress, and Joachim Kimmerle. Anger in wikipedia: Perceived intentionality and threat appraisal as mediators of anger about man-made attacks. *Journal of Applied Social Psychology*, 49(2):99–116, 2019.

[89] John D Griffin, Chad Kiewiet de Jonge, and Vania Ximena Velasco-Guachalla. Deprivation in the midst of plenty: Citizen polarization and political protest. *British Journal of Political Science*, 51(3):1080–1096, 2021.

[90] Frederic Guerrero-Solé. Community detection in political discussions on twitter: An application of the retweet overlap network method to the catalan process toward independence. *Social science computer review*, 35(2):244–261, 2017.

[91] Debby Guha-Sapir, Femke Vos, Regina Below, and Sylvain Ponserre. Annual disaster statistical review 2011: the numbers and trends. 2012.

[92] Samrat Gupta, Gaurav Jain, and Amit Anand Tiwari. Investigating the dynamics of polarization in online discourse during covid-19 pandemic. In *Conference on e-Business, e-Services and e-Society*, pages 704–709. Springer, 2021.

[93] Jheser Guzman and Barbara Poblete. On-line relevant anomaly detection in the twitter stream: an efficient bursty keyword detection model. In *Proceedings of the ACM SIGKDD workshop on outlier detection and description*, pages 31–39, 2013.

[94] P Sol Hart, Sedona Chinn, and Stuart Soroka. Politicization and polarization in covid-19 news coverage. *Science Communication*, 42(5):679–697, 2020.

[95] Marwan Hassani and Thomas Seidl. Using internal evaluation measures to validate the quality of diverse stream clustering algorithms. *Vietnam Journal of Computer Science*, 4(3):171–183, 2017.

[96] Brent Hecht, Lichan Hong, Bongwon Suh, and Ed H Chi. Tweets from justin bieber's heart: the dynamics of the location field in user profiles. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 237–246. ACM, 2011.

[97] Libby Hemphill, Aron Culotta, and Matthew Heston. Framing in social media: How the us congress uses twitter hashtags to frame political issues. *Available at SSRN 2317335*, 2013.

[98] Miriam Henríquez. La actual constitución no es compatible con las demandas sociales, 2020. URL `https://www.ciperchile.cl/2020/02/04/la-actual-constitucion-no-es-compatible-con-las-demandas-sociales/`.

[99] Julia Hirschberg and Christopher D Manning. Advances in natural language processing. *Science*, 349(6245):261–266, 2015.

[100] Sameera Horawalavithana, Kin Wai Ng, and Adriana Iamnitchi. Drivers of polarized discussions on twitter during venezuela political crisis. In *13th ACM Web Science Conference 2021*, pages 205–214, 2021.

[101] Muhammad Imran, Shady Elbassuoni, Carlos Castillo, Fernando Diaz, and Patrick Meier. Extracting information nuggets from disaster-related messages in social media. In *ISCRAM*, 2013.

[102] Muhammad Imran, Shady Elbassuoni, Carlos Castillo, Fernando Diaz, and Patrick Meier. Practical extraction of disaster-relevant information from social media. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 1021–1024. ACM, 2013.

[103] Muhammad Imran, Carlos Castillo, Ji Lucas, Patrick Meier, and Sarah Vieweg. Aidr: Artificial intelligence for disaster response. In *Proceedings of the companion publication of the 23rd international conference on World wide web companion*, pages 159–162. International World Wide Web Conferences Steering Committee, 2014.

[104] Muhammad Imran, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. Processing social media messages in mass emergency: A survey. *ACM Computing Surveys (CSUR)*, 47(4):67, 2015.

[105] Muhammad Imran, Prasenjit Mitra, and Carlos Castillo. Twitter as a lifeline: Human-annotated twitter corpora for nlp of crisis-related messages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1638–1643, 2016.

[106] Muhammad Imran, Prasenjit Mitra, and Jaideep Srivastava. Cross-language domain adaptation for classifying crisis-related short messages. In *13th International Conference on Information Systems for Crisis Response and Management, ISCRAM 2016*. Information Systems for Crisis Response and Management, ISCRAM, 2016.

[107] Insurance Information Instintute. Facts + statistics: Man-made disasters. `https://www.iii.org/fact-statistic/facts-statistics-man-made-disasters`, 2022. Accessed: 2022-06-22.

[108] Daniel J Isenberg. Group polarization: A critical review and meta-analysis. *Journal of personality and social psychology*, 50(6):1141, 1986.

[109] Michael I Jordan and Tom M Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260, 2015.

[110] Charles Kadushin. *Understanding social networks: Theories, concepts, and findings*. Oxford university press, 2012.

[111] Janani Kalyanam, Mauricio Quezada, Barbara Poblete, and Gert Lanckriet. Prediction and characterization of high-activity events in social media triggered by real-world news. *PloS one*, 11(12):e0166694, 2016.

[112] Krzysztof Kaniasty and Fran H Norris. Social support in the aftermath of disasters, catastrophes, and acts of terrorism: Altruistic, overwhelmed, uncertain, antagonistic, and patriotic communities. *Bioterrorism: Psychological and public health interventions*, 3:200–229, 2004.

[113] P Karthika, R Murugeswari, and R Manoranjithem. Sentiment analysis of social media network using random forest algorithm. In *2019 IEEE international conference on intelligent techniques in control, optimization and signal processing (INCOS)*, pages 1–5. IEEE, 2019.

[114] Leonard Kaufman and Peter J Rousseeuw. *Finding groups in data: an introduction to cluster analysis*, volume 344. John Wiley & Sons, 2009.

[115] Anastasia Kavada. Creating the collective: social media, the occupy movement and its constitution as a collective actor. In *Protest Technologies and Media Revolutions*. Emerald Publishing Limited, 2020.

[116] Prashant Khare, Grégoire Burel, Diana Maynard, and Harith Alani. Cross-lingual classification of crisis data. In *International Semantic Web Conference*, pages 617–633. Springer, 2018.

[117] Ashiqur R KhudaBukhsh, Rupak Sarkar, Mark S Kamlet, and Tom M Mitchell. We don't speak the same language: Interpreting polarization through machine translation. *arXiv preprint arXiv:2010.02339*, 2020.

[118] Ashiqur R KhudaBukhsh, Rupak Sarkar, Mark S Kamlet, and Tom Mitchell. We don't speak the same language: Interpreting polarization through machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14893–14901, 2021.

[119] Beata Beigman Klebanov, Eyal Beigman, and Daniel Diermeier. Vocabulary choice as an indicator of perspective. In *Proceedings of the ACL 2010 conference short papers*, pages 253–257, 2010.

[120] TUULI-MARJA KLEINER. Public opinion polarisation and protest behaviour. *European Journal of Political Research*, 57(4):941–962, 2018.

[121] Petra Kralj Novak, Jasmina Smailović, Borut Sluban, and Igor Mozetič. Sentiment of emojis. *PloS one*, 10(12):e0144296, 2015.

[122] Anna Kruspe, Jens Kersten, and Friederike Klan. Detection of informative tweets in crisis events. *Natural Hazards and Earth System Sciences Discussions*, pages 1–18, 2020.

[123] Anna Kruspe, Jens Kersten, and Friederike Klan. Detection of informative tweets in crisis events. *Natural Hazards and Earth System Sciences (NHESS)*, 2021.

[124] Yury Kryvasheyeu, Haohui Chen, Nick Obradovich, Esteban Moro, Pascal Van Hentenryck, James Fowler, and Manuel Cebrian. Rapid assessment of disaster damage using social media activity. *Science advances*, 2(3):e1500779, 2016.

[125] Shamanth Kumar, Geoffrey Barbier, Mohammad Ali Abbasi, and Huan Liu. Tweettracker: An analysis tool for humanitarian and disaster relief. In *ICWSM*, 2011.

[126] Mucahid Kutlu, Kareem Darwish, Cansin Bayrak, Ammar Rashed, and Tamer Elsayed. Embedding-based qualitative analysis of polarization in turkey. *arXiv preprint arXiv:1909.10213*, 2019.

[127] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. AcM, 2010.

[128] Mirko Lai, Viviana Patti, Giancarlo Ruffo, and Paolo Rosso. Stance evolution and twitter interactions in an italian political debate. In *International Conference on Applications of Natural Language to Information Systems*, pages 15–27. Springer, 2018.

[129] Mirko Lai, Alessandra Teresa Cignarella, Delia Irazú Hernández Farías, Cristina Bosco, Viviana Patti, and Paolo Rosso. Multilingual stance detection in social media political debates. *Computer Speech & Language*, 63:101075, 2020.

[130] Jörg Landthaler, Bernhard Waltl, Dominik Huth, Daniel Braun, Christoph Stocker, Thomas Geiger, and Florian Matthes. Extending thesauri using word embeddings and the intersection method. *ASAIL@ ICAIL*, 8(1):112–119, 2017.

[131] Chang Li, Aldo Porco, and Dan Goldwasser. Structured representation learning for online debate stance prediction. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3728–3739, 2018.

[132] Hongmin Li, D Caragea, X Li, and Cornelia Caragea. Comparison of word embeddings and sentence encodings as generalized representations for crisis tweet classification tasks. *en. In: New Zealand*, page 13, 2018.

[133] Jessica Lin, Michail Vlachos, Eamonn Keogh, and Dimitrios Gunopulos. Iterative incremental clustering of time series. In *International Conference on Extending Database Technology*, pages 106–122. Springer, 2004.

[134] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[135] Jacopo Longhini, Claudio Rossi, Claudio Casetti, and Federico Angaramo. A language-agnostic approach to exact informative tweets during emergency situations. In *2017 IEEE International Conference on Big Data (Big Data)*, pages 3739–3475. IEEE, 2017.

[136] Ramón López and Sebastian J Miller. Chile: the unbearable burden of inequality. *World Development*, 36(12):2679–2695, 2008.

[137] Valerio Lorini, Carlos Castillo, Francesco Dottori, Milan Kalas, Domenico Nappo, and Peter Salamon. Integrating social media into a pan-european flood awareness system: A multilingual approach.

[138] Alessandro Lovari and Shannon A Bowen. Social media in disaster communication: A case study of strategies, barriers, and ethical implications. *Journal of Public Affairs*, 20(1):e1967, 2020.

[139] Nicola Lunardon, Giovanna Menardi, and Nicola Torelli. Rose: A package for binary imbalanced learning. *R Journal*, 6(1), 2014.

[140] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967.

[141] Martin Maiden and Cecilia Robustelli. *A reference grammar of modern Italian.* Routledge, 2014.

[142] Mykola Makhortykh and Maryna Sydorova. Social media and visual framing of the conflict in eastern ukraine. *Media, War & Conflict*, 10(3):359–381, 2017.

[143] Jazmine Maldonado, Jheser Guzman, and Barbara Poblete. A lightweight and real-time worldwide earthquake detection and monitoring system based on citizen sensors. In *Proceedings of the Fifth Conference of Human Computation and Crowdsourcing*, pages 137–146. AAAI, 2017.

[144] Kanti V Mardia. Measures of multivariate skewness and kurtosis with applications. *Biometrika*, 57(3):519–530, 1970.

[145] Doug McAdam. Introduction: Opportunities, mobilizing structures, and framing processes-toward a synthetic, comparative perspective on social movements. *Comparative perspectives on social movements*, pages 1–20, 1996.

[146] Dawn McCaffrey and Jennifer Keys*. Competitive framing processes in the abortion debate: Polarization-vilification, frame saving, and frame debunking. *Sociological Quarterly*, 41(1):41–61, 2000.

[147] Maxwell E McCombs and Donald L Shaw. The evolution of agenda-setting research: Twenty-five years in the marketplace of ideas. *Journal of communication*, 43(2):58–67, 1993.

[148] Andrew J McMinn, Yashar Moshfeghi, and Joemon M Jose. Building a large-scale corpus for evaluating event detection on twitter. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 409–418, 2013.

[149] Julia Mendelsohn, Ceren Budak, and David Jurgens. Modeling framing in immigration discourse on social media. *arXiv preprint arXiv:2104.06443*, 2021.

[150] Marcelo Mendoza, Barbara Poblete, and Carlos Castillo. Twitter under crisis: Can we trust what we rt? In *Proceedings of the first workshop on social media analytics*, pages 71–79, 2010.

[151] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[152] Theophano Mitsa. *Temporal data mining*. CRC Press, 2010.

[153] Tun Lin Moe and Pairote Pathranarakul. An integrated approach to natural disaster management: public project management and its critical success factors. *Disaster Prevention and Management: An International Journal*, 2006.

[154] Tun Lin Moe, Fritz Gehbauer, Stefan Senitz, and Marc Mueller. Balanced scorecard for natural disaster management projects. *Disaster Prevention and Management: An International Journal*, 2007.

[155] Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, pages 31–41, 2016.

[156] Saif M Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. Stance and sentiment in tweets. *ACM Transactions on Internet Technology (TOIT)*, 17(3):1–23, 2017.

[157] Virginia Morini, Laura Pollacci, and Giulio Rossetti. Capturing political polarization of reddit submissions in the trump era. In *SEBD*, pages 80–87, 2020.

[158] Fred Morstatter, Jürgen Pfeffer, Huan Liu, and Kathleen M Carley. Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose. In *ICWSM*, 2013.

[159] Federico Navarro and Carlos Tromben. " estamos en guerra contra un enemigo poderoso, implacable": los discursos de sebastián piñera y la revuelta popular en chile. *Literatura y lingüística*, (40):295–324, 2019.

[160] Thomas E Nelson, Zoe M Oxley, and Rosalee A Clawson. Toward a psychology of framing effects. *Political behavior*, 19(3):221–246, 1997.

[161] Mark EJ Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical review E*, 74(3):036104, 2006.

[162] BBC News. Protestas en chile: La tortura, los malos tratos en comisarías y la violencia con connotación sexual son preocupantes, 2019. URL `https://www.bbc.com/mundo/noticias-america-latina-50178678`.

[163] Xiaodong Ning, Lina Yao, Xianzhi Wang, and Boualem Benatallah. Calling for response: Automatically distinguishing situation-aware tweets during crises. In *International Conference on Advanced Data Mining and Applications*, pages 195–208. Springer, 2017.

[164] Department of Humanitarian Affairs. Internationally agreed glossary of basic terms related to disaster management. URL `https://reliefweb.int/sites/reliefweb.int/files/resources/004DFD3E15B69A67C1256C4C006225C2-dha-glossary-1992.pdf`.

[165] International Federation of Red Cross and Red Crescent Societies. World disasters report 2020, 2020. URL `https://www.ifrc.org/sites/default/files/2021-05/20201116_WorldDisasters_Full.pdf`.

[166] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. Crisislex: A lexicon for collecting and filtering microblogged communications in crises. In *ICWSM*, 2014.

[167] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. Crisislex: A lexicon for collecting and filtering microblogged communications in crises. In *Eighth international AAAI conference on weblogs and social media*, 2014.

[168] Alexandra Olteanu, Sarah Vieweg, and Carlos Castillo. What to expect when the unexpected happens: Social media communications across crises. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 994–1009. ACM, 2015.

[169] Oluwafemi Oriola and Eduan Kotzé. Evaluating machine learning techniques for detecting offensive and hate speech in south african tweets. *IEEE Access*, 8:21496–21509, 2020.

[170] Jason Osborne. Improving your data transformations: Applying the box-cox transformation. *Practical Assessment, Research, and Evaluation*, 15(1):12, 2010.

[171] Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A Smith. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 380–390, 2013.

[172] Leysia Palen and Sophia B Liu. Citizen communications in crisis: anticipating a future of ict-supported public participation. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 727–736. ACM, 2007.

[173] Sergey Pashakhin. Topic modeling for frame analysis of news media. *Proceedings of the AINL FRUCT*, pages 103–105, 2016.

[174] V Pekar, J Binner, H Najafi, and C Hale. Selecting classification features for detection of mass emergency events on social media. In *Proceedings of the International Conference on Security and Management (SAM)*, page 192, 2016.

[175] Vanessa Peña-Araya, Mauricio Quezada, Barbara Poblete, and Denis Parra. Gaining historical and international relations insights from social media: spatio-temporal real-world news analysis using twitter. *EPJ Data Science*, 6:1–35, 2017.

[176] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[177] Mark Edward Phillips. Dallas Police Shooting Twitter Dataset. `https://digital.library.unt.edu/ark:/67531/metadc991469/`, 2019. [accessed May 29, 2020, University of North Texas Libraries, UNT Digital Library, https://digital.library.unt.edu].

[178] PIpsos. Movilizaciones sociales en chile, 2019. URL `https://www.ipsos.com/es-cl/movilizaciones-sociales-en-chile`.

[179] Barbara Poblete, Jheser Guzmán, Jazmine Maldonado, and Felipe Tobar. Robust detection of extreme events using twitter: worldwide earthquake monitoring. *IEEE Transactions on Multimedia*, 20(10):2551–2561, 2018.

[180] The Washington Post. Why is the world protesting so much? a new study claims to have some answers. `https://www.washingtonpost.com/world/2021/11/04/protests-global-study/`, 2022. Accessed: 2022-06-22.

[181] Hemant Purohit, Carlos Castillo, Muhammad Imran, and Rahul Pandey. Social-eoc: Serviceability model to rank social media requests for emergency operation centers. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 119–126. IEEE, 2018.

[182] Paz Radovic. ¿qué tan polarizados estamos?, 2020. URL `https://www.latercera.com/la-tercera-domingo/noticia/que-tan-polarizados-estamos/EKIQPDVZTZAQFK2VTYK5GZBUIU/`.

[183] Halim Rane and Sumra Salem. Social media, social movements and the diffusion of ideas in the arab uprisings. *Journal of international communication*, 18(1):97–111, 2012.

[184] Joseph P Reser. The experience of natural disasters: Psychological perspectives and understandings. In *International perspectives on natural disasters: Occurrence, mitigation, and consequences*, pages 369–384. Springer, 2007.

[185] Christian Reuter and Marc-André Kaufhold. Fifteen years of social media in emergencies: a retrospective review and future directions for crisis informatics. *Journal of Contingencies and Crisis Management*.

[186] Karl Rohe, Sourav Chatterjee, and Bin Yu. Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, 39(4):1878–1915, 2011.

[187] Arnon Rosenthal and José A Pino. A generalized algorithm for centrality problems on trees. *Journal of the ACM (JACM)*, 36(2):349–361, 1989.

[188] Giulio Rossetti, Letizia Milli, and Rémy Cazabet. Cdlib: a python library to extract, compare and evaluate communities from complex networks. *Applied Network Science*, 4(1):1–26, 2019.

[189] Martin Rosvall and Carl T Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the national academy of sciences*, 105(4): 1118–1123, 2008.

[190] Cinthia Mabel Sánchez Macías. Transfer learning for the multilingual and multi-domain classification of messages relating to crises.

[191] Peter M. Sandman and Jody Lanard. Explaining and proclaiming uncertainty: Risk communication lessons from germany's deadly e. coli outbreak. `http://www.psandman.com/col/GermanEcoli.htm`, 2011. Accessed: 2022-06-22.

[192] Hernan Sarmiento and Barbara Poblete. Crisis communication: a comparative study of communication patterns across crisis events in social media. In *Proceedings of the 36th Annual ACM Symposium on Applied Computing*, pages 1711–1720, 2021.

[193] Indira Sen, Fabian Flöck, and Claudia Wagner. On the reliability and validity of detecting approval of political actors in tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1413–1426, 2020.

[194] Saif Shahin. Affective polarization of a protest and a counterprotest: Million maga march v. million moron march. *American Behavioral Scientist*, page 00027642221091212, 2022.

[195] Stella Shen and Michael Shaw. Managing coordination in emergency response systems with information technologies. *AMCIS 2004 Proceedings*, page 252, 2004.

[196] David A Snow, Robert D Benford, et al. Ideology, frame resonance, and participant mobilization. *International social movement research*, 1(1):197–217, 1988.

[197] Nicolás M Somma. Power cages and the october 2019 uprising in chile. *Social Identities*, pages 1–14, 2021.

[198] Nicolás M Somma, Matías Bargsted, and Felipe Sánchez. Protest issues and political inequality in latin america. *American Behavioral Scientist*, 64(9):1299–1323, 2020.

[199] Nicolás M Somma, Matías Bargsted, Rodolfo Disi Pavlic, and Rodrigo M Medel. No water in the oasis: the chilean spring of 2019–2020. *Social Movement Studies*, 20(4): 495–502, 2021.

[200] Statista. Number of cumulative cases of coronavirus (covid-19) worldwide from january 22, 2020 to june 26, 2022, by day. `https://www.statista.com/statistics/1103040/cumulative-coronavirus-covid19-cases-number-worldwide-by-day/`, 2022. Accessed: 2022-06-22.

[201] Statista. Insured losses caused by man-made catastrophes worldwide from 1990 to 2020. `https://www.statista.com/statistics/281059/insured-losses-from-man-made-catastrophes-worldwide/`, 2022. Accessed: 2022-06-22.

[202] Peter Stefanov, Kareem Darwish, Atanas Atanasov, and Preslav Nakov. Predicting the topical stance and political leaning of media using tweets. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 527–537, 2020.

[203] Kevin Stowe, Michael J Paul, Martha Palmer, Leysia Palen, and Kenneth Anderson. Identifying and categorizing disaster-related tweets. In *Proceedings of The Fourth International Workshop on Natural Language Processing for Social Media*, pages 1–6, 2016.

[204] Cass R Sunstein. The law of group polarization. *University of Chicago Law School, John M. Olin Law & Economics Working Paper*, (91), 1999.

[205] Chris Sweeney and Maryam Najafian. A transparent framework for evaluating unintended demographic bias in word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1662–1667, 2019.

[206] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. Data mining cluster analysis: basic concepts and algorithms. *Introduction to data mining*, 487:533, 2013.

[207] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to data mining*. Pearson Education India, 2016.

[208] Yla R Tausczik and James W Pennebaker. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54, 2010.

[209] Johnny Torres and Carmen Vaca. Cross-lingual perspectives about crisis-related conversations on Twitter. In *Companion Proceedings of The 2019 World Wide Web Conference*, pages 255–261. ACM, 2019.

[210] John Twigg et al. Disaster risk reduction. 2015.

[211] Krishna S Vatsa. Risk, vulnerability, and asset-based approach to disaster risk management. *International Journal of Sociology and Social Policy*, 2004.

[212] Francisco Velásquez. El apruebo en las comunas donde ganó el rechazo, reflexionando sobre los privilegios, 2020. URL `https://www.revistadefrente.cl/el-tridente-del-fascismo-se-asoma-sobre-chile/`.

[213] Sarah Vieweg, Amanda L Hughes, Kate Starbird, and Leysia Palen. Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1079–1088, 2010.

[214] Ramon Villa-Cox, Ashiqur R KhudaBukhsh, Kathleen M Carley, et al. Exploring polarization of users behavior on twitter during the 2019 south american protests. *arXiv preprint arXiv:2104.05611*, 2021.

[215] Jia Wang, Yungang Feng, Elham Naghizade, Lida Rashidi, Kwan Hui Lim, and Kate Lee. Happiness is a choice: sentiment and activity-aware location recommendation. In *Companion Proceedings of the The Web Conference 2018*, pages 1401–1405, 2018.

[216] Stanley Wasserman, Katherine Faust, et al. Social network analysis: Methods and applications. 1994.

[217] Ingmar Weber, Venkata R Kiran Garimella, and Alaa Batayneh. Secular vs. islamist polarization in egypt on twitter. In *Proceedings of the 2013 IEEE/ACM international conference on advances in social networks analysis and mining*, pages 290–297, 2013.

[218] Melvin Stanley Whitley. *Spanish/English contrasts: A course in Spanish linguistics.* Georgetown University Press, 2002.

[219] Jaewon Yang and Jure Leskovec. Defining and evaluating network communities based on ground-truth. *Knowledge and Information Systems*, 42(1):181–213, 2015.

[220] Jie Yin, Sarvnaz Karimi, Bella Robinson, and Mark Cameron. Esa: emergency situation awareness via microbloggers. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 2701–2703. ACM, 2012.

[221] Jie Yin, Sarvnaz Karimi, and John Lingad. Pinpointing locational focus in microblogs. In *Proceedings of the 2014 Australasian Document Computing Symposium*, page 66. ACM, 2014.

[222] Tuukka Ylä-Anttila, Veikko Eranti, and Anna Kukkonen. Topic modeling as a method for frame analysis: Data mining the climate change debate in india and the usa. 2018.

[223] Tuukka Ylä-Anttila, Veikko Eranti, and Anna Kukkonen. Topic modeling for frame analysis: A study of media debates on climate change in india and usa. *Global Media and Communication*, page 17427665211023984, 2020.

[224] Robert B Zajonc. Feeling and thinking: Preferences need no inferences. *American psychologist*, 35(2):151, 1980.

[225] Aston Zhang, Zachary C Lipton, Mu Li, and Alexander J Smola. Dive into deep learning. *arXiv preprint arXiv:2106.11342*, 2021.

[226] Zhi-Hua Zhou. Three perspectives of data mining, 2003.