CHARACTERIZING AND MODELING CRISIS-RELATED CONVERSATIONS IN TWITTER

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF COMPUTER SCIENCE
AND THE COMMITTEE ON GRADUATE STUDIES
OF ESPOL UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Johnny Torres

May 2020

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Cristina Abad) Principal Co-Advisor

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Carmen Vaca) Principal Co-Advisor

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Luis Teran)

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Fabio Gonzalez)

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Denis Romero)

Approved for the ESPOL University Committee on Graduate Studies

Preface

...

Acknowledgments

...

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

On average, over a million earthquakes occur each year; of those more than one thousand register five points or more measured by the Richter scale [1]. Between 2001 and 2011, the death toll related to natural disasters was higher than 780 000, with 60% of these caused by earthquakes [16]. The global trends in urbanization [41] expose millions of people to the threat of natural disasters, thus increasing the need for rapid response and relief efforts. The advent of social media networks has opened the opportunity of harnessing the help of crowds on social networks for early disaster response [15]. For example, during mass emergency events (e.g., political crisis, elections, or natural disasters), crowdsourced data can rapidly provide news ahead of official or traditional media [25]. Citizens, being a self-organizing and collective intelligence force, can play an essential role with the support of social media services during relief efforts in emergencies due to natural disasters. Critical aspects need to be addressed on social media services to leverage public collaboration, such as analyzing quantity, quality, opinion drift, trustworthiness, and security of the information to become an effective tool [105].

Microblogging data, such as reporting live during natural disasters from the affected zones or neighboring cities, has been deemed unverifiable and untrustworthy by humanitarian relief organizations [130]. People are willing to provide help through activists and Non-governmental organizations (NGOs) because of government efforts often insufficient to organize resources and coordinate relief efforts in the early stages of the crisis events. However, several issues prevent social media platforms from being a useful tool in these scenarios. Current limitations include categorization, cross-lingual messages, report verification, automated report summarization, behavior prediction, scalability, and safety [42]. By harnessing the power of crowdsourced social media services (e.g., Twitter, Facebook, Google+), this work seeks to provide automated methods that can contribute to the relief efforts led by NGOs and activists. For instance, in Ecuador, after the 7.8 earthquakes last April, among

other initiatives, the twitter activist @KarlaMorales[1] asked for life straws to purify water that was very scarce in the affected areas. People started donating immediately, and international donors were searching for ways to transport life straws to Ecuador, avoiding additional costs. A couple of tweets from a conversation depict the situation:

KARLAMORALESR: Hay más de 50 life straws en Atlanta, DHL cobra tarifa regular por el envío ¿otra vía para que lleguen?[2].

ARIANNACEVALLOS: @KarlaMoralesR si logran enviarlos a Miami mañana sale un contenedor de Provex desde ahí[3].

Citizens are no longer passive information recipients, and they can build entire humanitarian assistance structures in hours. Following a natural disaster, local activists working in emergency relief operations can crowdsource, through social media, tasks related to water, sanitation, and hygiene, shelter organization, health, nutrition, local experts' transportation, among others. Additionally, each task usually involves additional logistics, such as transportation of goods (e.g., medicines, food, clothes, water) and local experts (e.g., doctors, chefs) to zones affected by natural disasters. Previous works on disaster management analyze tweets, related to natural disasters, individually, loosing useful context information such as the sequence of tweets in a conversation. Thus, applications and automated tools can help local volunteers and NGOs in relief efforts by gathering and analyzing user-generated data on social media.

1.2 Research Goals

The main goal is to develop a conversational model to help NGOs to cope with the overwhelming amount of data in the form of conversations, enabling citizens to contribute more efficiently during natural disasters. The specific research goals are to:

- Characterize crisis-related conversations.
- Design and evaluate a fine-grained taxonomy to classify crisis-related conversations.
- Design and evaluate cross-lingual models for crisis-related conversations.
- Design and evaluate semisupervised models for crisis-related conversations.
- Design and evaluate recommender models for users to join crisis-related conversations.

[1] https://twitter.com/KarlaMoralesR
[2] https://twitter.com/ariannacevallos/status/723234766653415425
[3] https://twitter.com/lapekepenia/status/723257439961927682

1.3 Research Questions

Most previous works focused only on analyzing individual sentences or phrases rather than whole conversations [59, 136, 111, 87, 17]. Broadly, this thesis tackles the problem of Characterizing and modeling crisis-related conversations occurring during natural disasters. This work seeks to answer the following research questions:

- Characterizing crisis-related conversations:
  - RQ1: What are the factors that ignite conversations?
  - RQ2: Does current taxonomies account for crisis-related conversations?
- Modeling of crisis-related conversations:
  - RQ3: How to deal with multiple languages that appear during a crisis event?
  - RQ4: Does conversational context help in downstream tweets classification tasks?
  - RQ5: How to leverage the massive amount of unlabeled social media data for supervised tasks?
  - RQ6: How to recommend users to join relevant conversations on social media?

1.4 Main Contributions

In this doctoral thesis, text data extracted from Twitter conversations regarding a natural disaster is analyzed and modelled. In doing so, contributions in different areas emerge:: novel Twitter conversation datasets, new tasks scenarios, machine learning models to automatically label the data. In this section, such contributions are presented in detail:

1.4.1 New task scenarios

Crisis-related conversational for contextual characterization: This thesis augments traditional crisis-related datasets with conversational context. The data collection process uses the Twitter REST API and Cassandra [74] as distributed storage. This work introduces a new conversational dataset about the earthquakes Ecuador 2016. The task scenario and datasets are available at Github[4].

New fine-grained classification for crisis-related events: This thesis investigates how to identify conversational acts (replies) to define outcomes that can be useful and relevant for crisis events. In doing that, this work introduces a new multi-label conversational act annotated corpus for

[4] https://github.com/johnnytorres/twconvch

disaster management on Twitter based on the novel, a fine-grained conversational act taxonomy. This new dataset enables single and multi-label classification tasks for individual tweets as well as the conversations outcome. A few-shot learning approach allows to handle the large number of labels with a long tail or few instances. The task scenario and dataset are available at Github[5].

Recommending user to join conversations: Recommendations are an essential component in downstream applications. As such, this thesis introduces a new large Twitter corpus (USERSEC) containing 10K conversations from three popular users from Ecuador. The dataset contains posts whose content is related to three domains: politics, sports, and crisis events activism. Then, this work proposes two novel recommendation tasks: recommending users to join a specific conversation and recommending a list of conversations to a particular user. The task scenario and dataset are available at Github[6].

1.4.2 Conversations characterizations

Relevance of factors igniting a conversation: This thesis investigates the importance of factors that influence the engagement of users in conversations and propose a language-independent model to identify seed tweets that have the potential to form conversations for different types of users. The results indicate that factors such as users' mentions and hashtags are predominant, and often generate a response (reply) from other users. Also, the retweet factor of an influencer usually triggers further interactions in conversations from users not in the followers graph of the user that started the conversation. This contribution has been peer-reviewed and published in the Proceedings of the Fourth IEEE/ACM International Conference on Big Data Computing, Applications, and Technologies [133].

Context-aware classification of crisis-related conversations: This thesis analyzes the context of crisis-related conversations on Twitter and how it can improve downstream NLP tasks such as classification and categorization, especially in the case of unlabeled tweets in the sample conversation. Through the augmentation of previous crisis-related datasets, this work analyzes labeled conversational tweets in the context of the conversations. The findings indicate that the conversational context improves downstream NLP classification tasks. Furthermore, the analysis determines what categories benefit the most from having a conversation context in classification tasks. This contribution is under review at the International AAAI Conference on Web and Social Media (ICWSM).

Cross-lingual crisis-related conversations This thesis investigates a multi-lingual scenario for crisis-related conversations, which is often the case of crisis events. Specifically, this work conducted

[5] https://github.com/johnnytorres/crisis_conv_crosslingual
[6] https://github.com/johnnytorres/recsys_twconv_s2s

a study of crisis-related tweets about the earthquake that occurred in Ecuador in April 2016 for cross-lingual tweets. To that end, this work introduces a new annotated dataset in both Spanish and English languages with approximately 8K tweets; half of them belong to conversations. Then, this research work evaluates a neural architecture to identify crisis-related tweets across multiple languages. Also, this work characterizes the conversations from locals and foreigners about the study case. This contribution has been peer-reviewed and published in the Companion Proceedings of The 2019 World Wide Web Conference [135].

1.4.3 New machine learning models

Leveraging unlabeled data for short-text classification: This thesis proposes a semi-supervised approach to learn how to categorize short-texts in a multi-label taxonomy using a small set of labeled data and leveraging the availability of large amounts of unlabeled data. Built upon a neural semi-supervised k-means clustering, the proposed new model modifies the normal objective function and adds a penalty term for labeled data. Then, it extended the neural semi-supervised clustering and applied it to multi-label settings. This contribution has been peer-reviewed and published in the Proceedings of the 2nd Workshop on Affective Content Analysis (AffCon 2019) co-located with Thirty-Third AAAI Conference on Artificial Intelligence (AAAI 2019) [132].

Recommending short-text conversations: This thesis proposes a neural learning architecture based on a sequence-to-sequence model to tackle the task of recommending conversations using the conversation context and the user's history. This research evaluates several types of recurrent neural networks that can be used in the proposed architecture and investigate their performance. This contribution has been peer-reviewed and published in the Journal Expert Systems with Applications [134].

1.5 Research outline

This thesis comprises of nine chapters. After a background chapter, there are six research chapters containing core contributions of this work plus a concluding chapter:

Chapter 2 - Background: This chapter introduces the background for the following chapters. The research in this thesis falls in the broader context of information retrieval and text mining for the crisis informatics domain. After a brief outline of the field, and of social media mining for crisis informatics, in particular, the chapter describes conversations characterization and taxonomies, cross-lingual and contextual classification, semisupervised learning and recommendations literature.

Chapter 3 - Characterization conversations: This chapter characterizes and analyzes the rise of conversational interactions in Twitter with an emphasis on crisis-related events. The characterization focuses on identifying the properties of conversations. Such properties include the type of conversations (wide or in-depth), the distribution of the replies, and the number of users participating. Aditional, these chapter aspects propose a model that identifies when a tweet could start a conversation by using user and tweet features.

Chapter 4 - Hierarchical multi-label taxonomy: This chapter analyzes the classification tasks of individual tweets found in the literature, which do not have enough context and cannot provide intuition on whether the interaction produces an outcome. This chapter introduces a new hierarchical multi-label taxonomy to categorize individual tweets as well the whole conversation.

Chapter 5 - Cross-lingual classification: Often, users from different countries around the world can interact in conversations about crisis events on social media. This chapter evaluates classification models suitable for handling multi-lingual scenarios for crisis-related conversations.

Chapter 6 - Context-aware classification: Current classification approaches have focused on the analysis of individual tweets, which do not have enough context to disambiguate information. This chapter analyzes conversations during several crisis events get insights into the use of conversational context in NLP classification tasks.

Chapter 7 - Semisupervised learning: Supervised learning often requires a lot of feature engineering or a significant amount of annotated data to achieve good results. This chapter introduces a semi-supervised neural architecture for multi-label settings, that combines deep learning representation and k-means clustering for text classification.

Chapter 8 - Recommending conversations: Since the amount of data on social media networks can overwhelm users, recommending relevant content becomes an essential task. This chapter describes a state-of-the-art recommendation model based on a sequence-to-sequence neural architecture for recommendation users or conversations on social media.

Chapter 9 - Conclusions: This chapter summarizes the main findings in this thesis and points out directions for future research.

Chapter 2

Background

This chapter provides the concepts and background needed in later chapters in this thesis. Section 2.1 provides a brief introduction to conversations on social media, and specifically crisis-related conversation on Twitter, which is the focus of this thesis. Section 2.2 introduces previous works on the literature related to the categorization of tweets in the domain. Section 2.3 summarizes the multilingual aspect of the crisis-related conversations on social media. Section 2.4 surveys background material on context-aware classification. Section 2.5 discuss methods to overcome the problem of low-resource languages because of the scarcity of labeled data. Section 2.6 detail preliminaries of content recommendation on social media.

2.1 Characterizing Conversations on Social Media

Understanding human conversations have been extensively studied and continue attracting the attention of researchers in the quest to achieve human-level reasoning and comprehension in machines. Conversation modeling has been studied previously using cellphone SMS corpus [53, 11], IRC chat corpora [37], and blog datasets [152].

There are several research directions in modeling human conversations. Amongst them, identifying conversation acts[1]. Several applications rely on acts identification, such as: conversational agents [147], dialogue systems [10], automated customer support service [103], virtual assistants [82], among others.

Conceived as a medium to share short personal status, Twitter rapidly evolved as a platform to interact with others with the novel use of '@' as a way of targeting other users to reply to a prior status or establish interactions in the form of conversations [51]. Prior work studied the use of the retweet as a means of engaging in conversations, and how dealt with different aspects such as authorship, attribution, and fidelity of the communication [21]. The authors found that in general,

[1]Known also as dialog acts.

conversations are messy, even when the interactions take place in a bounded group by location, timespan, and participant characteristics. In bounded groups, it is more likely to find cohesive conversations with turns (a set of contiguous tweets that belong to the same user) and references to previous messages, but that is unusual on unbounded groups where conversational structures are missing.

The aspect of information diffusion on Twitter depends on the analysis of the retweet mechanism on a large-scale analysis of the factors impacting the retweeting behavior for tweets [127]. This work identified a strong correlation of retweet behavior with content features (e.g., URLs and hashtags), as well as contextual features (e.g., number of followers and friends). Messages and breaking news on tweets sometimes propagate outside of the group of the originator, i.e., not restricted to the followers and depends on the user influence calculated by several metrics such as the number of followers, replies, retweets [153]. The process of content diffusion on Twitter can take the form of cascades when users reshare tweets. The study of the characterization and predictability of cascades of information show the predictability of temporal and structural features. Moreover, breath propagation rather than depth is a better indicator [28].

Previous works have studied the problem of predicting the popularity of messages [52]. Based on the future number of retweets, and how those influence the content propagation, proposed classification models include several content-based, contextual, and temporal features extracted from tweets. Additionally, models include network structure properties in its prediction model. Most of the previous studies focus on measuring the popularity of content based on the propagation of the content in the network. However, how a user writes a tweet can impact on its popularity and propagation. Previous works measure this factor by taking pairs of tweets posted with similar URLs and written by the same user but using different words. Previous findings indicate that some tweets can have more influence and popularity than others, depending on the words' choice [129].

In another aspect, some authors tackle the problem of predicting the popularity of the conversations on Reddit Threads [46]. Although it is a different social network, the authors tackled the problem of identifying the popularity of a conversation thread based on the content analysis using deep reinforcement learning. These prior works are closely related to the study but differ in the task and the metric used. This work focus on the number of replies received by tweets to predict whether a given tweet will generate interactions from other users.

2.2 Fine Grained Taxonomy for Crisis-related Conversations

The analysis of data from social networks has become critical for many application areas [101, 78, 117]. Immediately after a crisis or natural disaster has occurred, people use social media platforms to report the situation in the affected places, look for useful information, and request/offer help [57, 26]. Previous studies have analyzed the usefulness of extensive stream data from social media during crisis

events. It can be instrumental during natural disasters to gain insight into the situation as it unfolds for the relief efforts by organizations and activists [5, 120, 136].

Previous works implemented several systems to classify, extract, and summarize crisis-relevant information from social media; for a detailed survey see [57]. For example, the implementation of platforms for emergency awareness [24], using an SVM model to classify interesting tweets during natural disasters. Also, Naive Bayes and MaxEnt classifiers help to find situational awareness tweets from several crises [138]. Imran et al. implemented the AIDR platform to collect and classify Twitter data streams during crises [58]. This platform uses a random forest classifier trained in an offline setting. Although, after receiving every minibatch of 50 training examples, they update the older model with a new one.

The classification tasks used by previous work use single label taxonomy [57, 62, 100], which miss information in some tweets that refer to two or more labels at the same time. Moreover, previous works do not consider the label of the conversation itself, such as the outcome of deep conversations. This thesis proposed a new fine-grained taxonomy to account for hierarchical multi-label classification at a tweet and conversation level.

2.3 Cross-lingual Crisis-related Conversations

Previous works analyze the usefulness of extensive stream data from social media during crisis events. Immediately after a crisis or natural disaster has occurred, people use social media platforms to report the situation in the affected places, look for useful information, and request/offer help [57, 26]. It can be crucial during natural disasters to gain insight into the situation as it unfolds for the relief efforts by organizations and activists [120, 5, 136]. There are several datasets publicly available for crisis events such as crisisNLP datasets[2], which includes tweets from a large number of crisis-related events, such as earthquakes, floods, hurricanes, and cyclones. In each event, users often post tweets in their native language, where the event occurred. However, users from other countries and languages can post for the same event, leading to multi-lingual datasets.

Previous works proposed several learning methods to classify and categorize crisis-related data using single language approaches [138, 24, 57]. However, learning single language models means the model requires retraining for a different language, and this is even more critical for low-resource languages. Traditional classification approaches have several shortcomings due to the discrete word representations and the dependency on the trained data for a specific event and language, so they have poor performance classifying data for new events even more for new languages (out-of-event data) [62]. Recent approaches use neural architectures to deal with the issue the out-of-event data, specifically Convolutional Neural Networks (CNN) [100, 23] and semi-supervised learning [8, 7]. Social media reach worldwide audiences; therefore, learning methods have to deal with a myriad of

[2] http://crisisnlp.qcri.org/

languages for user-generated data during crisis events. Other approaches evaluated several scenarios of learning methods with cross-lingual data, e.g., using 30 cross-lingual datasets of crisis events where the model is trained on one language and tested in another language [68]. Their approach uses an SVM classifier with several semantic features in addition to the tweets' text. This thesis leverages the experimental setup proposed and evaluates an end-to-end multi-lingual model based on deep contextual embeddings and neural architectures.

2.4 Context-aware Classification of Crisis-related Conversations

Previous studies have analyzed the use of user-generated data in online social networks (OSNs) during crisis events. It can be instrumental during natural disasters to gain insight into the situation as it unfolds for the relief efforts by organizations and activists [136]. Several systems proposed in the literature perform tasks to classify, extract, and summarize crisis-relevant information from social media; for a detailed survey see [57]. For example, the implementation of platforms for emergency awareness [24], using the Support Vector Machine (SVM) model to classify interesting tweets during natural disasters. Imran et al. implemented the Artificial Intelligence for Digital Response (AIDR) platform to collect and classify Twitter data streams during crises [58]. This platform uses a random forest classifier trained in an offline setting. After receiving every minibatch of 50 training examples, they update the older model with a new one.

Traditional approaches require manually engineered features like cue words and TF-IDF vectors for learning [57, 62]. In contrast, recent approaches based on deep neural networks tackle some of the issues in traditional methods, such as the generalization error using word embeddings [100]. The classification approaches in previous works use learning techniques with either a discrete (TF-IDF) or dense (embeddings) representation of the text to classify crisis-related data. At the beginning of a disaster event, there is little to no labeled data available for training the models. Minutes or even hours later, the labeled data is ready in small batches depending on the availability of volunteers, often geographically dispersed. Once trained, the learning algorithms are dependent on the labeled data of the event for training. Due to the discrete word representations and the variety across events from which the big crisis data is accessible, they have poor performance when trained on the data from previous events (out-of-event data). Moreover, social media platforms allow users to interact in conversations that provide a context. Previous works use individual labeled tweets but do not consider the conversational context in their NLP approaches. This work aims to evaluate the effect of using conversation contexts in classification tasks.

2.5 Leveraging Unlabeled Data for Crisis-related Conversations

Traditional approaches to identifying conversation acts rely on manual human annotation. This process includes collecting and labeling acts in the dataset following an annotation guide. Although successful, this process can be very time consuming and costly to carry out and affects the analysis and use of low-resource languages such as Spanish. There are several approaches to mitigate manual annotation of the data. A straightforward approach is semisupervised learning to leverage the massive amount of unlabeled data available on social media. Previous work on semi-supervised clustering methods analyzes methods based on: constraints and representation. The constraint-based approaches use a small percentage of labeled data to restrict the clustering process [33]. Instead, the representation-based methods first learn a data representation model that satisfies the labeled data, and then use it to group both labeled and unlabeled data [13].

The hybrid approaches try to integrate both methods in a unified framework [18]. However, the use of linear projection for learning by representation has limitations to achieve a reasonable performance. In the last few years, several approaches use deep neural architectures to learn text representations that overcome the limitation of linear models [150]. However, the separation of the learning process of the data representation model and the clustering model restricts the benefits and is more similar to the techniques representation-based. In this work, the proposed model builds on an approach that combines into an integrated framework both the representation of deep learning and the clustering method [144].

2.6 Recommending Crisis-related Conversations

As popular microblogs such as Twitter add new functionalities to their platforms, there are new tasks to tackle on the field of social recommender systems (SRSs) and short-text analysis. A recent survey provides insights into the trends of academic literature reviews in the proposed context and presents a comparison of different research approaches [131]. They introduce a classification framework to better understand the methodologies and trends used in the development of RSs for microblogging. The authors identified nine types of recommendation: hashtags, mentions, news, points-of-interest, profile classification, retweets, tweets, URLs, and whom to follow. Recently, a related work proposed the task of recommending conversations to users, which is a slightly different task to the outlined in this work. The work presented focuses on a novel classification task: recommending users that might join a conversation. The task can help a user starting a conversation to find other users that might be interested in joining that conversation to achieve a goal, e.g., provide help during crisis events such as natural disasters.

There are various types of collaborative filtering (CF) techniques for RSs, dividing them into four categories: feedback-based, trust-based, matrix factorization-based, and nearest neighbor-based approaches [151]. Tweet-recommendation approaches use network, content, and retweet analyses for

making recommendations of tweets. The main advantage of the tweet-recommendation method is to recommend tweets that are not visible to the user through transitivity (friends-of-friends) relationship can determine useful recommendations. However, the disadvantage is that not always other users that might in the reach of the user's ego-network [9]. Another model extends the Word2Vec model to learn users and posts jointly representations [156]. The recommendation of posts to users depends on the cosine similarity of users and posts vectors. The method is simple but effective, and the method relies on a similar idea for similarity through vector representation but using more complex recurrent neural networks on top of the Word2Vec representations [90]. Other approaches use recommendation models based on a hybrid approach (CF + Topic modeling) that extract different aspects of the conversations. Although the empirical results show that the model provides better performance to prior methods, the disadvantage is the scalability and the generalization to new users or content [159].

In recent years, researchers have achieved substantial improvements in several machine learning tasks in areas such as computer vision, speech recognition, and machine translation, among others. These advances have been possible due to new deep learning architectures [77], sparked by the availability of massive datasets [34], increasing computational power [104], and the development of robust frameworks for training neural architectures [2]. Neural architectures have proven successful in improving RSs. For instance, incorporating deep learning techniques to existing CF methods improves the performance of recommendations by extracting content-based features from items [141]. For recommendations based on text content (e.g., news articles, blog posts, movie summaries, papers), sequence models, such as RNN, improve session-based recommendations where matrix factorization models are not accurate [49]. In conversational modeling, neural architectures require little feature engineering and are capable of generalizing unseen conversations [48]. Conversational modeling uses the ideas of statistical machine translation by learning phrase representations using RNN encoder-decoder models [116, 125]. This type of architecture can generate the representations of text sequences, such as conversation context and user profiles required for the tasks studied in this thesis. Seq2Seq models use a recurrent network (i.e., RNN or LSTM) to encode the input to a vector representation, and another sequence network to decode it in order to learn to rank [83] or to generate responses [123]. Seq2Seq models yield to better results compared to prior methods based on information retrieval for unstructured conversations [66]. This work presented in this thesis leverages Seq2Seq architecture to learn the text representation of the conversation context and user profiles for the task of recommending conversations.

Chapter 3

Characterizing Conversations

The previous chapter introduced the background material for this thesis. Starting in this chapter, the research questions listed in Chapter 2 are the main focus. This chapter addresses RQ1, which is concerned with characterization crisis-related conversations on Twitter and begins with an introduction about the motivation for characterizing conversations in Section 3.1 In Section 3.2, the characterization of conversations defines how to extract useful features and filter valid conversations, and then proposed a predictive model for identifying conversations. Section 3.3 describe the experimental settings and Section 3.4 presents the results. Finally, Section 3.5 outline the conclusions and research directions. The work presented in this chapter was first published at the Fourth IEEE/ACM International Conference on Big Data Computing, Applications, and Technologies [133].

3.1 Introduction

Amongst microblogging sites, Twitter has become one of the most popular worldwide. In this social network, its users share content publicly via short texts named tweets. Although most tweets generate little or no interaction with other users, sometimes the published content can ignite a long chain of replies and interactions from other users. This thesis denotes as seed tweets those that initiate conversations and seeks to understand the factors in seed tweets that contribute to igniting replies from other users. The published content on social networks can have an impact on different aspects of society, such as popular culture, brand communication, politics, activism, journalism, crisis communication, among others [146]. As the content generated increases on social networks, the factors that ignite conversations or discussions are of particular interest.

In the last decade, microblogging—specifically Twitter—has attracted the attention of the research community due to the open data access through its public APIs[1]. Several aspects of Twitter have been studied by researchers, including but not limited to its network structure, users' behaviors,

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

the content generated, and the infrastructure needed to handle its massive datasets. A particular aspect that calls the attention of researchers is related to the nature of the interactions that occur in this social network. Among the type of interaction are the conversations spontaneously occurring among users. Even though initially the idea of Twitter was to allow users to post what they were doing, soon, its users began to use @ symbol to interact with other users. This type of interaction often evolves into natural conversations that blur the border between conversations in private chats and public blogs [51].

In understanding human conversations, several aspects are essential. Some of them consist of identifying the structure and intent of conversations [115]. Predicting whether content posted on social networks will become popular or generate interest from users constitute another aspect in the analysis of conversations. The latter could be useful in many applications in the area of recommender systems (news feed, advertising placement). For instance, a user may be interested in reading several articles on different topics, for which a real-time news recommendation system should show relevant articles to fulfill users' preferences or generate interest from the user. Similarly, ads published on social networks, aim to generate attention (reading) or interactions (in the form of a like, retweet, or replies) from its audience. The work presented in this chapter examines the factors that contribute to sparking a conversation on Twitter, i.e., identifying whether a tweet that will generate replies from other users. The hypothesis is that contextual and content features extracted from the tweets can help to predict the likelihood of a tweet evolving into a conversation. The goal is not to predict the popularity level, but instead, if a tweet will evolve into a conversation.

The main contribution of this chapter is to:

- Introduce a new corpus for social media conversations.
- Characterize tweets that ignite conversations.
- Design and implement a classification model to identify seed tweets.

3.2 Methodology

The methodology in this chapter performs an exploratory analysis of conversational data collected from Twitter and then describes a predictive model for identifying when a tweet evolves into a conversation. The analysis in this chapter uses a dataset related to the earthquake in Ecuador in 2016. The Appendix A details the data acquisition, storage, and processing of the dataset.

3.2.1 Characterization

This section determines the features that identify and filter tweets that belong to conversations. This analysis helps to uncover features to filter out tweets that do not represent conversations between two or more users. Thus, the predictive model in Section 3.2.5 uses only valid conversations.

Like most human activities, short conversations are the bulk of the dataset, whereas there are few very long conversations. Figure 3.1 shows that the number of tweets in conversations follows a power law distribution [32].

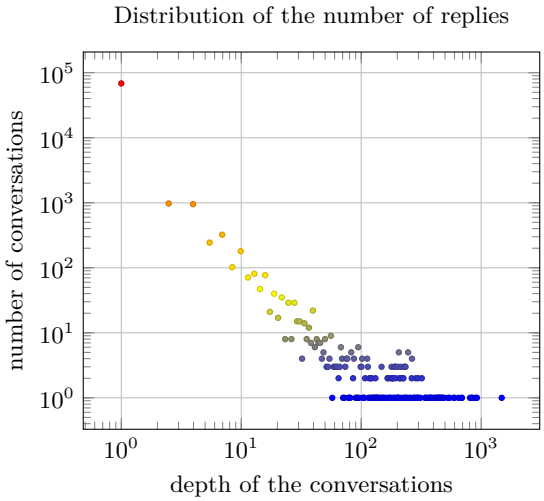


Figure 3.1: Distribution of the number of replies in conversations for the dataset.

Table 3.1 show that 64% of the tweets are non-conversational. Furthermore, the remaining tweets constitute conversational threads containing two or more tweets (length of the conversation). For conversational tweets (fourth column), 41% of them are short conversations (one tweet and one reply), similar to the results found in [115]. Lastly, the number of conversations with more than 5 replies are marginal, as the distribution depicted in Figure 3.1.

# of conv.	Length	% tweets	% conv.
1,747,374	1	0.64	0.00
401,274	2	0.15	0.41
144,078	3	0.05	0.15
86,512	4	0.03	0.09
61,935	5	0.02	0.06

Table 3.1: Conversations statistics.

Although, the majority of users in Twitter post tweets in one language, some users can use multiple languages. The information provided by the Twitter API in each tweet's metadata, specifically the field lang, allows detecting the language of the tweets. There is a total of 44 different languages detected, from which English and Spanish represent 85% of the tweets. The language was not available for a small percentage of the tweets (7% approximately), which have an indicator of undefined language. The content of tweets with undefined language is usually limited to mentions, hashtags, URLs, emoticons, or multimedia (i.e., images, videos). In addition to English and Spanish

languages, the further analysis includes the tweets marked as undefined when those are part of a conversation. Table 3.2 shows the distribution of the number of languages used for conversations in the dataset. The first column refers to the number of languages detected. The second column indicates the number of conversations. The third indicates the percentage of conversations by language, and the fourth the cumulative percentage. The majority of conversations contain tweets in one language (75%), and there is a 25% of conversations with two or more languages. However, up to three languages represent the cumulative 99% of the total number of conversations, which indicates that it is rare to have conversations with more than three languages.

# of lang.	# of conv.	% of conv.	cum. %
1	216,138	0.75	0.75
2	62,992	0.22	0.97
3	6,165	0.02	0.99

Table 3.2: Languages in conversations.

Geographically, this study focuses on countries in the American continent, but the interactions reach places all over the world. The dataset has approximately 12% of tweets with geolocated information associated, and 3.2% correspond to conversations. Although the geolocated tweets are a small percentage, Figure 3.2 shows users posting tweets all over the world, mainly in English or Spanish speaking countries.

Figure 3.2: Heatmap of geolocated tweets (12% in exploratory dataset). The markers, connected by the geodetic line, represent a conversation between two users in very distant places.

Although very distant conversations are not uncommon, conversations in the same point (zero meters from the origin) may be an indicator that the same user is self-replying or creating a message in multiple tweets. The analysis considers only conversations involving more than one user for further analysis, to avoid selecting tweets in the same spot as conversations. The analysis of the duration of conversations shows that the temporary distribution of tweets in conversations is uniform throughout the week. Figure 3.3 shows the density slightly increasing at night on Tuesday and Wednesday, as well as Friday morning, and it goes down the Saturday.

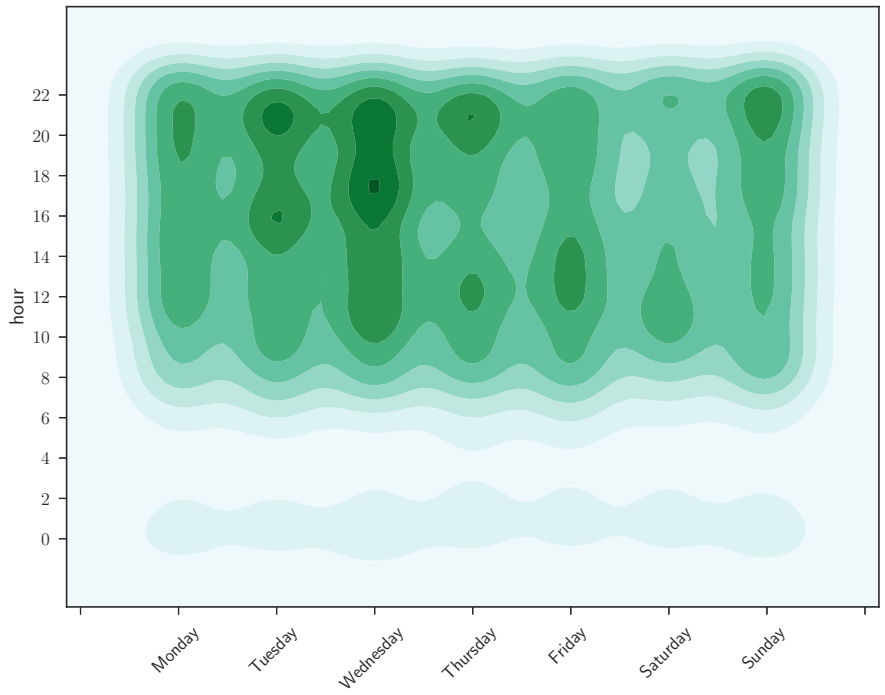


Figure 3.3: Temporal distribution for tweets initiating conversations.

Another aspect of interest is the duration of the conversations, as shown in Table 3.3. Most of the conversations are short-lived, i.e., have a duration of fewer than ten days. Also, the analysis show that concise duration replies are usually self replies created by third-party apps. For instance, the following tweets belong to the same user, created almost at the same time[2]:

tweet: 2017-07-12 21:00:24: @trendinaliaEC: '1. #ExperienciasElectoralesEC 2. #NoALaViolenciaDeGenero 3. Lula da Silva 4. #CPCCSMarcandoElCamino 5.Alfaro Moreno...'
reply: 2017-07-12 21:00:24: @trendinaliaEC: '6. Roger Federer 7. #LeyEficienciaTramites 8. Defensor del Pueblo 9.#EmergenciaCBQ 10. James Rodriguez.'

On the other hand, the increasing use of bots or spam accounts on Twitter can create noisy

[2]https://twitter.com/trendinaliaEC/status/885242460644888577

conversations that span several years. The following tweets illustrate this case[3]:

tweet: 2009-03-07 @finkd: 'Yes; this is the real Mark Zuckerberg. Thanks for following me!'
reply: 2016-08-13 @oropesa555: 'SOS SOS @$Pontifex_es$ @YourAnonGlobal @finkd ...'

Despite these cases, there are valid conversations with long duration. Usually, new friends or followers may visit tweets posted long ago and comment on them (using the reply option). For instance, the following conversation[4]:

tweet: 2009-10-25 @NARSissist: 'Eaten alive by a mosquito... Not fun'
reply: 2017-01-08 @GigiFreireA: '@NARSissist WTF NARS? xD'

This analysis filters out tweets that fall in the case of conversations containing sequential posts created by the same user in a short period. The other cases are more challenging to identify, so the analysis includes all tweets in long-duration conversations for further analysis.

Days >	≤	Conversations	%	cperc
0	10	277,222	0.97	0.97
10	20	4,548	0.02	0.98
20	30	1,138	0.00	0.99
30	40	627	0.00	0.99
40	50	379	0.00	0.99

Table 3.3: Duration of conversations.

The spatial and temporal analysis found that 80% of the conversations have two or more users involved, as shown in Table 3.4. There are some conversations where all tweets belong to the same user, therefore not included as valid conversations. For the conversations involving two users, the median of the conversation length is two tweets. However, there are some outliers, e.g., some sports journalists narrating football matches on Twitter using replies can from a conversation thread with more than 500 tweets. The predictive analysis considers only conversations having two or more users.

# Users	# Conversations	%	cum. %
1	20,476	0.07	0.07
2	245,417	0.80	0.87
3	29,128	0.09	0.96
4	7,029	0.02	0.98
5	2,420	0.01	0.99

Table 3.4: Users in conversations.

[3] https://twitter.com/finkd/status/1293412597
[4] https://twitter.com/NARSissist/status/5157432533

The next sections describe the predictive model for identifying seed tweets that evolve into conversations.

3.2.2 Prediction dataset

The predictive analysis uses a subset of tweets that allows understanding the features that might spark a conversation. The idea is to use this subset of tweets to extract the features, i.e., independent variables. Then, the preprocessing defines whether a tweet is a part or not of a conversation as the dependent variable in the model. From the exploratory dataset, the preprocessing selects randomly 1000 tweets that initiated a conversation (parent tweet of a conversation), stratified by the number of replies received. For these conversation tweets, the sample includes all replies. Then, the preprocessing select randomly 1000 tweets that do not evolve into conversations, i.e., with no replies. In total, the prediction dataset contains $10,805$ for tweets. Before further analysis, the preprocessing applies a log transformation to features with large values (number of followers, friends, tweets posted). These features usually follow a power-law distribution, while other features remain with the original range of values.

3.2.3 Features extraction

The analysis pipeline in this study employs a feature extraction step before using the prediction dataset in the conversational modeling. These features are content, contextual, and language-invariant attributes present in the text and metadata of each tweet. Table 3.5 shows two types of features: user-related and tweet related features. Those associated with the user level include metadata from the user profile. The tweet related features are the metadata and tweet's content itself.

User level features	
Statuses	# of statuses
Followers	# of followers
Friends	# of friends
Favorites	# of likes given to tweets by the user
Tweet level features	
Retweets	# of retweets received
Favorites	# of likes received
Urls	# of Urls in the tweet
Hashtags	# of hashtags in the tweet
Mentions	# of mentions in the tweet
Media	If there are images or video in the tweet
Replies	# of replies received

Table 3.5: Conversations features.

The feature extraction considers contextual features at the user level (e.g., the number of statuses posted, followers, friends, and favorites), and content features at the tweet level (e.g., the number of URLs, hashtags, mentions, and multimedia references associated). Some tweet attributes indicate the popularity of a tweet, such as the number of replies, retweets, and favorites. The focus of this study is conversations, as suchthe number of replies, is the target variable to predict. The feature extraction employs regular expressions for content attributes to identify words starting with @ (mentions), # (hashtags), or URLs. For contextual features, the feature extraction uses the tweets' metadata; while, the target variable is the count of all the replies for each tweet. Before the PCA and predictive analysis, the data cleansing of the dataset removes tweets with missing user profile features, which correspond to a few cases (0.3% of tweets). For user-level features, the feature extraction step applies log transformations to avoid large values dominating in the predictive analysis.

3.2.4 Principal components analysis

The pipeline performs PCA to reduce dimensionality and find possibly correlated to features shown in Table 3.5. PCA transforms the features into a small set of factors, identified as principal components. This technique aims to reveal the underlying data structure, and the weights each feature contributes to the data variance. Table 3.6 shows the principal components or factors in the dataset presented in descending order of importance. The second column contains the eigenvalues, i.e., the variance accounted by each factor. The third column shows the percentage of the eigenvalues for each feature and the cumulative percentage in the last column.

Factor	Eigenvalue	% Variance	% Cum. Var.
1	1.71	0.27	0.27
2	1.47	0.24	0.51
3	1.02	0.16	0.67
4	0.54	0.09	0.76
5	0.49	0.08	0.84
6	0.42	0.07	0.90
7	0.27	0.04	0.95
8	0.21	0.03	0.98
9	0.12	0.02	1.00

Table 3.6: Principal components analysis.

The number of factors used may influence the error variance if PCA retains too many factors while retaining a few factors risk leaving out valuable common variance. A criterion to determine the number of factors to retain is the Kaiser's criterion, which is a rule of thumb to retain that recommendation to retain the factor with eigenvalues greater than 1. The pipeline performs a scree test to avoid overestimating by choosing factors before the flattening of the slope of eigenvalues.

The pipeline uses the Kaiser's criterion as well the scree test [154] to determine the number of factors to retain. By combining the rules as mentioned earlier, PCA retains factors 1, 2, and 3 in Table 3.6. Together these factors represent 67% of the total variance of the features. Table 3.7 shows factor loadings (correlations) between the original features in Table 3.5 and each of the three factors retained in the previous step.

Feature	Factor1	Factor2	Factor3
fav given	0.35	−0.09	−0.08
followers	0.50	−0.20	0.12
friends	0.52	−0.01	0.07
statuses	0.53	−0.20	0.04
tokens	0.04	−0.17	−0.06
urls	0.02	0.01	0.04
hashtags	−0.02	0.25	0.95
mentions	0.27	0.91	−0.24
media	0.03	0.01	0.03

Table 3.7: PCA factors loadings.

To visualize the importance of each feature based on the correlation with factors, Figure 3.4 shows the factors in pairs. Factors represent the axis of the graphs, for instance, the factor vector for feature mentions is represented in the first graph with coordinates $(0.27, 0.91)$. Likewise, in the second graph, the same feature is represented with coordinates $(0.91, -0.24)$, illustrating the correlation feature-factors. The interpretation for the first graph is that it represents the networking and activity level of the user based on the high correlation with user profile features. The activity level is related to the number of favorites the user has given to other tweets or the number of tweets created. The networking level is related to user attributes that indicate network relationships (followers, friends). Meanwhile, the second graph represents the content patterns of the user. There is a slight negative correlation with the feature mentions. Other content-related features are slightly positive (e.g., the number of URLs, media, or hashtags). Content features, such as the number of tokens, have a negative correlation in the third factor. An important aspect is a fact that mentions to other users often generate a response (reply). Factor vectors of users' mentions and hashtags are predominant in both graphs of Figure 3.4.

3.2.5 Prediction model

Based on the underlying structure extracted in PCA, the predictive model aims to determine the likelihood of conversation arising for a given tweet, i.e. as binary classification task. To tackle this problem, the pipeline train and evaluate a supervised classifier using the prediction dataset. The classification model renders a set of prediction coefficients for each feature that predict the probability of a tweet initiating a conversation. This study defines the dependent variable as binary:

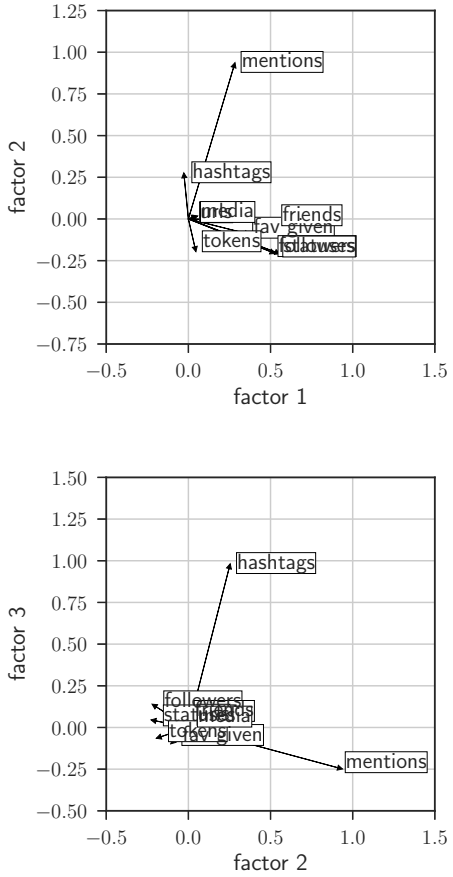


Figure 3.4: Factors mappings.

the tweet initiates or not a conversation, and transform the original feature number of replies (nr) into a binary feature, as follows:

$$c(n) = \begin{cases} 0 & \text{if } nr = 0 \\ 1 & \text{if } nr > 0 \end{cases} \tag{3.1}$$

3.3 Experimental Settings

The classification pipeline consists of extract and standardizes the features, apply PCA, stratified splitting cross-validation, and random grid search for hyper-parameters tuning. The pipeline uses two input datasets: content features only and full (user + content features). The first dataset considers only content features, i.e., tweet level features, while the second considers both users and content level features. This study evaluates the following classifiers: a) Logistic Regression b) Support Vector

Machine (SVM): RBF kernel c) Gaussian Naive Bayes (NB) d) Neural Net e) Naive Bayes

3.4 Results and Discussion

Table 3.8 shows the coefficients in the predictive model. These coefficients corroborate specific findings of the importance of certain features revealed through PCA analysis. Mainly, those related to the activity level (favorites granted, number of tweets posted), the interaction (mentions in tweets), the network (followers), and content (tokens in tweets). Friends have a negative coefficient, i.e., little influence in initiating conversations. Content features such as URLs or hashtags present in tweets are less critical in conversation forming, while mentions are slightly more critical.

feat	coef	std err	z	P>|z|
intercept	0.185	0.082	2.254	0.024
fav given	0.771	0.087	8.832	0.000
followers	1.074	0.098	10.934	0.000
friends	−0.214	0.070	−3.047	0.002
statuses	−0.092	0.094	−0.973	0.331
tokens	0.449	0.083	5.409	0.000
urls	−0.550	0.112	−4.906	0.000
hashtags	−0.471	0.071	−6.650	0.000
mentions	0.088	0.047	1.893	0.058
media	−0.046	0.137	−0.337	0.736

Table 3.8: Predictive model.

In Figure 3.5, the visualizations use the first two factors obtained in PCA. The results show that Logistic Regression performs consistently using both datasets: content only features or all features. Using all features, SVM has the best performance (0.80), followed by a Neural Net model (0.79). Neural Net model shows promising results, moreover if future works want to include for more complex tasks that involve analyzing textual and visual content.

Additionally, Figure 3.6 shows the performance of the best classifier by separating the dataset by percentiles (10th, IQ, 90th). This study uses the three features with a higher coefficient in Table 3.8. In the case of the feature number of followers, the classifier performs better for the 90th percentile as those are users with high popularity, as their tweets are more likely to generate replies. The feature favorites given that denote the activity level of the user for the 90th percentile has similar behavior, but interestingly for users with few activities (10th percentile) performs better than for average users. The number of tokens created by users has a similar behavior as favorites given., but this could be because IQ percentile contains noisier tweets than other percentiles.

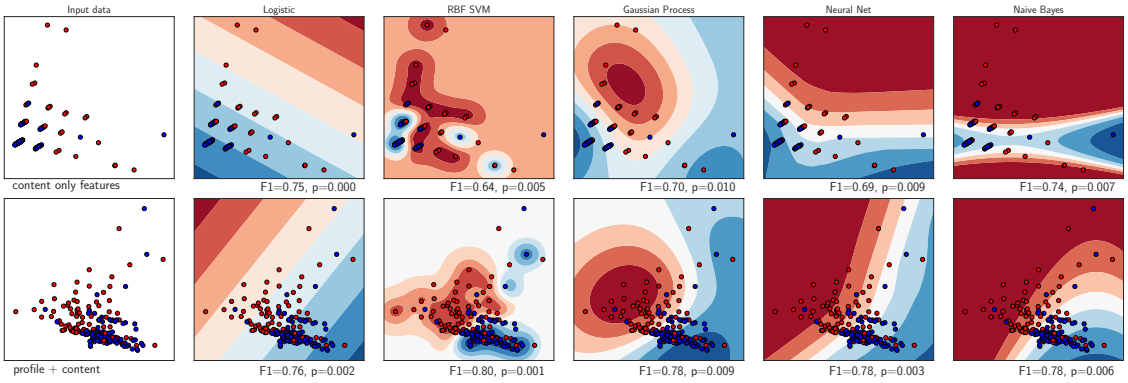


Figure 3.5: Classification models for identifying seed conversation tweets.

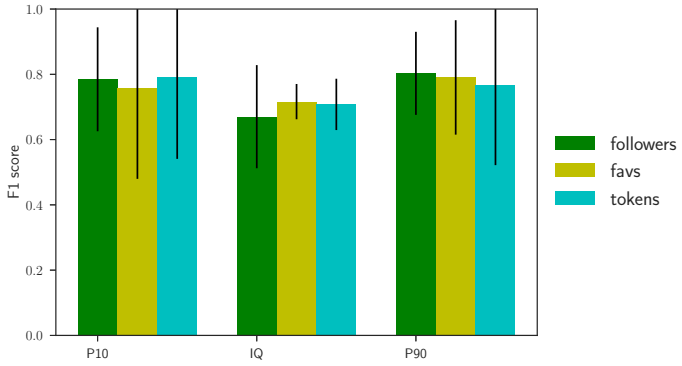


Figure 3.6: Classification comparison for different features percentiles.

3.5 Conclusion

This chapter analyzed the factors that may influence conversation forming from a given tweet. The characterization highlights the difficulties of working with noisy data found on Twitter and help to establish some considerations to avoid including noisy data in predictive analysis. For example, language, duration, distance, and the number of users in conversations can help to filter irrelevant and non-conversational tweets. The PCA analysis of both user and content features helped to establish their correlation and their influence on predictive models. The predictive analysis shows that the overall F1 score improves if the model uses both user profile features as well as tweets features. Future work is to explore large-scale analysis for massive Twitter datasets using distributed machine learning. Also, multimodal models can include additional features through the analysis of the textual and visual content of the tweets.

Chapter 4

Hierarchical Multi-label Taxonomy

This chapter addresses RQ2, concerned with extending previous taxonomies to consider the case of hierarchical multi-label settings. This chapter begins with an overview and motivation for hierarchical multi-label taxonomy in Section 4.1. Next, Section 4.2 describes the methodology and describes the conversation modeling framework. Then, Section 4.3 details the experimental settings in and Section 4.4 show the empirical results. Finally, Section 4.5 outlines conclusions and research directions.

4.1 Introduction

The content generated by users in social networks contains information that can be useful for disaster relief. One of the limitations in the management of social network data during Crisis events is related to the analysis of short and unstructured text messages, which are published by users and are directly related to disasters and emergencies [57, 26]. The analysis of messages posted on microblogging platforms such as Twitter can help humanitarian organizations (e.g., United Nations, Red Cross) or activists during crisis events. Social media allow them to become aware of the situation, know the urgent needs of affected people, assess critical damages in the infrastructure, identify medical emergencies in different locations, or coordinate relief or rescue actions [5, 139].

The automatic classification of tweets to identify useful and relevant information is a challenging task because: a) tweets are usually short text messages (limited to 140 characters[1]), which makes it difficult to understand them without sufficient context; b) tweets often contain abbreviations, informal (unstructured) language, misspellings, or present ambiguity; and, c) judging the relevance and usefulness of a tweet become a subjective analysis.

In order to categorize tweets into different topical classes, previous works have run into some issues because: a) tweets can contain information that belongs to one or more classes; b) choosing

[1]Twitter has doubled the number of characters in 2017 but this chapter analyses data from 2016.

the dominant category is difficult, even human scorers differ in their judgment about whether or not a tweet belongs to a specific category; and, c) the semantic ambiguity of the tweets, as well as the idiomatic phrases, sometimes makes it difficult to interpret them. Given these difficulties in tweets, text classification models generally do not score higher than the score in which human scorers agree with each other. Despite recent advances in natural language processing (NLP), the interpretation of the semantics of noisy short-texts remains a hard problem. Moreover, to extract useful information from Twitter, models must also understand complex interactions such as conversations. Analyzing the structure of a conversation regarding the different types of acts, such as statements or questions, can provide valuable insights into the flow, aspects, and outcome. Understanding conversations can be used as a first step to develop automated agents or support decision systems.

User	Tweet	Dialogs Act
miguelmomc02	Yo nunca siento nada, a mi solo me dicen corre .	Emotional, Expressive Negative
Lenniszumba	que maneraaaa	Sarcasm
miguelmomc02	solo senti cuando ya la casa se caia y estaba abajo	Infra Houses, Informative
Lenniszumba	pero el terremoto si lo sentiste? O noooo	Wh Question, Response Other
miguelmomc02	senti porque la casa se movia los postes todo y mis pies temblaban de hay nunca siento	Response Other
Lenniszumba	de leeey que si fue horribleeee que tragediaaaaas	Emotional, Expressive Negative

Table 4.1: Twitter conversation occurred after the earthquake in Ecuador 2016.

However, in order to understand and analyze conversations in the context of crisis events, it is necessary to extend previous coarse-grained or generic taxonomies. Table 4.1 shows an example of a conversation between several users on Twitter, where occur alternating turns to complain about a communication service. As shown in the dialogues, knowing that a turn is an act of type Statement or Request based on generic taxonomies is not enough to extract useful information relevant to a crisis. There is a need more detailed dialogue acts (labels), such as Informative Statement, Complaint, Offers, or Requests to capture the intention of the participants. Similarly, turns often belongs to multiple overlapping conversational acts, so that a multi-label approach is more precise than single label approaches.

Conversational acts prediction provides a guide for automatic response generation systems and to develop analysis tools for disaster management. Predicting conversational acts in real settings can be leveraged by an automated agent to generate responses in future scenarios. Moreover, meaningful patterns can emerge from the analysis of the fine-grained acts in conversations for a post-prediction setting. For example, if a user does a follow-up with specific actions in response to an urgent need request dialogue act, this could be seen as a successful relief outcome. The analysis of large numbers of conversational act sequences allows correlate to specific outcomes and derive several rules, e.g., Requesting certain types of services or goods in a conversation often leads to a positive outcome.

This chapter seeks to predict conversational acts to identify outcomes that can be useful for crisis

events. It addresses several challenges, such as: building a multi-label conversational act annotated corpus, which to the best of the knowledge, is not available for disaster management on Twitter. The task of conversational acts annotation is subjective, and existing taxonomies do not capture the fine-grained information that would be helpful for the crisis domain. Although tweets are concise, usually, human annotators can associate several overlapping conversational acts to the same tweet.

The contribution of the work is three-fold. First, design a fine-grained conversational act taxonomy. Second, build an annotated crisis-related corpus based on the proposed taxonomy for Spanish tweets. A third, propose a deep learning architecture based on Convolutional Neural Networks for multi-label for act prediction, as well as the conversation outcome.

This chapter first expands upon previous work about conversational act taxonomies [103] and crisis domain [60]. Then, it develops a fine-grained set of conversational acts for crisis management and conducts a systematic user study to identify conversational acts in the corpus containing 518 conversations about the Ecuador earthquake. The aim is to understand the conversation flow between users using the proposed taxonomy through a neural architecture [100] to predict the fine-grained conversational acts for a conversation as well as the potential outcomes, such as prevention, situational awareness, and relief coordination.

4.2 Methodology

This chapter focuses on the crisis management domain on Twitter in the context of fine-grained dialogue act classification. It provides recommendations about useful conversations, based on a real scenario data. The underlying goal of this chapter is to show how a well-defined taxonomy of dialogue acts to summarize semantic information about the flow of a conversation. The Fine-grained taxonomy allows to derive meaningful insights into the outcome of the interaction and then to extract actionable knowledge to the implementation of automated agents and tools. Figure 4.1 shows the methodology pipeline.

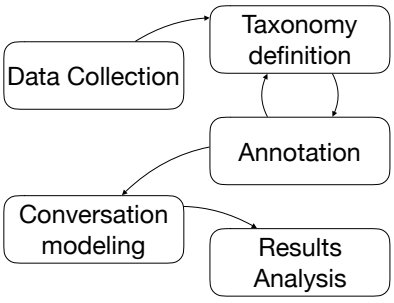


Figure 4.1: Methodology pipeline.

1. Data Collection and Taxonomy Definition: Extend previous works by defining a taxonomy of fine-grained dialogue acts suited to the crisis management domain and use this taxonomy to gather annotations for crisis conversations on Twitter in an iterative process.

2. Conversation Modeling: Develop a neural model to identify the different dialogue acts in a conversation during a crisis event and using a novel multi-label approach to capture different intents contained in a conversational turn (reply). The evaluation compares the model performance in different settings to better understand the differences between user interactions and relief efforts.

3. Conversation Outcome Analysis: Use the model to provide actionable knowledge for the development of support decision tools. The aim is to allow systems to answer questions such as, What is the correlation between conversation flow concerning the dialogue acts used?, What is the overall collaboration achievement for urgent needs/problems, and situational awareness? or, What information can users extract from conversations during crisis events for automated systems or support decision tools?

4.2.1 Taxonomy Definition

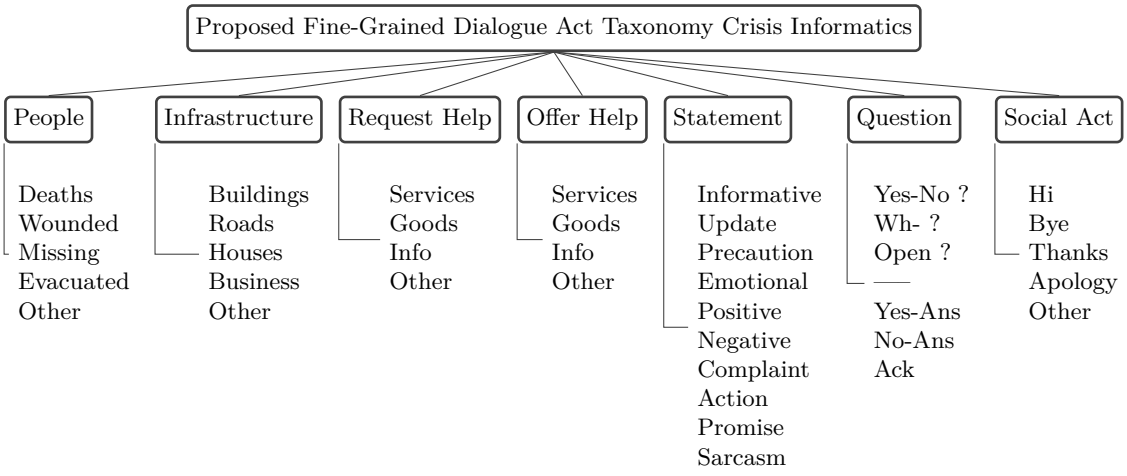


Figure 4.2: Proposed fine-grained dialogue act taxonomy for crisis-related conversations.

The coarse-grained nature of existing taxonomies presents several shortcomings concerning what type of information the models can learn by performing classification with this type of taxonomy. Although, the number of labels in existing taxonomies helps achieve good agreement between annotators. For instance, a kappa score of 0.87 between the three expert annotators [63]. However, it is unable to offer deeper semantic insights toward specific intents of each act for many of the categories. For example, the Informative act, which often comprises the most significant percentage

of turns, is a broad category that fails to provide useful information from an analytical perspective. Likewise, the Request or Offers categories do not specify any user intent behind the act, and there is room for improving the analysis of social media data. For this purpose and motivated by previous work seeking to develop conversational act taxonomies appropriate for different domains [64, 69], and convert the list of conversational acts presented by the literature into a hierarchical taxonomy, as shown in Figure 4.2.

This study builds upon prior taxonomy for instant messaging chat conversations [64, 69], but also on conversational acts observed in the crisis management domain, including people, infraestructure, and needs. The proposed taxonomy organizes labels into seven high-level conversational categories: People, Infrastructure, Request Help, Offer Help, Statement, Question, and Social Act. Then, the taxonomy includes subcategories iteratively through several iterations in the annotation process. The proposed taxonomy does not restrict which party in the conversation may contribute to each label. Some acts are more frequent or sometimes non-existent in usage, depending on whether the speaker is a general user or an organization. For example, the Sarcasm category rarely shows up in Government or NGO turns. The taxonomy includes Other acts as the broadest of the categories to account for gaps in conversational act selections for annotators. While the proposed taxonomy fills in many gaps from previous work in the domain, this study does not claim to cover all possible acts in the crisis domain. The proposed taxonomy allows to identify the intent and motivation behind each turn more closely, and ultimately extract knowledge from conversations.

4.2.2 Data and Annotation

The Appendix A details the data acquisition, storage, and processing of the dataset. The preprocessing performs several steps, as described next. It filters out conversations non-alternating users' turns (single turn per user), has less than three or more than ten turns[2]. Also, it filters out tweets that have less than 60 words in total, and if a turn in the Twitter conversation ends in ellipses followed by a link (which indicates that the turn has been truncated due to length and spans another tweet). After the preprocessing, the dataset contains 518 conversations, spanning 3,572 turns.

This study conducts the annotation study with three undergraduate students as annotators and presenting them with data consisting of conversations between users during the crisis event. The annotators use a definition of each of the conversational acts along with a sample annotated conversation for reference. For each turn in a conversation, the annotators select as many labels as required to characterize the intent of the conversational turns entirely. Additionally, annotators have to answer three questions at the end of each conversation, stating that they agreed, disagreed, or could not tell:

- By the end of the conversation, does the users acknowledge prevention action in anticipation

[2]The lower bound allows at least one turns per user. Then, it defines the upper-bound after finding that 98% of the conversations had ten or fewer turns.

of a crisis event or precaution about the current crisis?

- By the end of the conversation, does the users get situational awareness regarding the crisis?
- At any point in the conversation, does the users achieve collaboration to relief efforts (match need-offers)?

The annotation process asks three students to annotate each conversation. The list of conversational acts (labels) for each tweet is the list of any acts that have received majority-vote labels (at least 2 out of 3 annotations). It is important to point out that an important choice is how to handle conversational acts tagging for each turn. This study found that each turn may contain more than one dialogue act, and it is vital to give its full meaning. The lines differentiating these conversational acts are not very well defined, and applying segmentation would not necessarily help in clearly separating each intent. Therefore, similar to previous works, there is no specific segmentation task on the tweets due to the overall brevity of tweets in general and to avoid the overhead of requiring annotators to provide segment boundaries, and instead ask for all appropriate conversational acts. Also, contrary to previous works [85, 157], this study characterizes each tweet as a single unit composed of different, often overlapping, conversational acts.

Label	Example	% Turns	% Annot.
sarcasm	pobre infeliz no e digas lo q no sientes. Preocupa [...]	27.09	20.44
informative	no pero manifiestas que se tiene que volar sin ten [...]	22.89	17.27
response_other	lo sé, yo también he pasado por algo similar y sí [...]	21.11	15.93
emotional	Lo se, me lo imagino, y ojalá no vengan más réplic [...]	13.36	10.08
expressive_positive	necesitamos cadena d oración x todo el país.Suplic [...]	9.12	6.88
wh_question	vos estas bien?	7.16	5.40
complain	??!! Según Ud Ecuador paralizado si está de viaje [...]	5.24	3.96
expressive_negative	Fue terrible, hay destrozos en varios lugares de G [...]	4.20	3.17
thanks	Gracias hermanos. Hoy nos toca pasar momentos grav [...]	3.65	2.75
insult	Aprenda a leer. *Imbécil*	2.92	2.20
All Others		15.78	11.91

Table 4.2: Detailed distribution of top 12 fine-grained conversations acts derived from annotations.

Figure 4.3 shows the distribution of the number of times each conversation act, selected by majority vote between the annotators. Table 4.2 shows a more detailed summary of the distribution of the top 12 dialogue acts according to the annotation experiments, extended from the work presented by Ivanovic et al [63]. The measurement of the agreement in the annotations uses different methods. Since each tweet in the annotation process can have multiple labels, the first method evaluates an agreement metric that accounts for how frequently each annotator selects the acts that agree with the majority selected labels for the turns they annotated. The second method measure Fleiss' Kappa [39] agreement between annotators in two ways. First, by normalizing the annotation results into binary-valued items indicating annotators' votes for each label contain within each turn. The

Fleiss-κ values for each label uses the categories defined by Landis and Koch to bin the speech acts based on agreement [76]. Table 4.3 show that the per-label agreement varies from an almost perfect agreement of $\kappa = 0.871$ for lexically defined categories such as Apology and Thanks, with the only slight agreement of $\kappa = 0.01 - 0.2$ for less clearly-defined categories, such Response other, Sarcasm, Others).

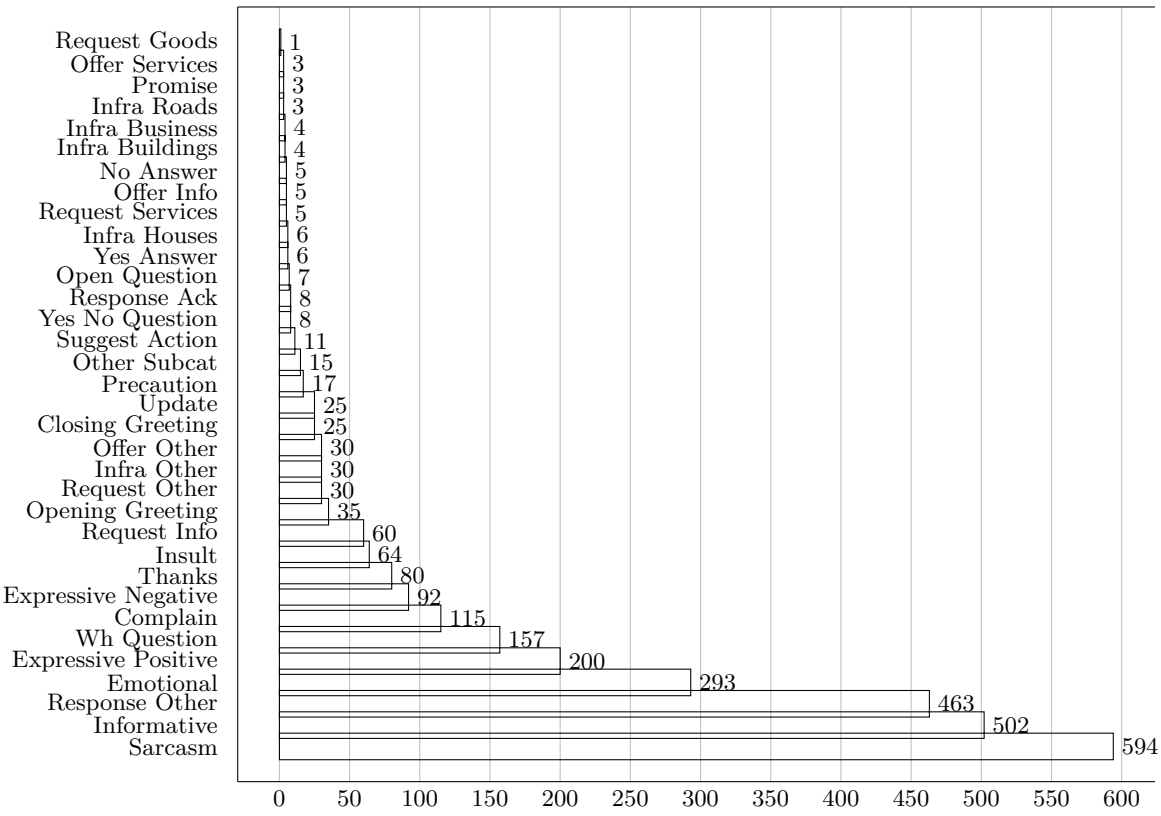


Figure 4.3: Distribution of annotated dialogue act labels.

The hypothesis is that tweet turns are often characterized by more than one distinct dialogue act label by measuring the percentage overlap between frequent pairs of labels. Table 4.4 shows the distribution percentage of the most frequent pairs, which highlights the feasibility of using the multi-label approach as there are conversational acts that contain more than one label.

4.2.3 Conversation Modeling

This section describes the few-shot learning approach for modeling conversations on the data collected and annotated on the proposed fine-grained taxonomy. The aim is to design an end-to-end deep learning model with minimal feature engineering. Although the proposed taxonomy provide more detailed information, new challenges arise given the large number of labels. Using all classes

Agreement	Dialog Act
(0.0, 0.2]	Sarcasm, Insult
(0.2, 0.4]	People Other, Response Ack, Response Other, Other Subcat
(0.4, 0.6]	Infra Other, Request Other, Offer Other, Update, Yes Answer, No Answer, Apology
(0.6, 0.8]	Outcome Prevention Ack, Outcome Situational Awareness, Outcome Relief Coordination, People Missing, People Evacuated, Infra Buildings, Request Info, Request Goods, Request Services, Offer Info, Offer Goods, Offer Services, Informative, Expressive Positive, Expressive Negative, Complain, Suggest Action, Promise, Yes No Question, Wh Question, Open Question, Opening Greeting, Closing Greeting, Thanks
(0.8, 1.0]	Crisis Related, People Deaths, People Wounded, Infra Roads, Infra Houses, Infra Business, Precaution, Emotional

Table 4.3: Dialogue act agreement in fleiss-κ bins.

Dialog Acts Pair		% Turns
Emotional	Expressive Positive	8.53
Emotional	Expressive Negative	3.88
Sarcasm	Response Other	3.19
Sarcasm	Insult	2.42
Informative	Emotional	1.69
Informative	Expressive Positive	1.32
Wh Question	Response Other	1.32
Informative	Sarcasm	1.19
Emotional	Response Other	1.14
Response Other	Thanks	1.00

Table 4.4: Distribution of the 10 most frequent dialogue act pairs for turns with more than 1 label.

for a classification task in a real scenario, there is a large tail in the number of classes.

To overcome the long-tail problem, the proposed model uses few-shot learning approach base on siamese networks [75]. Siamese networks have two or more identical sub-networks as depicted in Figure 4.4. Siamese networks perform well on similarity tasks and have been used for tasks like sentence semantic similarity, recognizing forged signatures and other tasks [95].

The tweets' text are the inputs to the network, and are zero-padded sequences of word indices. The inputs are vectors of fixed length, where the model ignores the zeros entries and the nonzeros indices uniquely identify words. The input vectors are then fed into the embedding layer, which looks up the corresponding embedding representation for each word and build a embeddings matrix. The proposed model rely on the use of word embeddings based on FastText [89] which improve Word2Vect [88] subwords learning approach that is more suitable for noisy microblogging data to represent the syntactic and semantic irregularities . To further handle noisy data, the model use a double channel for embeddings based on characters beside word embeddings.

The model have two embedded matrices that represent a candidate of two similar tweets that belong to a same label. Then the model feed them into the LSTM and the final state of the LSTM for each question is a n-dimensional vector ($n = 300$). It is trained to capture semantic similarity

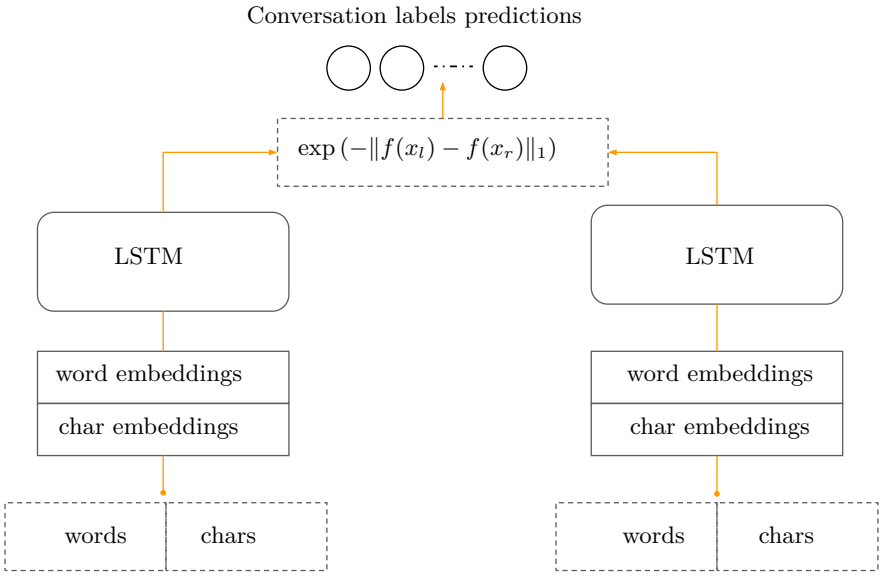


Figure 4.4: CNN architecture for crisis-related conversational modeling.

between tweets that share the same label even if a label have a few instances in training. Then, the two vectors that hold the semantic similarity of each tweet go through a similarity function:

$$\text{sim}\,(x_l, x_r) = \exp\left(-\|f(x_l) - f(x_r)\|_1\right) \tag{4.1}$$

where $\text{sim} \in [0, 1]$, $\text{sim} \in [0, 1]$ is the L_1 norm, and f is the function corresponding to the application of the cloned sequential network to the left/right input. The exponent of the negative the output (the prediction) will be between 0 and 1.

4.3 Experimental Settings

Data Preprocessing: The data preprocessing step normalizes all characters to their lower-cased forms, truncate elongations to three characters, convert every digit to `D`, anonymize twitter usernames to `userID`, and all URLs to `HTTP`. The tokenization of the tweets uses the NLTK toolkit [19], and remove all punctuation marks except periods, semicolons, question, and exclamation marks.
Data Settings: This study focuses on a particular event: earthquake occurred in Ecuador in 2016. The splitting strategy for the data is to use 80% for training (using 10-fold cross-validation for parameter tuning) and hold out 20% of the data for tests. The splitting uses the ski-learn toolkit's module [107], which ensured that the class distribution remains balanced in each subset.
Feature Extraction: The feature extraction uses unigram features from the text of the tweets. For non-neural models, the features are converted to TF-IDF vectors, considering each tweet as

a document. For neural models, the feature extraction creates a dictionary of the vocabulary to map pre-training embeddings. The features contains mainly Words: represented it as binary bag-of-word (BOW) unigrams or dense embeddings, and Punctuation that is present in tweets as part of the BOW or dense embedding. Punctuation features indicates the existence of a question mark or exclamation mark in a turn, e.g., Questions or Greeting Dialog Acts.

4.3.1 Non-neural model settings

This category includes traditional algorithms such as Support Vector Machines (SVM) or Linear Regression (LR), which uses sparse learning representation. The evaluation includes Logistic Regression [38] to compare the neural models with the traditional approaches. The implementation uses sci-kit-learn toolkit [107].

4.3.2 Neural Models Settings

The proposed model trains LSTM networks and optimizes the binary cross-entropy loss using the gradient-based online learning algorithm [71]. The learning rate is the default value, as suggested by the authors, and the number of epochs was set to 25 as the learning curve flattened out. The model uses dropout [126] after embeddings and hidden units to avoid overfitting. A early stopping occurs based on the accuracy of the validation set. The experiments evaluates several dropout rates ($\{0.0, 0.2, 0.5\}$) and minibatch sizes ($\{32, 64, 128\}$), and consider the most common $P\%$ ($P \in \{80, 85, 90\}$) words in the training corpus to limit the vocabulary size (V) . The initialization of the word embeddings uses pre-trained embeddings. For the activation functions, the model uses rectified linear units (ReLU) and filters (f) $\{100, 250, 500\}$ with window size (L) of $\{3\}$. Then, the model applies a pooling length (p) of $\{3\}$, and $\{250, 500\}$ dense layer units. The experiments include fine-tuning all the hyperparameters on the validation set. Beside the proposed model, the experiments evaluate standard CNN and LSTM architectures for text classification.

4.3.3 Evaluation metrics

The evaluation compares non-neuronal models (with scattered text representation) and neural models (with a distributed representation of text) using textual characteristics as inputs for both models. The experiments use a multi-label configuration, given an overlapping of multiple dialogue acts for each tweet. The experiments perform a binary classification task for each label depending on the class sets defined in Section 4.2.3. For each class classification task N, the model predicts each turn in the conversation as belonging to the label or not. This configuration assigns a binary value to each turn in the conversation for each label (that is, for the experiment of 6 classes, each shift receives a binary value indicating whether it belongs to each of the labels). Therefore, for each class experiment N, there are N binary tags, for example, a turn can be Informative, Offer, Request.

Next, the experiments aggregate the N binary predictions for each turn, then compare the resultant prediction matrix for all turns to the majority-vote ground-truth labels, where at least 2 out of 3 annotators have selected a label to be valid for a given turn. The task difficulty increases as the number of classes N increases due to the number of labels to learn. For instance, for the 6-class problem, there are six classification tasks per turn, for the 8-class problem, there are 8, and so on for 10-class and all-class problems). The experiments use weighted F-macro as an evaluation metric to calculate the final scores to account for the inherent imbalance of label-distribution in the data (Figure 4.3). For each feature set, F-macro finds the average of the metrics for each label weighted by the number of actual instances for that label [108].

4.4 Results and Discussion

This section presents the experimental results and discusses the findings as detailed next.

4.4.1 Is this conversation related to a crisis event?

This question seeks to know if a given conversation is related to a crisis event, i.e. the experiments only consider the label that indicates if the whole conversation is crisis-related. To that end, the experiments perform a binary classification task using a specific label in the annotation taxonomy that categorizes a conversation as crisis-related. Table 4.5 presents the results of binary classification using baseline models as well as siamese LSTM neural architecture for few-shot learning. The results show the standard LSTM is slightly better than LR and other neural models, while the proposed model performs better that LR and CNN models. Although LSTM has a better overall F_1 score, the improvement is marginal over the proposed model. Due to the short and noisy nature of the Twitter data, non-sequential models perform well for classification tasks, even although they do not account for the order and sequential nature of the data.

Model	Set	Precision	Recall	F1
LR	cv	65.47	62.31	63.24
LR	test	63.67	61.05	61.90
CNN	cv	66.61	66.02	66.28
CNN	test	62.45	62.87	62.64
LSTM	cv	65.89	67.05	66.28
LSTM	test	66.04	66.51	66.25
LSTM-s2s	cv	64.99	66.04	65.27
LSTM-s2s	test	65.03	65.40	65.14

Table 4.5: Models performance for the task of identifying crisis-related conversations.

Additional experiments evaluate the ability of the models to identify whether a conversation is

related to a crisis considering: only the initial tweet (root), only the replies, or the whole conversation. This study hypothesizes that replies do not have enough context information to find out if a conversation is related to a crisis event. Figure 4.5 supports the hypothesis that it is challenging to try to find out if isolated tweets (replies) are crisis-related. For initial tweets is more feasible as those usually give enough information or context to infer that it is a crisis related tweet. The performance improves when the experiments consider all tweets.

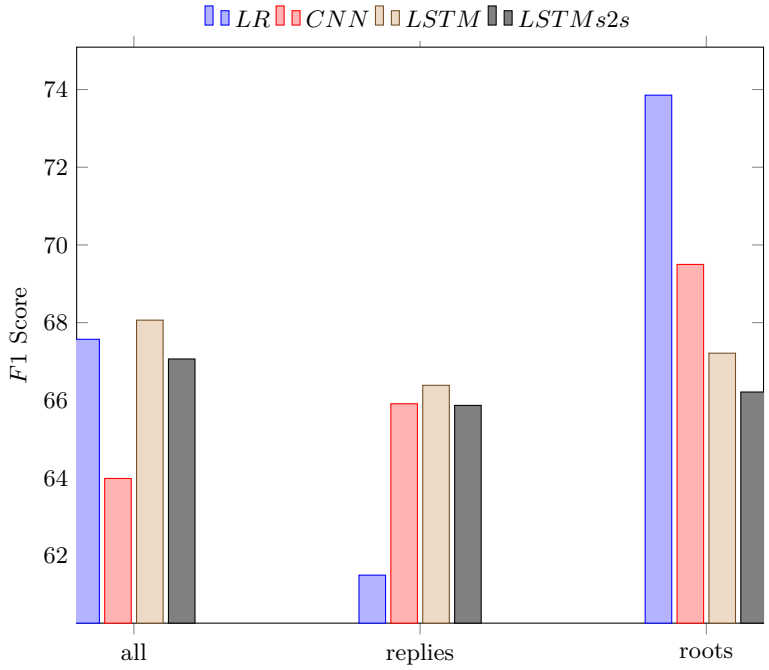


Figure 4.5: Identify crisis-related conversations.

4.4.2 Identifying more than one dialog acts per tweet

This study performs experiments to train models to learn to identify multiple dialogue acts per tweet in the dataset. The experiments test each classifier on each of the four class sets and reporting weighted F-macro for each experiment. The Table 4.6 shows the group of dialog acts or labels used for the experiments. The grouping of the labels into categories si based on the distribution of label observed as a result of the annotation process (Table 4.2). The binary category is an exception, and this class indicates whether the conversation is related or not to a crisis. The top four classes form the category 4 class, followed by 8 class which in addition to the previous 4 class include four additional classes. The 12 class includes the four more classes, and the final category includes all classes.

Figure 4.6 shows the results of the experiments and the performance drop as the number of

Experiment Group	Dialogue Acts
2 class (binary)	Binary
4 classes (easy)	Sarcasm
4 classes (easy)	Informative
4 classes (easy)	Response Other
4 classes (easy)	Emotional
8 classes (medium)	Expressive Positive
8 classes (medium)	Wh Question
8 classes (medium)	Complain
8 classes (medium)	Expressive Negative
12 classes (hard)	Thanks
12 classes (hard)	Insult
12 classes (hard)	Request Info
12 classes (hard)	Opening Greeting
all classes (hard+)	All others

Table 4.6: Dialogue acts used for experiments.

classes increases. For LSTM, the predictions improve for the all-Class setting. The performance in the 8-Class category is steady because of the addition of lexically distinct classes Request and Thanks with the addition Even with the addition of straightforward classes in 12-class category the performance drops for all models but the proposed LSTMs2s. The results show the effectiveness of few-shot learning approach for handling large number of labels.

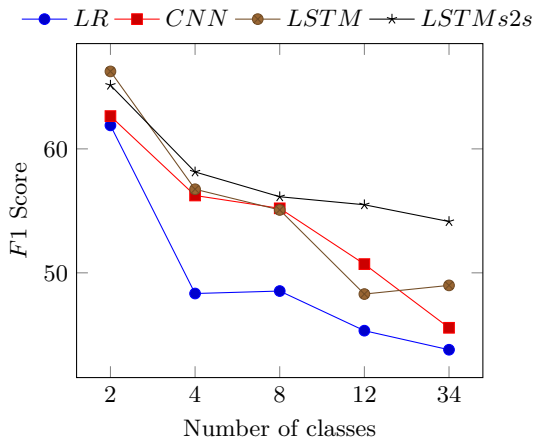


Figure 4.6: Multi-label classification of dialog acts for crisis-related conversations.

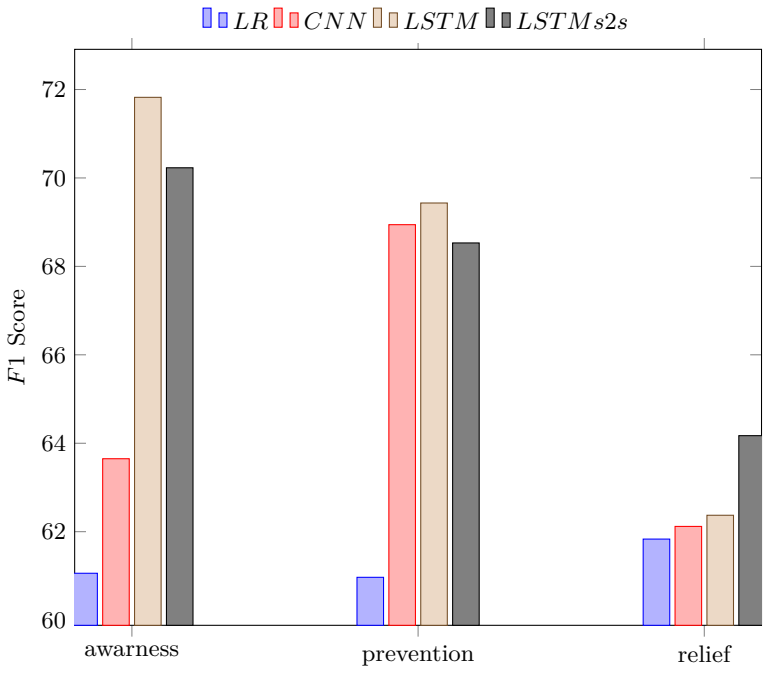


Figure 4.7: Identify crisis-related conversations outcomes.

4.4.3 Conversation outcome analysis

The next set of experiments aims to show the ability of the models for dialogue act classification as a method for inferring semantic intent in a conversation to derive insights to improve automated systems or support decision tools for crisis events. The data used include the conversation context, i.e. the whole conversation. The models predict on one of three problem outcomes: acknowledge prevention, situational awareness, and relief coordination. Figure 4.7 shows the results for each outcome, with neural architectures performing well for prevention and situational awareness outcomes, but for relief coordination where there is still room for improvement due few instances.

4.4.4 Discussion

There are several issues in conducting experiments that are important to mention. Given the conversational nature in the Twitter platform, the conversation might have several overlapping conversations, i.e., conversations can have several branches, the experimental settings treat each branch as a conversation, which can cause overlapping tweets. The labels are a scarcity in languages other than English limit the capacity to use additional annotated data from other events to augment the dataset. Also, increasing the number of labels bring difficulties for human annotator, which causes the agreement score to drop.

The ambiguity in some tweets can be challenging to spot even for human annotators, and a

limitation of the annotation process is the number of annotations per tweet compared to previous works. Similar to previous works on classification tasks, the corpus has a long tail of classes after the annotation process. Because of the size of the dataset and the labels imbalance, it is difficult for models to learn to predict from very few instances for some classes. The proposed model help to improve the performace of LSTM model to handle this situation.

The dataset splitting strategy uses random splitting of the tweets, but further experiments should consider time-based splitting, as it is a more difficult scenario. In this sense, models should be able to learn from incomplete conversations until more replies arrive. A final aspect that deserves consideration is the models' performance, as there is still room for improvement using different architectures.

4.5 Conclusion

This chapter explores how to identify dialog acts in crisis-related conversations on Twitter and extract insights concerning their outcomes. It introduced a novel taxonomy of fine-grained conversation acts tailored for the crisis management domain and gather annotations for Twitter conversations in Spanish about the earthquake that occurred in Ecuador in 2016. The annotation process shows that conversation acts are often semantically overlapping, and the experiments train a multi-label classification model to predict labels for each turn under a different number of classes. To that end, this chapter proposes a few-shot learning approach based on siamese neural architecture, and the results show that outperforms baseline approaches in settings with large number of labels. Future research directions include improving the performance of the models through the exploration of different neural architectures and using recent improvements in learning representations.

Chapter 5

Crosslingual Classification

This chapter addresses RQ3 related to handling the scenario of multiple languages for a specific crisis-related event. This chapter starts with an overview and motivation for cross-lingual classification in Section 5.1. Next, Section 5.2 provides a description of the methodology and the cross-lingual classification task evaluation. Then, Section 5.3 details the experimental settings in and Section 5.4 show the empirical results. Finally, Section 5.5 outlines conclusions and research directions. The work presented in this chapter was first published at The 2019 World Wide Web Conference [135].

5.1 Introduction

During crisis events like earthquakes, user-generated data on social media can provide valuable information to humanitarian organizations such as the United Nations, Red Cross, and also activists working in the relief efforts. Social media leverage the power of the crowds to provide awareness of the situation often faster than traditional media, allows them to respond quickly to the urgent needs of affected people, assess damages in the buildings and infrastructure, identify medical emergencies, or coordinate relief actions [57, 26]. With the worldwide reach of social media, users from different countries and languages can react and interact in any crisis event. This chapter aims to identify the crisis-related tweets in a multi-lingual scenario and characterize them in the context of conversations. To that end, this chapter introduces a new annotated dataset about the earthquake that occurred in Ecuador on April 16, 2016. The corpus contains 8360 tweets annotated for English and Spanish languages. Table 5.1 shows an example of a conversation between several users on Twitter. Based on the metadata (in-reply-to-status-id field) provided by Twitter, it is possible to collect all the tweets that belong to conversations.

Despite recent progress in natural language processing (NLP), the semantic interpretation of noisy short-texts remains a hard problem. A multi-lingual scenario complicates, even more, the task of natural language understanding. For instance, prior annotation taxonomies [60] define that

User	Tweet	In reply to
BBCBreaking	Ecuador declares state of emergency in six provinces after powerful earthquake kills at least people.	
RNexists	is the current count!	BBCBreaking
MaJoJovi	The current count is , I'm from Ecuador. Please pray for us.	Rnexists
1SHeRA1	I am and will, my dear. I hope we can help in some concrete ways as well.	MaJoJovi
MaJoJovi	we need all the help we can get. Your prayers helps too. Thank you.	1SHeRA1

Table 5.1: A sample conversation about the 2016 earthquake in Ecuador initiated by an organization's Twitter account.

a tweet can belong to one of several categories such as: statistics about affected people, emotional support, or helping through donations, goods, or volunteers. To categorize tweets, human annotators have to deal with some issues associated with social media data that include: a) associating a tweet to a category can be difficult due to ambiguity, even human annotators may differ in their judgment about whether or not a tweet belongs to a specific category, b) the noisy nature of the tweets, as well as the idiomatic phrases, can make it difficult to train models and infer across languages.

This chapter evaluates neural architectures to identify crisis-related tweets across multiple languages and it uses the Ecuadorian earthquake as study case. The contribution of the work is as follows. a) introduce an annotated corpus of crisis-related tweets for Spanish and English language, b) evaluate deep contextual neural architectures for the multi-lingual classification task at hand, and c) characterize the conversations from locals and foreigners about the study case earthquake. The dataset and the code are available at Github[1].

5.2 Methodology

5.2.1 Data

The Appendix A details the data acquisition, storage, and processing of the dataset. The Table 5.2 summarizes some statistics about the dataset. The first section shows all the tweets in the dataset split by language: Spanish, English, and other languages.

The conversations row refers to the number of conversations by language, identifying each conversation by its initial tweet. In total, 55% of the tweets in the dataset belong to a conversation. Although the average of the number of replies and users vary across languages, the median resulted similar for all languages (num replies = 3 and num users = 2) (the minimum is due to the constraint in the preprocessing of the conversations. The average is far from the median for the number of

[1] https://github.com/johnnytorres/crisis_conv_crosslingual

	Spanish	English	Other	Total
Tweets	93,405	38,533	20,331	152,269
Users	50,758	25,880	10,387	87,025
Avg. tweets	1.84	1.49	1.96	1.75
Conversations	4,632	1,092	377	6,101
Replies	50,747	17,989	9,506	78,242
Tweets	55,379	19,081	9,883	84,343
Avg. replies	11.96	17.47	26.21	13.82
Avg. users	8.34	13.65	16.36	9.79

Table 5.2: Statistics of the tweets by language. The first section refers to all tweets in the dataset. The second section refers to the tweets that belong to conversations with at least one reply.

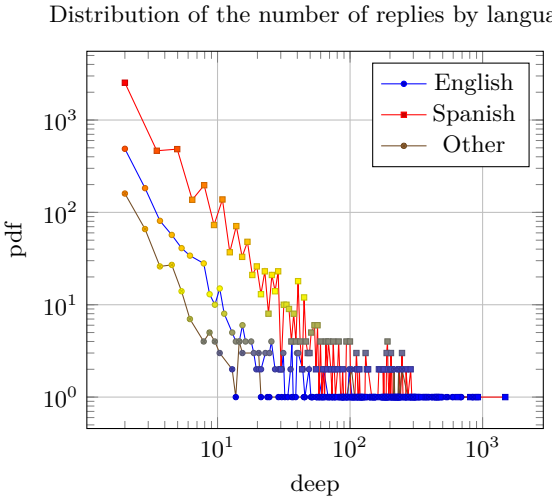


Figure 5.1: Distribution of the number of replies by language for the dataset.

replies and users, indicating some outliers (i.e., some popular conversations often initiated by influencers). The average of other languages is higher than Spanish or English due to outliers in the number of replies to a specific type of conversation (e.g., games or sports). Although tweets in other languages belong mostly to conversations in English, the initial tweet contains multimedia (images or video) or limited text that difficult language detection by the Twitter API. Figure 5.1 depicts the distribution of the number of replies on the conversations.

Further analysis performs several steps of preprocessing to the conversations, e.g., filtering out conversations with non-alternating users' tweets, i.e., replies from different users. Also, filtering out conversations with less than 3 tweets and more than 10 tweets[2]. After the preprocessing, the dataset contains 518 Spanish and 172 English conversations.

[2]The lower bound allow at least two replies in the conversation, and the upper-bound is replies ≤ 10 which accounts for 84% of the conversations.

5.2.2 Annotation

The annotation task uses 518 Spanish, and 172 English conversations, including the replies, account in total 2193 and 730 tweets Spanish and English, respectively. The annotation task relies on the multi-class taxonomy [60] where each tweet can belong to one of the following categories: a) Injured or dead people b) Missing, trapped or found people c) Displaced people and evacuations d) Infrastructure and utility damage e) Donation needs or offers or volunteering services f) Caution and advice g) Sympathy and emotional support h) Other useful information, and i) Not related or irrelevant

The annotation task involves two undergraduate students as human annotators and presented them with the tweets selected for annotation. The annotators have instructions describing each of the categories and associated examples. For each tweet, the annotators select only one category for each tweet-based taxonomy. A third annotator resolves the discrepancies or disagreements between the two annotators. Similar to annotation procedures in previous works [60], the annotator to make judgments based on the text only, even if Twitter APIs truncated the text during the data collection. The annotators are not allowed to open any link inside the text of the tweets as the experiments seek to use only the available text for training the machine learning models.

The annotation quality analysis uses the percentage of agreement (p_o) between the two annotators is an average of the coincidences because each tweet in the annotation process can have only one label. Additionally, the annotation process uses Fleiss' Kappa (κ) [39] to measure the quality of the annotations, defined as $\kappa = (p_o - p_e)/(1 - p_e)$, where p_o define the empirical probability of agreement (i.e. the observed agreement percentage), and p_e accounts for a random agreement between annotators. Table 5.3 shows that the per-label agreement varies from $\kappa = 0.75$ for lexically well-defined categories (such as caution and advice and infrastructure damage) and agreement $\kappa = 0.45$ for less clearly-defined categories (e.g., displaced and evacuations). The categories with few instances often have low scores because of an error in the annotation effects by a large margin of the agreement score. Also, in cases where tweets belong to several categories, the agreement score might decrease even for categories with a large number of instances.

5.2.3 Model

This section evaluates different approaches for modeling crisis-related tweets on multi-lingual settings for the annotated corpus. First, it describes the learning representation for the tweets, then the neural architecture for classification used, and finally, the experimental settings.

For tweets' text representation, traditional approaches require manually engineered features like cue words or sparse vector representation such as TF-IDF used in previous works [57]. This chapter relies on the use of dense representation such as word embeddings Word2Vec [88]. [88] proposed an unsupervised language model using two log-linear models for computing dense representations from large (unlabeled) corpus efficiently: a) bag-of-words model CBOW that predicts the current word

	Spanish				English		
Category	N	*po*	*pe*	κ	N	*po*	*pe*	κ
Injured or dead people	624	0.69	0.12	0.65	1,165	0.64	0.12	0.59
Missing or found people	30	0.77	0.27	0.68	4	0.75	-	0.75
Displaced people and evacuations	12	0.50	0.08	0.45	7	0.57	-	0.57
Infrastructure damage	157	0.64	0.07	0.61	73	0.75	0.11	0.72
Needs or offers	234	0.67	0.11	0.63	207	0.64	0.12	0.59
Caution and advice	61	0.77	0.08	0.75	39	0.64	0.18	0.56
Emotional support	451	0.67	0.10	0.63	325	0.62	0.10	0.59
Other useful information	753	0.69	0.10	0.66	426	0.59	0.13	0.53
Not related or irrelevant	1,846	0.69	0.11	0.66	1,946	0.66	0.11	0.61

Table 5.3: Agreement statistics for Spanish and English tweets.

based on the context words, and b) a skip-gram model that predicts surrounding words given the current word. The approach shows that both models can learn high accuracy syntactic and semantic regularities and overcome the issues in sparse representation models.

Word2Vec [88] represents each word in the corpus like an atomic entity and generates an embeddings vector for each word. In this aspect, Word2Vec and Glove [109] are similar; both define words as the smallest unit to train. However, Word2Vec does not take advantage of the global context. Both CBOW and Skip-Grams are predictive models and only use local contexts during training. In contrast, GloVe leverage the same intuition but uses a neural method to decompose the co-occurrence matrix into more expressive and dense word vectors. N-gram feature is a critical improvement in FastText [89] compared to Word2Vec, and it aims to solve the out-of-vocabulary (OOV) issue. FastText enables word embeddings to encode sub-word information and produce more accurate vectors than Word2Vec. Most recently, deep contextualized embeddings such as ELMo [110] and Flair [6], generate embeddings for a word based on the context, thus generating slightly different embeddings for each word depending on the context of its occurrence. The models leverage recent advances of the learning representation models that use the pre-trained embeddings and then fine-tune the embeddings to the specific dataset.

This chapter is interested in identifying if a given tweet is related to a crisis event and frame the problem of detecting crisis-related tweets as a multi-class classification task. To that end, the classification tasks evaluate Convolutional Neural Networks (CNN) and LSTM sequence models using word embeddings as learning representation. Due to the unstructured, short, and noisy nature of the Twitter data, CNN models have shown to perform well for short-text classification task [100]. The training of the models optimizes the binary cross-entropy loss using the adaptive gradient-based learning algorithm [71], with the learning rate and parameters set to the values suggested by the default. The number of epochs is 10 for the case of the random embeddings and 5 when using the stacked embeddings with Glove-Flair [6]. The models use dropout [126] after embeddings and hidden units to avoid overfitting. The initialization of the word embeddings in L is random

in the case of CNN. LSTM model uses word embeddings with random initialization and Stacked Glove-Flair pre-trained.

5.3 Experimental Settings

5.3.1 Data Preprocessing

The preprocessing step normalizes all characters to their lower-cased forms, truncate elongations to three characters, convert every digit to `D`, anonymize twitter usernames to `userID`, and all URLs to `HTTP`. The preprocessing step remove all punctuation marks except periods, semicolons, question, and exclamation marks. The tokenization of the tweets uses the NLTK toolkit [19].

5.3.2 Label Grouping

Due to the imbalance of the labels, it is necessary to group the labels for a binary classification task to identify whether a tweet is related to a crisis event or not. A given tweet is related to a crisis if it belongs to any of the categories but Not related or irrelevant. The Table 5.4 shows the grouping labels used for the experiments. The grouping into two classes achieves a reasonable balance for both Spanish and English languages.

Label	Spanish (es)	English (en)
Crisis related	2322	2249
Not related	1846	1946
Total	4168	4195

Table 5.4: Dataset for the binary classification task.

5.3.3 Splitting Strategy

The splitting strategy for the dataset is to use 80% for training and development (using 10-fold cross-validation) and hold out 20% of the data as a test set. A random split of the dataset into train and test sets to ensure that the class distribution remains reasonably balanced in each set.

5.3.4 Classification Tasks

There are two sets of classification tasks for the experiments, as described next. First, the experiments train and test models using the same language: train and predict on the Spanish dataset, and similarly, train and predict on the English dataset. Second, the language is different for training and testing, e.g., training a model on the Spanish dataset and predicting on English dataset.

model		LR			LSTM			CNN			LSTM Stacked	
lang	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
es-es	85.70	85.48	85.52	85.78	85.12	85.18	86.30	85.71	85.77	81.49	81.17	79.73
en-en	93.30	93.21	93.21	92.89	92.95	92.92	94.30	94.04	94.05	91.06	91.35	91.04
es-en	79.05	78.45	78.47	76.23	74.68	74.59	79.74	79.71	79.65	85.90	85.90	85.88
en-es	57.73	45.10	29.97	56.32	44.62	28.39	52.27	44.67	29.31	79.57	77.50	77.49

Table 5.5: Performance of the models in single language and multi-lingual classification.

5.3.5 Classification Models

The experiments evaluate a baseline classifier based on Linear Regression (LR) as well as models based on neural architectures such as CNN and LSTM. Neural models use Stacked Embeddings (Glove word embeddings + Flair deep contextual embeddings).

5.3.6 Evaluation Metrics

The experiments use *Precision*, *Recall*, and macro F_1 metrics to evaluate the performance of classification models and report the results on the test set.

5.4 Results and Discussion

The Table 5.5 shows the results of the classification task. The first column shows the source language in which is trained the model and the target language that predicts. Each of the models have three columns associated that represent the metrics precision (P), recall (R), and F_1 score.

For the first experiment, the CNN model outperforms other models training and predicting for a single language. The performance of the baseline model LR is on par with CNN and better than the LSTM model. The results mean TF-IDF based models are strong baselines in noisy short-text classification under a single language and single event but often fail to generalize for new events due to the out-of-vocabulary (OOV) issue. The LSTM and CNN are using randomly initialized embeddings and fine-tuned during training, which hinders the performance compared to using pre-trained embeddings. In this experiment, the LSTM with multi-lingual embeddings did not perform well, mainly because it does not apply the fine-tuning to the embedding weights.

The second experiment aims to train a single model for predicting tweets in another language (row 3 and 4 in Table 5.5). Traditional approaches as LR using TF-IDF fail to generalize, and the performance fall in the case es-en and drops drastically in the case en-es. The reason that the first case is not as critical as the second is not apparent and requires further analysis. There is a small percentage (4%) in the Spanish dataset with a different language, but not enough to affect the training of the model. In the case of the English dataset, there is 10% of tweets with a language different from English. However, the model LSTM with multi-lingual Stacked Embeddings

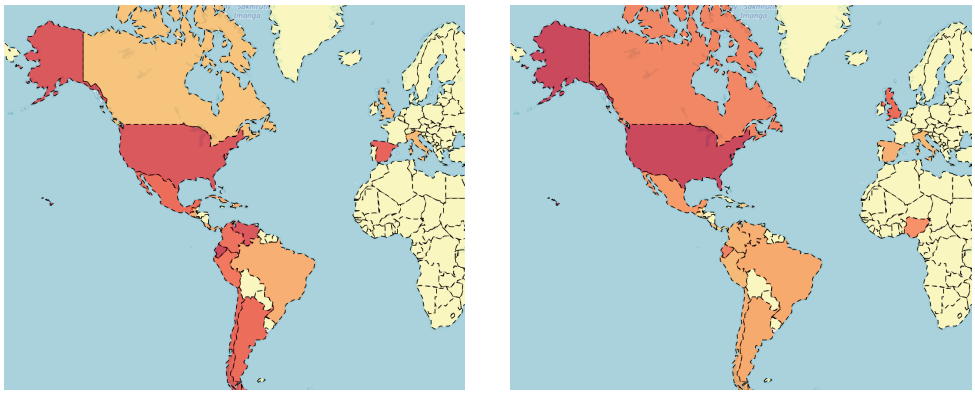
generalizes well in this setting and outperforms other models, which is promising for tackling the multi-lingual scenarios in detecting crisis-related tweets. The next sections detail some of the findings related to the crisis-related conversations, cross-lingual analysis, and current limitations in this research.

5.4.1 Crisis-related conversations

This study explores the conversational nature of the interactions on Twitter and how to improve the extraction of insights during crisis events by analyzing entire conversations. The analysis focuses on both Spanish and English conversations. By considering only the root or initial tweets in the conversation, approximately 21% of them are not related to the earthquake, while the rest of the root tweets belong to the other categories. The replies to root tweets not related to the crisis (97%) are also not related to the earthquake. However, there is a remaining 3% of replies annotated as donations, emotional support, and other useful information even when the root or parent tweet of the conversation is not related to the crisis. The small percentage of non-related conversations indicates that analysis might lose some information even when the root tweet is not related to the crisis. A similar situation occurs in the replies to root tweets that are related to the crisis, for instance, 24% of the replies annotated as not related to the crisis, in the case of English conversations up to 35%.

In this case, these tweets could help to analyze the objective or outcome of the conversation.

5.4.2 Cross-lingual analysis



(a) Spanish crisis-related tweets. (b) English crisis-related tweets

Figure 5.2: Spatial distribution of crisis-related tweets, darker color indicates larger percentage of tweets.

The further analysis uses the best model to categorize the complete dataset to identify tweets related to the crisis event. First, the analysis determines the country of the users in the conversations using the location field in the metadata of the tweets. Approximately 30% of the users do not provide

the location, and 15% contain noisy information in location; therefore, this analysis considers only those with valid location information. The Figure 5.2 shows the spatial distribution for each of the languages. The Figure 5.2a show the crisis-related tweets in the Spanish language with Ecuador as the predominant country, which is where the earthquake occurred, followed by the United States, Venezuela. There is a significant percentage in Spain, where there is a large number of Ecuadorian immigrants. The Spanish dataset contains the location information for approximately 49% of the tweets. The Figure 5.2b shows that the Spanish-speaking countries decrease their participation in the English dataset. The United States is the country where most tweets in English are posted, followed by England and Canada. The English dataset contains the location information for approximately 77% of the tweets.

Another aspect is the percentage of the annotated tweets associated to each category depending on the language and location. Most of the tweets (3.57%) related to Needs / Offers for goods and services came from the affected country, while 3.24% came from other countries. The majority of the tweets related to statistics about deaths and injured people come from other countries (13.87), while the tweets in the affected country are 2%. The tweets related to Needs / Offers for goods and services that came from the affected country is almost inexistent, while 4% came from other countries. The majority of the tweets related to statistics about deaths and injured people come from other countries (29.46%), while tweets in the affected country are less than 1%.

5.4.3 Limitations

This study identifies several issues while conducting experiments that are important to mention. Due to the limited size of the annotated dataset, further analysis is necessary to characterize the tweets from local users where the event occurred vs. foreigners, and the cross-lingual differences. Given the conversational structure in the Twitter platform, the conversations might have several overlapping tweets, i.e., conversations can have several branches, and this study treats each branch as an independent conversation, which could cause duplicated tweets in the dataset. The dataset splitting strategy performs a random splitting, but further experiments should consider time-based splitting, as it is more similar to a real scenario. In this sense, models should be able to learn from incomplete conversations until more replies arrive. The use of the conversational context for learning to predict a new tweet could improve the performance of the classification models; thus, it constitutes a future research direction, especially using attention mechanisms [137].

Due to the ambiguity in some tweets, assigning labels can be confusing even for human annotators, and a limitation in the annotation process is the number of annotations per tweet compared to previous works. Similar to previous works on crisis-related datasets, the corpus has a long tail of classes with few instances. Due to the imbalance in the labels, it is difficult for classification models to learn from very few instances. Some strategies to overcome the imbalance include the implementation of Zero-shot [162] or Few-shot [114] learning for labels with few instances. The scarcity of

annotated data in the Spanish language limits the capacity to use additional annotated data from other events to augment the dataset. Further evaluation of new cross-lingual embeddings and language models are essential for transfer learning approaches to overcome the scarcity of annotated data in low-resource languages [54].

5.5 Conclusion

This chapter introduced an annotated corpus in Spanish and the English language for the earthquake that occurred in Ecuador on April 16, 2016. The annotated corpus considers not only isolated tweets but those that belong to conversations regarding the earthquake, which enable future research for a more in-depth understanding of that type of interaction and their outcome. The findings show that tweets often overlap semantically, approximately 15% of the tweets that can belong to multiple labels, an indication that a multi-label annotation would be more suitable for a more in-depth understanding of more complex interactions such as conversations. The experiments explore how to identify crisis-related tweets on multi-lingual settings leveraging advances on multi-lingual deep contextual embeddings. The results show that multi-lingual embedding outperforms other approaches based on sparse representation; however, for a single language, a simpler model still performs better. Future research directions include a comparative analysis of cross-lingual modeling of crisis-events using additional datasets publicly available. Also, future work could tackle the learning of labels with few instances, as in the corpus, through zero-shot or few show learning methods.

Chapter 6

Context-aware Classification

This chapter addresses RQ4 that seeks to evaluate the effect of using conversational context in classification. This chapter begin with an overview and motivation for context-aware classification in Section 6.1. Next, Section 6.2 provides a description of the methodology and the crosslingual classification task evaluation. Then, Section 6.3 details the experimental settings in and Section 6.4 show the empirical results. Finally, Section 6.5 outlines conclusions and research directions.

6.1 Introduction

The analysis of messages posted on microblogging platforms such as Twitter can help humanitarian organizations (e.g., United Nations, Red Cross) or activists during crisis events. Social media platforms allow them to gain situational awareness, know the urgent needs of affected people, assess critical damages in the infrastructure, identify medical emergencies in different locations, or coordinate relief or rescue actions [139].

The content generated by users in social networks contains information that can be useful for disaster relief. One of the limitations in the management of social network data during Crisis events is related to the analysis of short and noisy text messages, which are posted by users during crisis events such as disasters and emergencies [26].

Previous works have used or analyzed tweets individually in downstream NLP tasks [57, 100]. Often tweets belong to conversations as depicted in Figure 6.1, which represent a graph of interactions for a set of sample conversations during a crisis event (Pakistan earthquake in 2013). Conversation seed tweets (blue nodes) have links to their replies (orange nodes). The annotated tweets can either be a conversation starter tweet (large blue nodes) or a reply (large orange nodes). This chapter seek to evaluate and gain insights on the potential of using the conversational context in NLP tasks.

In a more detailed example, Table 6.1 shows a sample conversation for the Nepal earthquake in 2015, in which one of the tweets has been collected and tagged (category column) in the CrisisNLP

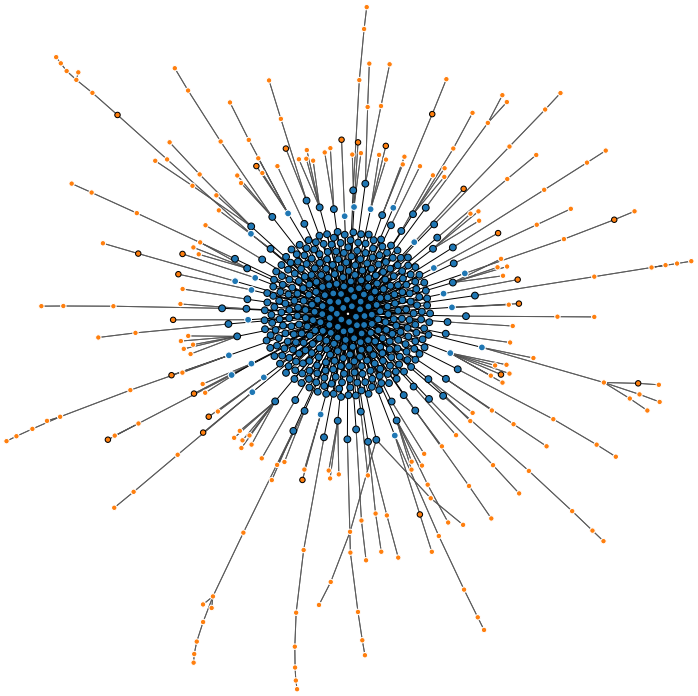


Figure 6.1: Network graph of sample conversations during a crisis event (Pakistan earthquake in 2013). Conversation tweets (blue nodes) have links to their replies (orange nodes). The annotated tweets can either start a conversation (large blue nodes) or belong to the replies (large orange nodes).

corpus [61].

user	tweet	deep	category
Jaimarie13	#Nepal #HELPNEPAL #NepalEarthquake #NepalQuake #Vegas #LasVegas #RescueNet https://t.co/rRDuBDZ6iE	1	
RescueNetOnline	@Jaimarie13 Thank you for the mention! Our team is now on the ground in #Nepal helping w/the relief efforts. #NepalEarthquake #NepalQuake	2	
Jaimarie13	@RescueNetOnline so glad you made it safely! Let me know how I can help! Keep up the great work & be the voice for us back home #NEPAL	3	help_efforts
RescueNetOnline	@Jaimarie13 thank you, we will stay connected and let you know how things are from the field as we're able!	4	

Table 6.1: Sample conversation in the conversational CrisisNLP dataset. The history (highlighted tweets) helps to understand the labeled tweet (indicated by the category column) and could improve downstream NLP tasks.

This study hypothesize that the context of the conversation can improve downstream NLP tasks, such as the case of tweets before the annotated tweet in the sample conversation, therefore seek to

answer the following research questions:

RQ1 Does previous crisis-related datasets sampling contain conversational tweets and wherein the conversation is?

RQ2 Does the conversational context improve downstream NLP classification tasks, and to what extent?

RQ3 What are the categories that best benefit from having conversation context in downstream NLP classification tasks.

6.2 Methodology

To answer the research question RQ1, this study introduce the Conversational CrisisNLP dataset, which augment the original CrisisNLP dataset [60] with the conversational context for the tweets. The Appendix A details the data acquisition, storage, and processing of the dataset.

6.2.1 Data

code	event	ltweets	utweets	tweets	convs	users	langs	ut	$g\%$	$c\%$
EV00	2013-PK-eq	1,881	687	2,568	1,880	1,253	24	2.05	36.52	0.05
EV01	2014-CL-eq	2,685	1,843	4,528	2,671	2,955	13	1.53	68.64	0.52
EV02	2014-CL-eq-en	2,376	2,689	5,065	2,369	2,698	23	1.88	113.17	0.30
EV03	2014-Ca-eq	1,884	230	2,114	1,882	1,384	6	1.53	12.21	0.11
EV04	2014-IN-fl	1,820	1,362	3,182	1,818	1,569	17	2.03	74.84	0.11
EV05	2014-MX-hu-en	1,447	349	1,796	1,441	1,148	8	1.56	24.12	0.42
EV06	2014-PH-ty-en	11,742	5,778	17,520	11,551	11,929	30	1.47	49.21	1.65
EV07	2014-PK-fl	1,769	1,539	3,308	1,768	2,171	25	1.52	87.00	0.06
EV08	2015-NP-eq-en	12,506	23,788	36,294	12,438	21,950	42	1.65	190.21	0.55
EV09	2015-VU-cy-en	2,610	3,358	5,968	2,595	2,644	29	2.26	128.66	0.58
	ALL	40,720	41,623	82,343	40,413	49,701	217	1.66	102.22	0.76

Table 6.2: The conversational crisisNLP dataset statistics.

Table 6.2 describes the Conversational CrisisNLP dataset per event, assigning a code to each event later reference. The name of the event contains the [year]-[country or state]-[event type]-[language] format, when the language is not present means the native language of the country where the event occurred. The augmented dataset have the tweets (ltweets) in the original CrisisNLP [61] corpus plus the unlabeled tweets (utweets) that are either child or parent tweets of some of the labeled tweets in the original dataset. For all events, the total number of tweets ($tweets = ltweets + utweets$) increases compared to the original dataset; in some cases, the percentage of increase ($g\%$) is higher than 100%. The number of conversations (convs) is less than the number of original tweets (ltweets),

which implies that a percentage ($c\%$) of the original tweets belong to the same conversation. Also, most users in the dataset have less than three tweets on average ($ut = tweets/users$).

Due to some inconsistencies in the original CrisisNLP dataset between the labeled data annotated by paid workers and the labeled data annotated by volunteers, this study structured the categories for the Conversational CrisisNLP dataset as detailed in Table 6.3. The agumented dataset increase the total number of tweets (tweets column) for all categories compared to the original tweets (ltweets). The percentage of increase ($g\%$) is significant for some crisis-related categories (help-efforts,people-displaced,sympathy-or-emotional) but also for the non-related tweets.

category	ltweets	tweets	$g\%$
caution-and-advice	1,342	1,725	28.54
help-efforts	4,523	9,415	108.16
help-request	253	471	86.17
infrastructure-damage	2,138	3,289	53.84
not-related	15,947	42,320	165.38
other	9,385	14,771	57.39
people-dead	2,951	3,242	9.86
people-displaced	602	1,298	115.61
people-missing	476	641	34.66
sympathy-or-emotional	3,103	8,615	177.63
ALL	40,720	85,787	47.47

Table 6.3: Category statistics in conversational crisisNLP.

Table 6.4 shows the statistics of the categories for each of the events in the Conversational CrisisNLP dataset. Some categories (help-request, people-missing) have none or few instances, which makes difficult NLP tasks such as classification.

category	EV00	EV01	EV02	EV03	EV04	EV05	EV06	EV07	EV08	EV09
caution advice	69	140	325	76	41	91	348	51	41	160
help efforts	299	581	12	82	45	198	436	463	2,018	389
help request		73	1				71		108	
infra. damage	28	146	43	366	66	466	356	86	348	233
not related	312	161	342	119	498	47	7,006	25	6,720	717
other	724	514	835	948	206	321	3,031	618	1,507	681
people dead	326	453	178	197	893	31	66	220	420	167
people displaced	12	26	63	4	27	101	130	85	89	65
people missing	5	42	6	6	14	28	8	103	243	21
sympathy emotion	106	549	571	86	30	164	290	118	1,012	177

Table 6.4: Category statistic by events in the conversational crisisNLP.

Figure 6.2 shows the temporal distribution of the tweets for each of the events in the Conversational CrisisNLP dataset. Some of the events begin their conversational tweets (blue line) before the date the labeled tweets were collected, due mostly to tweets non-related to the crisis event.

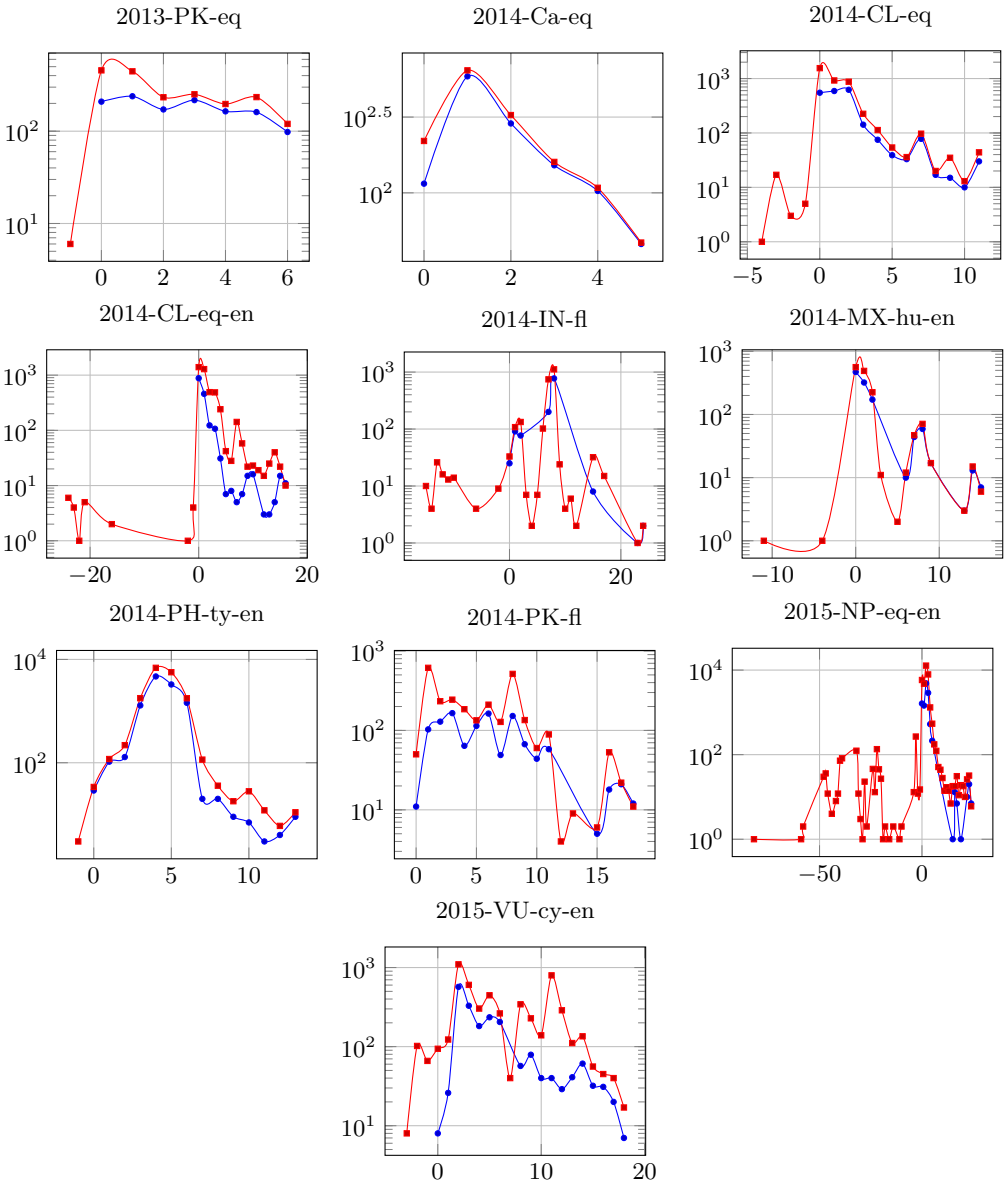


Figure 6.2: Temporal distribution of the tweets in the Conversational CrisisNLP dataset. As expected, the occurrence of the conversational tweets (blue line) begins before the labeled tweets (orange line) for most of the crisis events. The conversational tweets (blue line) contain both labeled and unlabeled tweets.

Figure 6.3 shows the distribution of the number of replies (red line) for conversations in the Conversational CrisisNLP dataset. As expected, the distribution follows a power-law distribution as most human activities. Most conversations are short (i.e., few replies), but some events (2014-CL-eq-en,2015-VU-cy-en) contains a few very long conversations with more than ten tweets.

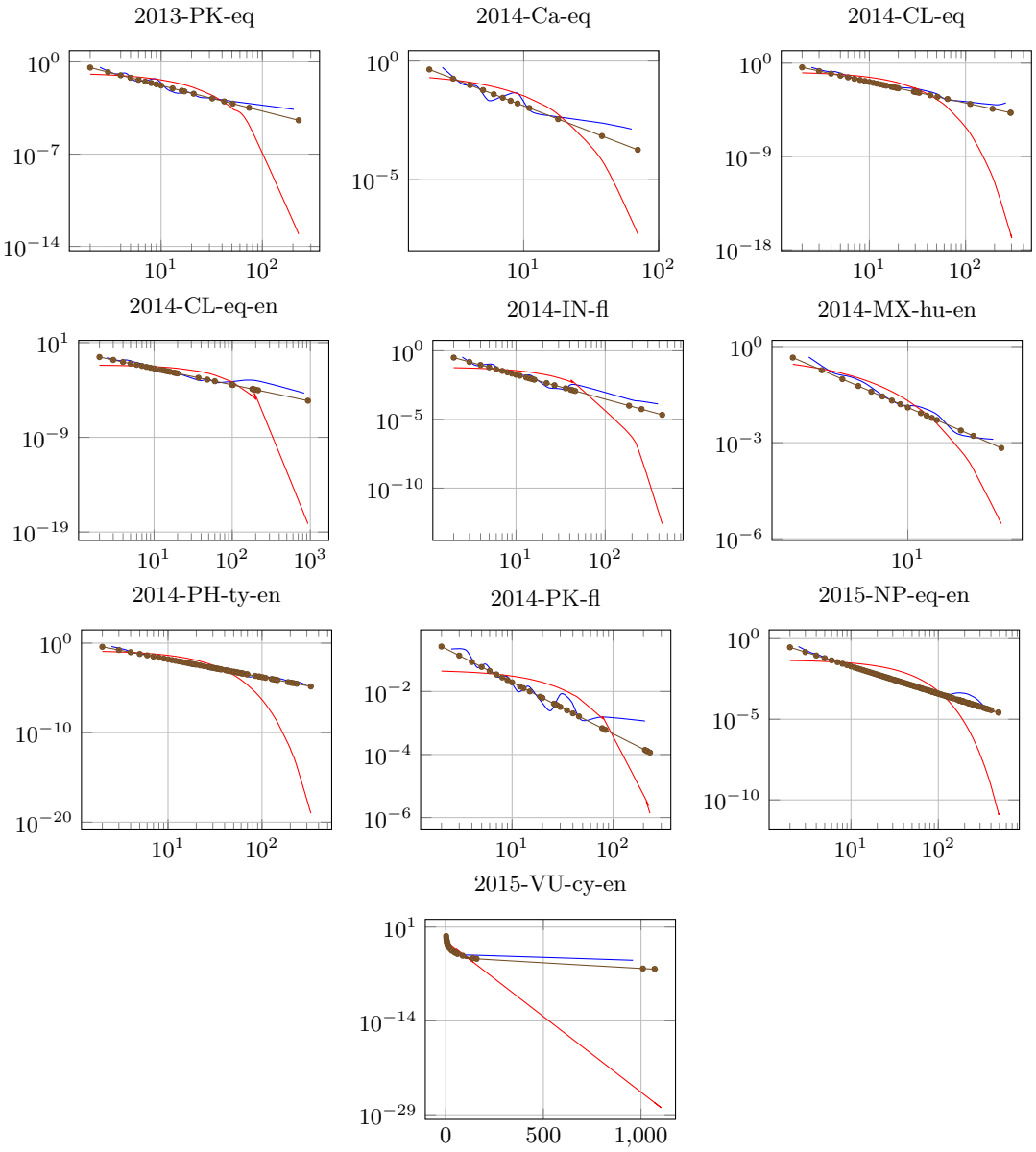


Figure 6.3: Distribution of the number of replies for conversations. The empirical distribution (brown line) follows a power-law distribution (blue line) rather than an Exponential distribution (red line).

6.3 Experimental Settings

This section describe the settings for the experiments conducted. The focus is to know the effects of conversational context in NLP classification tasks (RQ2, RQ3) using existing classification models.

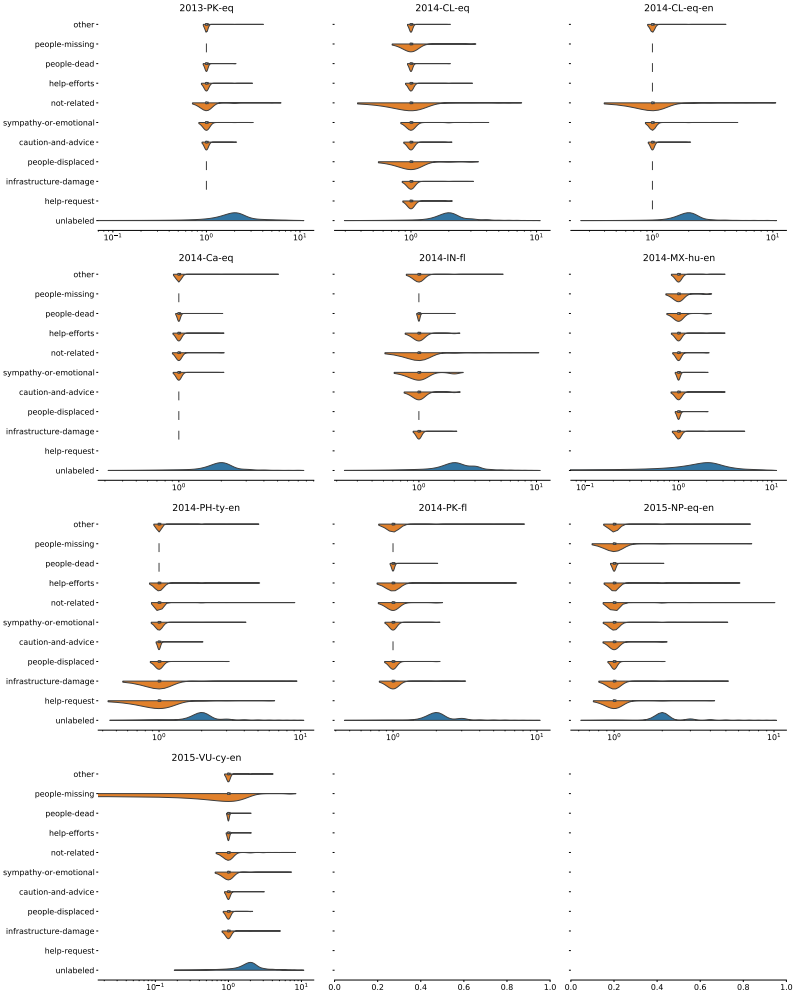


Figure 6.4: Distribution of the deep of labeled (orange curves) and unlabeled tweets (blue curves) in the conversations for each event in the conversational CrisisNLP dataset. The Y axis show the categories and the X axis indicates the probability density estimate of the deep of the tweets.

To that end, the experiments select several models used in previous works [140, 100] suitable for Binary or Multinomial distributed discrete data such as texts dataset.

Naive bayes (NB) parametrizes distribution of the data by $\theta_y = (\theta_{y1}, \ldots, \theta_{yn})$ for each class y, where n is the the size of the vocabulary and θ_{yi} is the probability $P(x_i \mid y)$ of feature i appearing in a sample belonging to class y [113].

Logistic regression (LR) represents the probabilities of the possible outcomes uses the logistic function $f(x) = L/(1 + e^{-k(x-x_0)})$ and minimizes the following cost function

$$\min_{w,c} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \log(\exp(-y_i(X_i^T w + c)) + 1) \tag{6.1}$$

The experiments use the L-BFGS-B optimization method [161].

Random forest (RF) is an averaging algorithm based on randomized decision trees [22]. RF uses n classifiers created with randomness in the construction (the experiments use $n = 5$). The prediction of the ensemble is the averaged prediction of the individual classifiers.

Support vector machine (SVM) constructs a hyper-plane or set of hyper-planes in a high dimensional space. It aims to achieve a good separation between classes by the hyper-plane that has the most considerable distance to the nearest training data points of other classes (the more significant the margin, the lower the generalization error). The experiments use RBF kernel [27].

Convolutional neural network (CNN) is a regularized version of multilayer perceptrons. CNN is fully connected networks; that is, each neuron in one layer connects to all neurons in the next layer. The fully-connectedness of these networks makes them prone to overfitting data. The experiments use the vanilla implementation in [70]

To evaluate the classifiers, the experiments use the following metrics: (a) Precision (b) Weighted F_1 (c) AUC The weighted F_1 metric account for the imbalance of the labels. The experiments use stratified cross-validation ($kfolds = 5$) and report the average of the scores in the empirical results in the next section. The categorical classification task filters the categories with less than $kfolds$ instances. The experiments perform standard preprocessing steps such as replacing tokens such as URLs, Mentions, Reserved words (RT, FAV), Emojis, or Smileys. The preprocessing maintain Hashtags since removing the symbol (#) can be used as an input feature, mainly because there are some tweets that most of their content is hashtags. For the non-neural methods, the experiments use TF-IDF vectorization to represent the features and the implementation uses Sklearn [107]. While the neural models use word embeddings, and the implementation uses Keras [29].

6.4 Results and Discussion

This study conduct several experiments to determine if there are improvements in classification tasks (RQ2) through the use of contextual information provided by conversation history on Twitter. The experiments perform two types of classification tasks: (i) Binary, and (ii) Categorical. Since the Conversational CrisisNLP dataset contains several events, this section report the results per event.

event	NB	LR	RF	SVM	CNN	NB+	LR+	RF+	SVM+	CNN+
EV00	76.63	78.71	80.92	80.27	79.81	76.50	78.76	81.54	80.15	79.68
EV01	91.10	91.10	91.47	91.08	91.10	91.10	91.10	91.45	91.08	91.10
EV02	79.96	87.59	89.17	88.98	93.42	79.44	87.74	89.64	89.57	93.93
EV03	90.63	90.63	90.34	90.72	90.63	90.63	90.63	90.90	90.72	90.63
EV04	90.83	94.15	92.32	94.09	92.67	90.53	94.10	93.07	93.95	92.90
EV05	95.16	95.16	95.22	95.16	95.16	95.16	95.16	95.37	95.16	95.16
EV06	71.49	69.07	65.03	68.57	69.58	71.27	68.62	66.12	67.95	70.65
EV07	97.89	97.89	97.89	97.89	97.89	97.89	97.89	97.89	97.89	97.89
EV08	73.80	72.30	67.06	71.95	71.01	73.75	72.29	67.20	71.70	72.18
EV09	72.84	85.48	83.35	86.65	89.17	73.92	85.78	83.70	86.91	88.19

Table 6.5: Models comparison for the binary classification task using Weighted F_1 metric.

event	NB	LR	RF	SVM	CNN	NB+	LR+	RF+	SVM+	CNN+
EV00	85.69	86.63	72.05	87.25	86.00	86.10	86.77	75.58	87.25	86.37
EV01	71.93	77.68	65.20	76.96	75.55	71.37	78.11	63.12	77.04	75.36
EV02	93.74	94.25	85.11	93.95	95.48	93.93	94.35	87.12	93.98	95.11
EV03	75.29	77.63	58.14	73.49	78.07	76.02	77.71	58.00	73.70	77.92
EV04	98.32	98.48	95.42	97.74	98.04	98.27	98.43	96.87	97.73	98.25
EV05	79.26	80.72	58.66	80.52	76.37	78.79	80.41	57.07	80.70	76.17
EV06	78.01	80.22	70.76	77.08	80.48	77.81	80.22	71.36	76.91	80.81
EV07	73.19	72.54	54.53	65.84	65.81	75.83	72.93	52.17	67.11	70.81
EV08	81.13	82.64	73.30	81.57	82.65	80.79	82.39	73.44	81.24	82.32
EV09	92.68	91.46	85.33	91.52	93.78	92.68	91.58	84.77	91.54	92.66

Table 6.6: Models comparison for the binary classification task using *AUC* metric.

category	NB	LR	RF	SVM	CNN	NB+	LR+	RF+	SVM+	CNN+
caution-and-advice	20.00	39.44	27.39	34.22	26.75	13.33	33.90	21.69	31.84	22.83
help-efforts	59.52	62.12	54.90	58.48	53.80	63.44	61.05	53.10	56.38	48.51
help-request	0.00	20.00	31.51	45.33	36.76	0.00	13.33	19.86	38.00	24.84
infrastructure-damage	11.75	53.76	33.97	47.63	28.39	15.00	55.98	30.20	47.54	28.88
not-related	32.09	43.96	48.66	51.36	50.32	42.52	53.25	51.94	58.52	58.17
other	47.55	46.67	41.50	49.89	43.58	44.60	44.82	39.77	47.36	39.55
people-dead	55.56	62.16	50.27	58.66	67.43	52.33	53.60	39.53	49.97	62.45
people-displaced	0.00	5.00	7.47	27.81	17.61	0.00	6.67	18.95	36.74	17.58
people-missing	5.00	17.96	17.42	26.53	20.87	6.67	22.94	17.89	28.21	17.23
sympathy-or-emotional	46.58	67.79	41.69	65.47	50.86	32.87	70.94	41.28	67.42	48.12

Table 6.7: Models comparison for the categorical classification task using *precision* metric.

6.4.1 Binary classification task

The Binary classification task uses just two categories for the tweets in the Conversation CrisisNLP dataset: (i) Not related, and (ii) Related. The Related category contains all the categories defined in Table 6.3 but not-related. The results are in separate tables for each metric Weighted F_1 (Table 6.5) and *AUC* (Table 6.6). The event column uses the code for each event defined in

category	NB	LR	RF	SVM	CNN	NB+	LR+	RF+	SVM+	CNN+
caution-and-advice	1.38	11.70	12.90	18.15	24.70	0.92	8.52	10.09	13.44	19.97
help-efforts	46.23	54.27	46.74	55.80	53.00	40.11	48.00	40.86	50.35	48.08
help-request	0.00	2.58	18.00	24.62	25.65	0.00	1.72	14.96	20.58	17.87
infrastructure-damage	0.95	9.19	20.41	21.78	24.81	0.91	10.70	17.37	22.20	24.24
not-related	37.30	38.95	47.35	44.19	48.49	49.58	51.83	54.23	55.69	56.65
other	40.18	43.45	37.43	42.28	40.18	36.07	39.24	35.75	37.26	37.79
people-dead	42.84	59.21	53.58	60.14	62.84	30.07	48.98	41.62	50.50	52.20
people-displaced	0.00	0.37	5.61	5.51	11.08	0.00	0.49	11.19	8.13	12.14
people-missing	0.20	7.77	15.79	16.66	14.05	0.27	10.40	15.50	19.06	15.46
sympathy-or-emotional	31.50	45.92	40.78	48.16	51.09	20.92	40.71	36.21	42.56	47.10

Table 6.8: Models comparison for the categorical classification task using F_1 metric

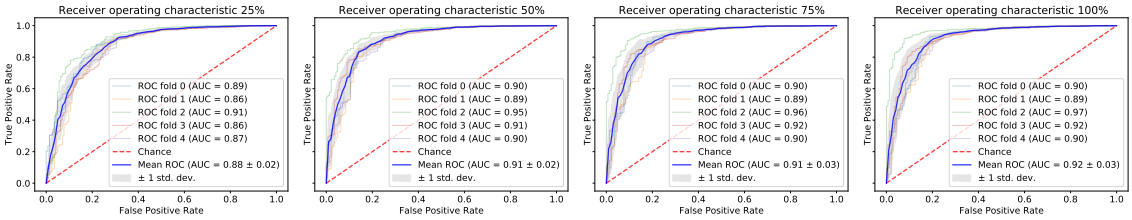


Figure 6.5: Analysis of the effect of using different percentiles $(25, 50, 75, 100)$ from the conversation history for the binary classification task (SVM model) in one of events (2015-VU-cy-en) of the Conversational CrisisNLP dataset.

user	text	deep	category
Jaimarie13	xxbos nepal helpnepal nepalearthquake nepalquake vegas xxunk xxunk url	1	
RescueNetOnline	mention thank you for the mention our team is now on the ground in nepal helping w the relief efforts nepalearthquake nepalquake	2	
Jaimarie13	mention so glad you made it safely let me know how i can help keep up the great work be the voice for us back home emoji emoji emoji nepal	3	help_efforts
RescueNetOnline	mention thank you twe will stay connected and let you know how things are from the field as we re able emoji	4	

Table 6.9: Interpretability model for the neural classifier for a sample conversation.

Table 6.2, while the remaining columns represent each of the models. The line in the middle of the tables separates the models, to the left models that do not use conversational context. To the right of the line are those that use the conversational context. The experiments show mixed results

in the binary classification task for both metrics. For some of the events, the model with the best score (highlighted cells) uses a conversational context. In general, some models improve their scores (blue) compared to their counterparts that do not use conversational context.

6.4.2 Categorical classification task

To answer the research question RQ3, the experiments allow to know the performance of the models across categories. To that end, the experiment defines a specific classification task that uses all the categories as described in the original CrisisNLP dataset [61]. The experiments use the same format as the Binary classification task, i.e., best model per category is highlighted and blue color denotes improvement of models that use conversation history compared to those that do not use it. This task calculates the metrics *precision* (Table 6.7) and F_1 (Table 6.8), and the scores per category are averaged across events.

The results show that there is a significant decrease in the performance across models because the Categorical classification task is more difficult compared to the binary task. The difficulty is due to the imbalance of the labels per category that require zero or one-shot learning approaches [56]. The most problematic categories are help-request, people-displaced, and people-missing. For the F_1 metric, the CNN model outperforms other models in most categories and the worst-performing model is NB. Overall, the results show a similar trend that in the binary classification task, that is in several cases the models do improve their predictions when using the conversation context as opposed to using only the labeled tweet. However, critical categories such as help-efforts and help-request did not improve as expected.

6.4.3 Discussion

This study aims to know to what extent using conversational context could improve downstream NLP tasks (RQ2) thought additional experiments for the binary classification task.

6.4.4 Context sensitivity

The SVM classifier is one of the models with most improvements (see Table 6.6) when using conversational context, thus the experiments use SVM for further experiments that measure the sensitivity of varying the length of the conversation context (history). Figure 6.5 shows the analysis of the effect of using different percentiles $(25, 50, 75, 100)$ for the conversation history in the binary classification task in one of the events (2015-VU-cy-en). Our findings indicate that using a more extended conversation context did improve the classification task from percentile 25 to 50. However, there are not significant improvements using percentile 75 of the history, and finally, using the complete history of the conversations improves w.r.t percentile 75 by 1% on AUC metric. Despite

the gains are not large, using context information can help to disambiguation or handling difficult cases for the models in downstream NLP tasks beyond the Crisis Informatics domain.

6.4.5 Models interpretability

A interpretability model, applied to the sample conversation presented in Table 6.1, allows to validate the initial hypothesis through the use of the UMLFit neural learning framework [54]. UMLFit extracts an interpretation of the classification model based on input sensitivity detected through an external mechanism. The interpretability results fit closely with the initial hypothesis that classifiers can leverage additional data provided by conversation contexts to improve the understanding of the target tweet (labeled tweet) as depicted in Table 6.9. The most important or relevant words (highlighted close to dark red color) are in the target tweet as expected but also in the second tweet (e.g., the word efforts). The lighter colors indicate less importance or prevalence to the minimum level (light yellow and green color), The interpretability model also shows that the model is paying too much attention to unknown words (xxunk tag) or unnecessary words (e.g., in, for, the, it) which indicates the necessity of filtering those as part of the preprocessing step.

6.4.6 Limitations

This study have some limitation listed next. Most of the events in the Conversational CrisiNLP dataset contain tweets in several languages (Table 6.2), which generate noisy representations for the classification models. There is not evaluation of the cross-lingual aspect in the experiments, but it could have an impact on the conversational aspect. Also, the experiments do not evaluate cross-event scenarios, which is an aspect important for real-world scenarios, and some proposed approaches tackle the issue [7].

6.5 Conclusions

This chapter gains insights on the use of conversational context for the NLP classification task for crisis-related conversations on Twitter. This study introduced a new Conversational CrisisNLP dataset and evaluated several text classifiers. Empirical results show leveraging conversational context does improve the prediction scores of the classifiers. Future research directions can evaluate the sample aspect for image-grounded conversations in the crisis domain.

Chapter 7

Semisupervised Learning

This chapter addresses RQ5 that deals with the use of massive unlabeled data for classification tasks. This chapter begins with an overview and motivation for semisupervised learning in Section 7.1. Next, Section 7.2 provides a description of the methodology and the cross-lingual classification task evaluation. Then, Section 7.3 details the experimental settings in and Section 7.4 show the empirical results. Finally, Section 7.5 outlines conclusions and research directions. The work presented in this chapter was first published at the 2nd Workshop on Affective Content Analysis (AffCon 2019) co-located with Thirty-Third AAAI Conference on Artificial Intelligence (AAAI 2019) [132].

7.1 Introduction

The classification or grouping of short texts is critical in various tasks in text mining and the retrieval of information in the context of social networks or data generated by users on the web. Specifically, these tasks aim to categorize or group similar texts, so that texts with the same label or group are similar to each other and different from texts in other categories or groups. Traditional classification or grouping models often use a sparse representation for text data, such as the bag of words (BOW) or TF-IDF [84].

However, the characteristics of the short texts create some problems for both conventional unsupervised and supervised models. Usually, the number of unique words in each short text is small (90% of the instances of the text in tweets in the datasets used in this work have less than 23 words), and as a result, the problem of lexical shortage generally leads to poor grouping quality [36].

An alternative to address lexical shortages is to enrich text representations by extracting characteristics and relationships with sources such as Wikipedia [14] or ontologies [40]; however, this approach requires written knowledge, which also depends on the language. Another alternative is to code texts in distributed dense vectors [90] with neural networks [150].

Another problem is the definition of the labels for a specific task and the number of manually annotated instances for each label. Unsupervised methods learn the categories from the data, but the resulting groupings may not be related to the expected labels. Supervised methods have predefined labels but often require a considerable number of labeled instances to learn to categorize. Semi-supervised approaches offer an alternative to solve these problems by using a small amount of labeled data according to predefined classes, at the same time, take advantage of the massive unlabelled data availability [13].

This chapter investigates the research question: How can a semi-supervised approach learn to categorize short texts in a multi-label taxonomy using a small set of labeled data and leveraging the availability of large unlabeled data? To that end, the proposed build upon neural semi-supervised k-means clustering that modifies the normal objective function and adds a penalty term for labeled data [144]. The proposed model extended the neural semi-supervised clustering and applied it to multi-label settings. The results show that semi-supervised k-means outperform other baseline unsupervised models for multi-label classification tasks.

7.2 Methodology

In unsupervised learning, k-means is an algorithm for clustering data used in many applications, including text mining tasks [30]. The k-means algorithm divides the data into a K number of clusters, so that minimizes the distance of each point to the centroids of the clusters, assigning it to the nearest cluster. The input to the clustering model are the set of short texts $\{s_1, s_2, s_3, ..., s_N\}$ represented by the data points $\{x_1, x_2, x_3, ..., x_N\}$, where x_i is a sparse or dense vector.

The k-means algorithm defines a set of binary variables $r_{nk} \in \{0, 1\}$ for each data point x_n, where $k \in \{1, ..., K\}$ specifies the cluster assigned. For example, $r_{nk} = 1$ if x_n is assigned to cluster k, and $r_{nj} = 0$ for $j \neq k$. The objective function in k-means is defined as:

$$J_{unsup} = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|x_n - \mu_k\|^2 \tag{7.1}$$

where μ_k is the centroid of the k-th cluster. The k-means algorithm learns the values of $\{r_{nk}\}$ and $\{\mu_k\}$ such that optimizes J_{unsup}. To minimize the objective function, k-means utilizes the gradient descent approach with an iterative procedure [99].

Each iteration involves two steps: assign clusters and estimate centroids. In the assign clusters step, k-means minimizes J_{unsup} with respect to $\{r_{nk}\}$ by keeping fixed $\{\mu_k\}$. In this case, J_{unsup} is a linear function for $\{r_{nk}\}$, so to optimize each data point separately by merely assigning the n-th data point to the closest cluster centroid.

In the estimate centroids step, k-means minimizes J_{unsup} with respect to $\{\mu_k\}$ by keeping $\{r_{nk}\}$ fixed. In this case, J_{unsup} is quadratic function of $\{\mu_k\}$, and minimization sets to zero the derivative

for $\{\mu_k\}$, as follows:

$$\frac{\partial J_{unsup}}{\partial \mu_k} = 2 \sum_{n=1}^{N} r_{nk}(x_n - \mu_k) = 0 \tag{7.2}$$

Then, it solves $\{\mu_k\}$ as

$$\mu_k = \frac{\sum_{n=1}^{N} r_{nk} x_n}{\sum_{n=1}^{N} r_{nk}} \tag{7.3}$$

Thus, μ_k corresponds to the mean of all the data points assigned to the cluster k.

7.2.1 Neural Semi-supervised Clustering

The classical k-means algorithm uses unlabeled data to solve the clustering problem based on an unsupervised learning approach; however, the clustering results may not be consistent with the expected labels. This study extend the semi-supervised approach in [144], which injects some supervised information into the learning process to produce useful and coherent clusters. Similar to the classic k-means algorithm, the training steps for the neural semi-supervised k-means are:

1. Initialize $\{\mu_k\}$ and $f(\cdot)$.
2. Repeat until convergence:
   (a) assign clusters: Assigns each short-text to its nearest cluster centroid based on its neural representation.
   (b) estimate centroids: Estimates the clusters' centroid based on the cluster assignments from previous step.
   (c) update parameters: Updates the neural networks parameters according to the objective function by keeping fixed the centroids and cluster assignments.

7.2.2 Representation learning

The model represents each short text entry s_i as a sequence of word indices and, together with the initial centroids, form the input data to the semi-supervised neuronal clustering model. Then, the embedding layer maps each word in the sequence as a dense vector $x = f(s)$, using word embeddings initialized randomly or from pre-trained embeddings [90, 89]. In this approach, rather than training the text representation model independently, the semi-supervised clustering integrates it with the k-means algorithm training process.

7.2.3 Objective function

The neural semi-supervised clustering uses a small number of labeled instances to guide the clustering process and minimizes the objective function defined as:

$$
\begin{aligned}
J_{semi} = \sum_{c=1}^{C} \{ \alpha \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \| f(s_n) - \mu_k \|^2 \\
+ (1-\alpha) \sum_{n=1}^{L} \{ \| f(s_n) - \mu_{g_n} \|^2 + \\
\sum_{j \neq g_n} [l + \| f(s_n) - \mu_{g_n} \|^2 - \| f(s_n) - \mu_j \|^2]_+ \} \}
\end{aligned}
\tag{7.4}
$$

where $\{(s_1, y_1), (s_2, y_2), ..., (s_L, y_L)\}$ denote the labeled data, and the unlabeled data is $\{s_{L+1}, s_{L+2}, ..., s_N\}$. The label y_i specify the cluster for each short-text s_i. The outer sum iterates over the number of labels C defined in the taxonomy; thus, extending the original objective function in [144]. The objective function contains two terms:

1. The first term is the objective function in the classic k-means algorithm (Equation (7.1)), and the second term penalizes depending on how far are the predicted clusters from the ground-truth clusters for labeled data. The experiments use the factor $\alpha \in [0, 1]$ to tune the importance of unlabeled data.

2. The second term contains two sub-terms:

   (a) The first sub-term penalizes depending on the distance between each labeled instance and its correct cluster centroid, where $g_n = G(y_n)$ indicates the cluster ID given by the label y_n. The mapping function $G(\cdot)$ uses the Hungarian algorithm [96].

   (b) The second sub-term specifies a hinge loss function with a margin l, where $[x]_+ = max(x, 0)$. This term incurs in some loss if the distance to the ground truth centroid is larger (by a margin l) than the distances to the wrong centroids.

7.2.4 Model training

The parameters in J_{semi} are: the clusters' assignment for each text $\{r_{nk}\}$, the clusters' centroids $\{\mu_k\}$, and the neural network weights $f(\cdot)$. The goal is to find the values of $\{r_{nk}\}$, $\{\mu_k\}$, and parameters in $f(\cdot)$ that minimizes J_{semi}. Based on the k-means algorithm, the semi-supervised model iteratively minimizes J_{semi} with respect to $\{r_{nk}\}$, $\{\mu_k\}$, and parameters in $f(\cdot)$.

First, the model initializes the clusters' centroids $\{\mu_k\}$ with the k-means method [12], and also initializes randomly the parameters in the neural network. Then, the model iteratively carries

out three steps (assigns clusters, estimates the centroids, and updates the parameters) until J_{semi} converges.

The assign clusters step minimizes J_{semi} with respect to $\{r_{nk}\}$ by keeping fixed $f(\cdot)$ and $\{\mu_k\}$ to assign a cluster ID for each data point. The second term in Equation (7.4) has no relation with $\{r_{nk}\}$. Thus, the model only needs to minimize the first term, by setting the nearest cluster centroid to each text, i.e., is identical to the assign clusters step in the k-means algorithm. In this step, the model also calculates the mappings between the ground-truth clusters specified by $\{y_i\}$ and the cluster assignments for the labeled data.

The estimate centroids step minimizes J_{semi} with respect to $\{\mu_k\}$ by keeping $\{r_{nk}\}$ and $f(\cdot)$ fixed, which corresponds to the estimate centroids step in the classic k-means algorithm. It aims to estimate the cluster centroids $\{\mu_k\}$ based on the cluster assignments $\{r_{nk}\}$ from the assign_cluster step. In the Equation (7.4), the second term considers each labeled instance in the process of estimating cluster centroids. Solving $\partial J_{semi}/\partial \mu_k = 0$, gives

$$\mu_k = \frac{\sum_{n=1}^{N} \alpha r_{nk} f(s_n) + \sum_{n=1}^{L} w_{nk} f(s_n)}{\sum_{n=1}^{N} \alpha r_{nk} + \sum_{n=1}^{L} w_{nk}} \tag{7.5}$$

$$\begin{aligned}
&w_{nk} = (1-\alpha)(I_{nk}^{'} + \sum_{j \neq g_n} I_{nkj}^{''} - \sum_{j \neq g_n} I_{nkj}^{'''}) \\
&I_{nk}^{'} = \delta(k, g_n) \\
&I_{nkj}^{''} = \delta(k, j) \cdot \delta_{nj}^{'} \\
&I_{nkj}^{'''} = (1 - \delta(k, j)) \cdot \delta_{nj}^{'} \\
&\delta_{nj}^{'} = \delta(l + \|f(s_n) - \mu_{g_n}\|^2 - \|f(s_n) - \mu_j\|^2 > 0)
\end{aligned} \tag{7.6}$$

where $\delta(x_1, x_2)$=1 if x_1 is equal to x_2, otherwise $\delta(x_1, x_2)$=0; and $\delta(x) = 1$ if x is true, otherwise $\delta(x) = 0$. In the numerator of Equation (7.5), the first term represents the contributions from all data points, and the weight of s_n for μ_k is αr_{nk}. The second term represents labeled data, and w_{nk} is the weight of an instance s_n for μ_k.

The update parameters step minimizes J_{semi} with respect to $f(\cdot)$ by keeping $\{r_{nk}\}$ and $\{\mu_k\}$ fixed, with no counterpart in the k-means algorithm. The primary goal is to learn the parameters of the text representation model. The training uses J_{semi} as the loss function and employs the Adam algorithm to optimize it [71].

7.3 Experimental Settings

This chapter evaluates the models on the dataset for the Ecuador earthquake introduced in previous chapters. The Appendix A details the data acquisition, storage, and processing of the

dataset. For training, the experiments use a small subset of labeled data and a large subset of unlabeled dataset [67]. Table 7.1 summarizes the number of labeled and unlabeled text instances for training, as well as the number of instances in the test set. For the experiments, the splitting strategy is to randomly sample 80% of labeled instances for training (training set) and remaining 20% instances for validation (validation set). The unsupervised and semisupervised models use unlabeled instances for training (unlabeled set). The experiments train the models using k-fold cross-validation ($k = 10$) on the training set and report the results for the validation set using the metric F_1.

Dataset	Labeled	Unlabeled	Test	Total
2016-EC-eq	1,500	150,076	693	152,269

Table 7.1: Statistics for the dataset.

This subsection compares the proposed semi-supervised approach with unsupervised and supervised models.

7.3.1 Unsupervised learning:

All unsupervised models use k-means for clustering, with the number of clusters $k = 2$ to map the values for each label $(0, 1)$. For learning representation, the experiments evaluate the following methods: a) BOW represents each short-text as a sparse vector based on term frequency (TF). b) TF-IDF similar to BOW, uses a sparse vector to represent each short-text based on term frequency-inverse document frequency, and c) AVG-EMB uses word embeddings vectors to represent each short-text and then calculate the average.

7.3.2 Supervised learning:

The experiments evaluate several supervised models for the classification task; the representation learning used depends on each model as described next:

LR: uses a sparse vector representation that feeds a logistic regression classifier.

FastText: uses dense word vectors representation (embeddings layer), followed by a Global Average Pooling layer, which averages the word embeddings, and then uses a Dense layer with sigmoid activation to predict the labels.

CNN: uses dense word vectors representation (word embeddings layer) followed by a Dropout layer, then a convolutional layer, and an output layer with sigmoid activation.

LSTM: similar to CNN, but the word embeddings layer feeds a recurrent LSTM layer, which is more suitable for sequence modeling such as texts.

BiLSTM: uses two LSTM networks to model the text sequences in both directions, followed by a Dropout layer with a rate of 0.5 and then a dense layer with sigmoid activation.

CNN-LSTM: leverage the advantage of the CNN layer to capture salient features and sequence modeling capability of LSTM.

7.4 Results and Discussion

Type	Model	Precision	Recall	F1
unsupervised	BOW	62.68	54.72	58.43
unsupervised	TF-IDF	62.10	68.75	65.26
unsupervised	EMB-AVG	63.50	75.91	69.16
supervised	LR	87.02	86.80	86.91
supervised	FastText	88.28	88.91	88.60
supervised	CNN	88.56	88.92	88.74
supervised	LSTM	89.45	86.97	88.19
supervised	BiLSTM	89.05	87.32	88.18
supervised	CNN-LSTM	88.48	86.85	87.66
semi-supervised	CNN	89.03	87.10	88.06

Table 7.2: Models performance for the task of identifying crisis-related conversations.

Table 7.2 summarizes the scores of the models on the test set. The models fall into three categories (type): unsupervised, supervised, and semi-supervised. The metrics are precision, recall, and F_1, and report the scores for each label. The columns show the total weighted score of the metrics for each model. The results show that the supervised systems outperform unsupervised models by a small margin. Among the supervised learning, deep neural models perform better than the baseline method (LR). The semi-supervised model shows promising results, as it achieves scores close to the supervised models. Therefore, classification tasks can achieve good predictions results with less annotation effort.

Neural architectures introduce several hyper-parameters like the output dimension of the text representation models, while semi-supervised k-means clustering has α in Equation (7.4). The next sections analyze the impact of some of the hyper-parameters and determine the configuration for further experiments.

7.4.1 Embeddings dimension

The experiments evaluate the effectiveness of the output dimension in text representation models. To that end, the experiments use embeddings size of {50, 100, 200, 300, 500, 1000}, while maintaining all other parameters fixed. Figure 7.1 show that the score F_1 drops if the size is ≤ 100 and the curve

falls if the size is $\geq$ 500. Based on the results, further experiments use 300 as the size of the embedding.

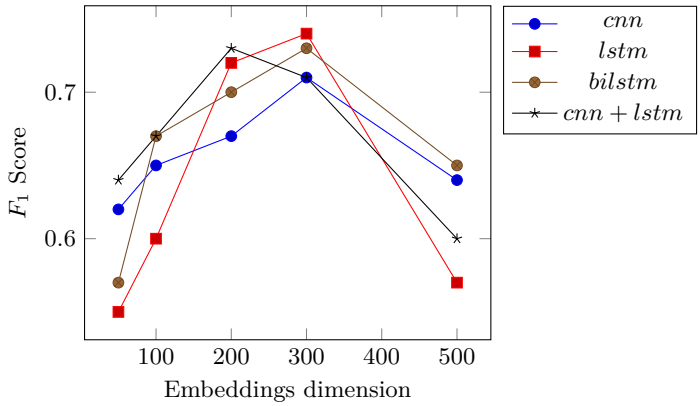


Figure 7.1: Influence of the dimensionality of the text learning representation.

7.4.2 Alpha

The experiments evaluate the effect of α in Equation (7.4), which controls the importance of unlabeled data in the performance of the model. The experiments evaluate α with values of: {0.00001, 0.0001, 0.001, 0.01, 0.1}, and maintain the other parameters fixed. Figure 7.2 shows that the performance decay for small α values. By increasing the value of α, there are improvements and reach a peak F_1 score with $\alpha = 0.1$. Further experiments use the value of $\alpha = 0.1$ as it maximizes F_1.

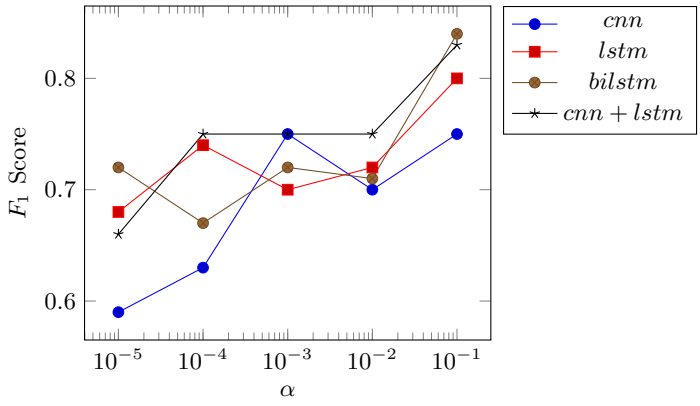


Figure 7.2: Influence of unlabeled data, where the x-axis is α in Equation (7.4).

7.4.3 Labeled set size

This parameter controls the influence of the size of the labeled data. The experiments evaluate the ratio of labeled data for training between $[1\%, 10\%]$, and kept the other parameters fixed. Figure 7.3 illustrates the performance improvement as the size of labeled data increases and confirms the importance of labeled data for training.

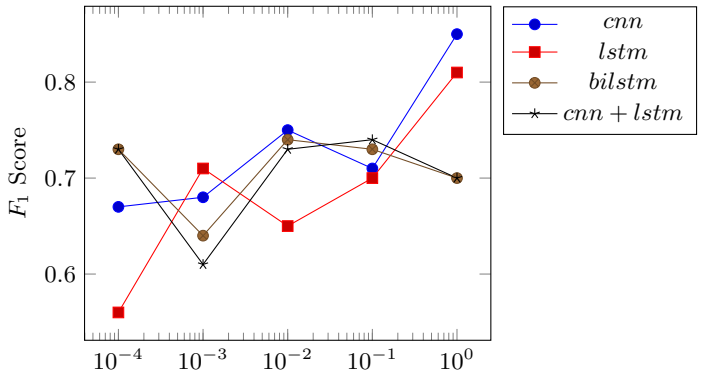


Figure 7.3: Influence of the size of labeled data used for training.

7.4.4 Pre-training

This aspect measures the effect of the pre-training embeddings for neural architectures. The first option uses pre-trained embeddings in the models for the classification task with labeled data. These experiments evaluate several pre-trained embeddings such as Word2Vec, Glove, FastText. Figure 7.4 shows that pre-trained embeddings achieve superior performance compared to random embeddings; further experiments use FastText.

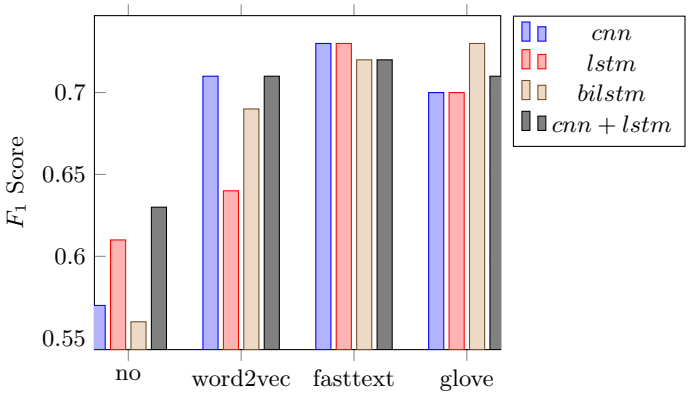


Figure 7.4: Influence using pre-training embeddings in neural models.

7.5 Conclusions

The work presented in this chapter builds on the neural semi-supervised clustering that integrates a neural representation learning for short-texts, and the k-means clustering into a unified framework. To that end, the model utilizes a small percentage of labeled data to guide the intention for clustering. The proposed extended the model to use it in the muti-labeled clustering of short-texts. The results show that the proposed neural semi-supervised clustering is more effective than baselines unsupervised, and it close to the supervised models. Therefore, the results show the potential to overcome critical issues, such as scarcity of labeled data, and leverage the availability of massive unlabeled data.

Chapter 8

Recommending Conversations

This chapter addresses RQ6 related to the recommendation of content on social media, in particular, recommending conversations. This chapter begins with an overview and motivation for recommending conversations in Section 8.1. Next, Section 8.2 provides a description of the methodology and the recommendation task evaluation. Then, Section 8.3 details the experimental settings in and Section 8.4 show the empirical results. Finally, Section 8.5 outlines conclusions and research directions.

8.1 Introduction

Online social networks (OSNs) are increasingly receiving considerable attention from the research community, as the content available in these networks contains valuable information for event detection, crisis management, forecasting, recommendations, among others. The near real-time user interactions within OSNs, make new relevant content (e.g., news) available to the public often faster than traditional media [73, 45]. Thus, the content recommendation is an essential task for web companies and organizations that are looking to reach an audience, such as advertising [3, 79].

The work presented in this chapter focuses on recommending content on OSNs as part of intelligent systems in the domain of crisis management [106]. Due to the increasing number of crisis events worldwide, it is an important task to recommend content that may elicit interactions from the crowds to help other people, especially in the context of natural disasters. Nevertheless, the proposed approach applies to other application domains. Previous works focused on analyzing individual tweets for recommendation tasks. In the field of recommendation systems, previous works addressed several types of recommendation tasks, including hashtags, mentions, news, points-of-interest, profile classification, retweets, tweets, URLs, and whom to follow [131].

Twitter conversations may span a wide range of domains such as politics, religion, health, fitness, sports, food, or fashion. Conversational recommendations could be particularly useful during crisis

events, such as earthquakes when people use social media platforms to look for relevant information, request for help, offer support, contact relatives, and other types of inquiries. This chapter presents a novel recommendation task for OSNs, particularly recommending users to join relevant multi-agent and multi-turn conversations. This task aims to recommend users to join a set of conversations on different topics they might be interested, even when no link connections are available (i.e., users have no relationship between them). Moreover, recommending short-text conversations can be modeled using neural architectures such as sequential models [50]. This chapter proposes a neural learning architecture based on a sequence-to-sequence model to tackle the task of recommending conversations using the conversation context and the user's history.

In conversational modeling, it is feasible to train supervised machine learning models without human-annotated data, since the data itself provides the ground truth as in the case of the prediction of the next utterance [83]. The experiments use a large Spanish Twitter corpus of Ecuadorian Users (USERSEC) containing 10K conversations from three popular users from Ecuador. The dataset contains posts whose content is related to three domains: politics, sports, and crisis events activism.

Empirical results show that the model outperforms several baseline models, such as TF-IDF and state of the art model based on collaborative filtering (CFT). The experiments highlight the performance of the bi-LSTM variant of the model, which outperforms other variants such as RNN or LSTM.

The main contributions of this chapter are: (i) the USERSEC corpus, a conversational dataset of short-text conversations on Spanish, (ii) a novel task for recommending users in short-text conversations, and (iii) a state-of-the-art neural architecture to recommend users or conversations on microblogs.

8.2 Methodology

8.2.1 Data

This chapter analyzes conversations on Twitter, which is one of the most popular microblogging sites. In some circumstances such as crisis events, conversations on Twitter occur almost in real-time as the case of IRC or chat rooms. An advantage of using Twitter is the limit in the number of characters of each message, which makes it comparable to conversations found in chat rooms. Although in November of 2017, Twitter announced that it had increased the limit to 280 characters[1], most of the tweets collected in the corpus have the prior limit of 140 characters.

Twitter conversations begin with a user posting a message, not directed to any user in particular but broadcast to the user's audience (followers). Some users in the network interact by replying to the initial tweet or another tweet in the conversation thread. Since the Twitter APIs provide an

[1] https://blog.twitter.com/official/en_us/topics/product/2017/tweetingmadeeasier.html

identifier to each tweet, it is not necessary a preprocessing step to disentangle utterances (replies) in conversations as it is the case for other data sources [142].

Figure 8.1 illustrates the types of conversations occurring on Twitter. The conversation begins with a root tweet (initial tweet) from a user (blue node). Wide conversations have just a single reply to the root tweet (orange nodes). Some of the wide conversations evolve into Deep conversations that can contain several replies in a cascade forming conversational threads with three or more tweets (green nodes). In this context, non-followers might be interested in joining the conversation; thus, finding the right profiles for a set of conversations highlights the scope and relevance of the tasks proposed in this chapter.

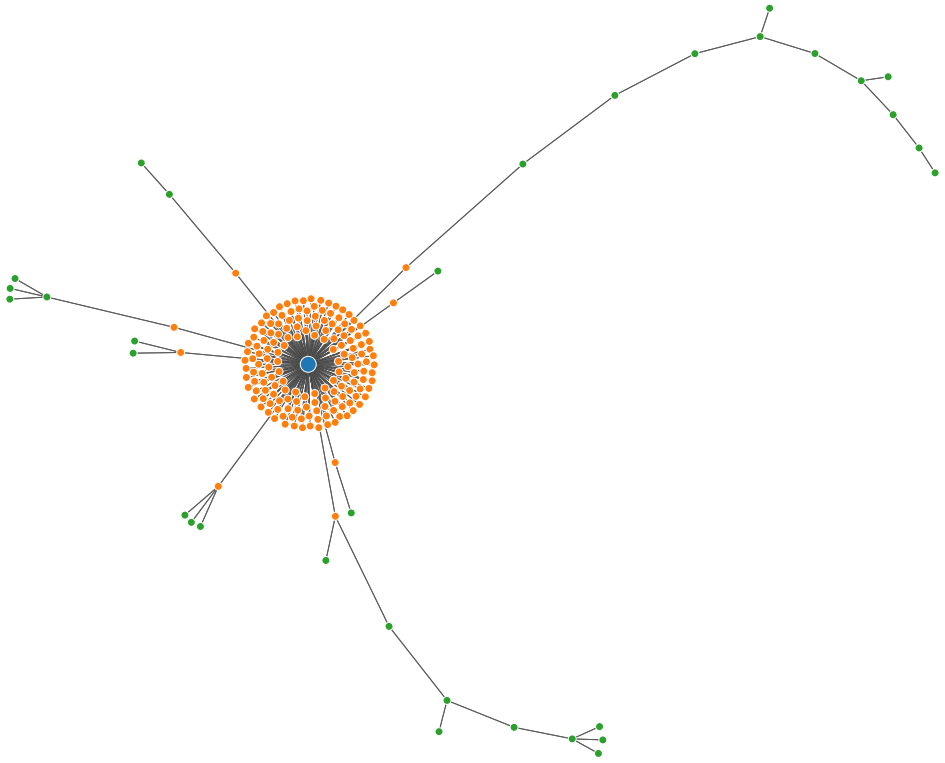


Figure 8.1: A sample conversation from the USERSEC corpus. The large blue node at the center is the initial tweet of the conversation. The orange nodes around are direct replies to it (wide conversations), while some of the tweets (green nodes) evolve into deep conversational threads.

The work presented in this chapter introduces the USERSEC conversational corpus that includes the following properties: • Large number of conversations, on the scale of 10^4, appropriate for training neural architectures. • Multi-turn conversations with more than two turns. • Multi-agent conversations as opposed to two-way conversations in chats datasets. • Conversations on several domains. • Conversations on the Spanish language.

Unlike the work of [115], the dataset does not restrict the conversations to the initial two or three utterances. Without the limitation in the number of utterances, one can train models to learn context-aware conversations, which is essential for the recommendation of relevant conversational threads. The dataset contains all tweets in conversations initiated by three popular users in Ecuador. These users belong to different expertise domains: politics (@MashiRafael), sports (@aguschmer), and humanitarian aid activism (@KarlaMorales). By using different domains, the analysis shows that there are users with mutually exclusive interactions as well as inter-domain interactions (green nodes), as shown in Figure 8.2. In the dataset collected, there are at least 10% of users interact with two or more in different domains.

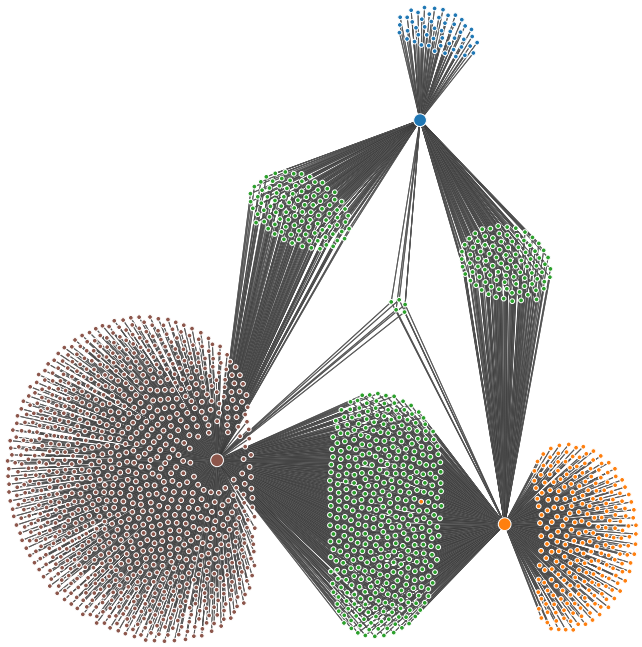


Figure 8.2: Clusters of interacting users. Three popular users (MashiRafael, KarlaMorales, aguschmer) in the USERSEC dataset and a sample of their audience forming clusters (in color Red, Blue, Orange) of users interacting exclusively with them. The users interacting with several popular users constitute the intersecting clusters (green).

The data collection uses Twitter's Streaming API[2] to collect the initial tweets of fashionable users. The tweets collected contain all the replies to the tweets of the popular users, as well as the

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

timelines of users participating in the conversations collected with the Twitter REST API[3]. The reply field (in-reply-to-status-id) in tweets' metadata specifies when a tweet is a reply by another user and allows to collect all parent tweets in the conversation threads through recursive navigation to the root tweet. Then, the data collection gets the tweets in all possible conversation threads, navigating recursively down to the child tweets. The timelines dataset contains the last two hundred tweets posted by all users participating in the conversations.

The grouping of the tweets per conversation uses the root tweet ID of each conversation and include only replies within a time frame of 10 days to avoid noisy tweets generated by bots or spam accounts since 90% of the conversations on Twitter last less than ten days [133]. Also, the preprocessing discard replies posted in a short period (less than five seconds), given that those are usually tweets created by third-party applications, bots, or spam accounts. To allow the models to learn the context of the conversations, the preprocessing filter out conversations with less than two utterances. In the case of consecutive utterances, if they belong to the same user, the preprocessing concatenates them in a single turn. Finally, the preprocessing discard tweets in languages other than Spanish using the language identifier in the tweet meta-data.

8.2.2 Characterization

The dataset contains tweets created from March 2015 to December 2017. Approximately 35% of the tweets are conversational, which is similar to the values obtained in prior work [115]. Table 8.1 summarizes the most important characteristics of the collected corpus. There are two datasets, the conversational tweets, and the tweets that belong to the timeline of the users participating in the conversations (the timelines dataset).

The conversations dataset contains approximately 520K tweets for 13K conversations started by three popular users selected as the seed users. To build the conversation threads, the augmentation task uses the methodology described in [133] to increase the number of conversations given that each conversation can have multiple ramifications (threads). Additionally, the same conversation thread can generate several overlapping contexts to increase the number of instances in the training set. The number of turns is less than the number of utterances because of the grouping of consecutive utterances that belong to the same user in a conversation thread. The average length of the conversations is approximately five turns, constrained to a minimum of two and involving two or more users. An important statistic is the number of words for the contexts of the conversation threads, as it defines the maximum length of the input layer in the model, as explained in subsection 8.2.4.

The timelines dataset contains the last 200 tweets crawled for each of the users participating in the conversations. The number of unique users engaging in conversations with popular users is 120K approximately. For those users, 71K (60%) have a description in their profile, although the experiments do not use that information given that it is very noisy. The distribution of the number

[3] `https://dev.twitter.com/rest/public`

Conversations		Timelines	
popular users	3	users	120,220
conversations	12,909	descriptions	71,142
tweets	520,263	tweets	14,860,508
words	588,039	words	13,607,961
conversation threads	65,274		
utterances	345,277		
turns	331,971		
minimum turns	3.00	minimum tweets	3.00
average turns	5.09	average tweets	123.61
maximum turns	64.00	maximum tweets	200.00
median turns	4.00	median tweet	74.00
minimum context words	5.00	minimum words	1.00
average context words	68.23	average words	1,571.77
maximum context words	1,720.00	maximum words	652,282.00
median context words	57.00	median words	834.00

Table 8.1: USERSEC dataset statistics

of replies in conversations follows a power law ($alpha = 1.4368, xmin = 2$), as shown in Figure 8.3. Most conversations are short (i.e., few replies), but the dataset contains deep conversations with more than ten tweets mostly initiated by the three popular users.

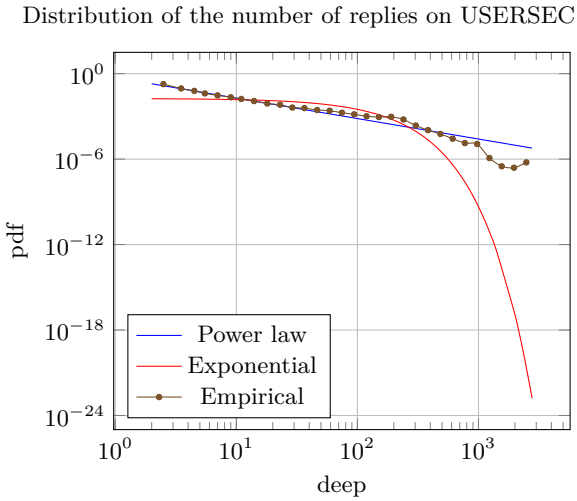


Figure 8.3: Distribution of the number of replies in conversations for the USERSEC dataset. The empirical distribution (dotted brow line) follows a power-law distribution (blue) more closely than an exponential distribution (red).

8.2.3 Problem Definition

This study frames the problem of recommending short-text conversations on microblogs as a ranking task. Give a set of users $U = \{u_1, u_2, \ldots, u_n\}$ and a set of conversations $C = \{c_1, c_2, \ldots, c_n\}$, it defines two subproblems:

1. Recommend a user U_i to join a conversation C_{target} from a subset C_{sample}, where $C_{sample} \subset C$. In this case, the task consists of predicting which conversation a user will join from a list of candidate conversations, which includes the ground truth and several distractors conversations.

2. Recommend a conversation C_i to a user U_{target} from set U_{sample}, where $U_{sample} \subset U$. In this case, the task consists of predicting the user that will join a conversation from a list of candidate users, which includes the ground truth user and several distractors users.

Each conversation context C_i is the concatenation of a percentage of the tweets in a conversation thread $Ct = \{ct_1, ct_2, \ldots, ct_n\}$. While each user profile U_i consists of the concatenation of the set of tweets $Ut = \{ut_1, ut_2, \ldots, ut_n\}$ that belong to Ui prior to the conversation C_i.

The two subproblems generalize as a more generic content-to-content recommendation task: recommending a source item S_i to a target item T_{target} from a subset T_{sample}, where $T_{sample} \subset T$. In this case, S can be either the user profiles U or the conversation contexts C and similarly for T.

8.2.4 Proposed Approach

Neural architectures rely on large datasets for training, which highlights the value of using a sizeable conversational corpus, as previous works [116, 83, 125]. The experiments leverage the large-scale data collected to train and evaluate a deep neural model based on a sequence-to-sequence architecture. The model presented leverage sequence-to-sequence models that use an encoder-decoder approach for the problem of dialog response selection [83]. This chapter focus on recommending users according to the likelihood that they might join a conversation or vice-versa, as defined in Section 8.2.3. Figure 8.4 depicts the proposed model based on a sequence-to-sequence architecture for the case of recommending users to join a conversation. The model learns to rank source users U_i that will join a target conversation C_{target} based on the prediction score, in this case $U_i = S$ and $C_{target} = T$.

For the activation units in the recurrent neural networks, the experiments use several variants as detailed next. RNNs extends traditional neural networks to allow for time-delayed directed cycles between units [86]. This characteristic leads to the formation of an internal state of the network, h_t, which allows it to model time-dependent data. The network updates the internal state at each time step as some function of the observed variables x_t, and the hidden state at the previous time step h_{t-1}. The matrices W_x and W_h represent the input and hidden state, respectively, as shown in

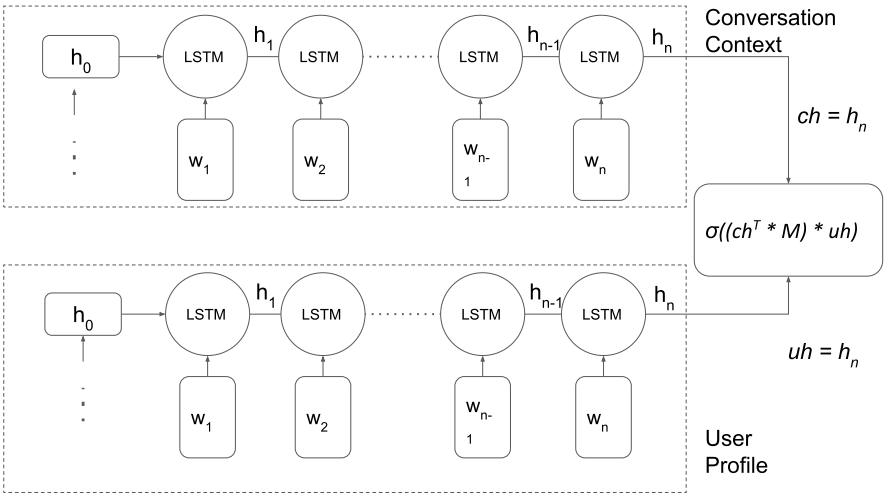


Figure 8.4: Diagram of the Seq2Seq model, where w_i are the word vectors of a conversation context C_i and a user profile U_i. The values ch, ph are the last hidden states from the sequence models and correspond to the source S and target T in the generic task.

Equation 8.1.

$$h_t = f(W_h h_{t-1} + W_x x_t) \tag{8.1}$$

RNNs are the building block of many current neural language models, including those on encoder-decoder architectures [116, 125]. In the encoder-decoder approach, the first RNN is used to encode the given sequence. The initial hidden state in the second RNN uses the final hidden state from the first RNN to generate a candidate's sequences optimized by beam-search or similar search strategies. The model consisting of two siamese RNNs to model the conversation context and the user profiless, respectively. The input layer receives the text of these two sets of tweets. Then, the embedding layer maps the sequences of words to a lower-dimensional representation ($w \in \mathbb{R}^d$). For the initialization of the word embeddings, the experiments use the pre-trained vectors [89], which consist of 2 million-word vectors trained on Common Crawl (600B tokens). The input to the RNN networks is the output of the embedding layer. At each step, the siamese RNNs update the hidden states and obtain two final hidden states that represent a summary of the conversation contexts and the user profiles. Then, the model learns the probability of matching the correct pair of final hidden states from both RNNs, as shown in Equation 8.2.

$$p(\text{flag} = 1|ch, uh, M) = \sigma(ch^T * M * uh + b), \tag{8.2}$$

where the bias b and the matrix $M \in \mathbb{R}^{d \times d}$ are parameters of the model. Given an input profile, the model generates a context with the product $ch' = M * uh$ and measures the similarity to the

actual context using the dot product. Then, the sigmoid function converts it to a probability. The training of the model tries to minimize the cross-entropy [155] of all triples (context, profile, flag), as shown in Equation 8.3.

$$\mathcal{L} = -\sum_{n} \log p(\text{flag}_n | ch_n, uh_n, M) + \frac{\lambda}{2}||\theta||_F^2, \tag{8.3}$$

where $||\theta||_F^2$ is the Frobenius norm of $\theta = \{M, b\}$. The initialization of the weights and the optimizer are critical for the performance of the RNNs networks. The experiments initialize the W_h matrix using orthogonal weights [121] and W_x with values from uniform distribution in the range of $[-0.01, 0.01]$. The RNN model uses one hidden layer with 256 neurons and updates the gradients using the Adam optimizer [71].

The first variant is to use LSTM units in the same architecture shown in Figure 8.2. RNN networks suffer from vanishing gradients when dealing with long sequences. LSTM models tackle the problem of vanishing gradients and allow the model to learn longer dependencies in sequential data such as text [50]. To handle longer dependencies, LSTM uses a series of gates that determine whether the network should either, remember a new input, forget (and retain the old value), or utilize as output. The error signal can be feedback indefinitely into the gates of the LSTM unit. This error signal helps to overcome the vanishing and exploding gradient problems in standard RNNs, where the error gradients would otherwise decrease or increase at an exponential rate.

The second variant is to use bi-directional LSTM (bi-LSTM) networks [122], which consists of two LSTM networks that read the input sequence from left to right and vice-versa. This architecture allows us to capture a more accurate meaning of the sequences and relationships of the words in both conversation contexts and user profiless. Bi-LSTM networks provide two final hidden states, and the model average them to get the vectors p and c for Equation (8.2).

8.3 Experimental Settings

8.3.1 Preprocessing

The preprocessing applies several standard preprocessing steps to the data, as described next. The tokenization uses the NLTK[4] library to split each tweet into words. Then, further preprocessing uses the library Preprocessor[5] to replace with generic tags any specific tokens present in the tweet's text like mentions, URLs, and hashtags, emojis and reserved words (RT, FAV). To reduce the computational cost of the experiments, the preprocessing truncates the $70th$ percentile (120 words) as the maximum text length for the conversation contexts and the user profiles. Also, the preprocessing excludes the timeline data of follower users as well as images or other multimedia data associated

[4] www.nltk.org/
[5] https://github.com/s/preprocessor

with the tweets given that the focus of this chapter is to analyze text-based conversations.

8.3.2 Splitting

For the main experiments, the data splitting uses the same methodology as in the work of [159]. For each conversation, the experiments use 80% for training, 10% for validation, and test sets, respectively. The creation of the training, as well as evaluation sets, are explained next. For the task of recommending users, the preprocessing build the training set by extracting triplets (context, user, flag) from conversations in the corpus. The experiments consider each utterance after the second, as a potential candidate specifying the user that joined the conversation, and the previous utterances as the conversation context. The context contains a sequence of utterances occurring in a conversation thread, truncated to a specific number of tweets. The user profile consists of the tweets in the user's timeline. The flag indicates whether or not the user joined the conversation, i.e., whether the user is the one who replied to the context.

Each conversation context forms a pair of triples. The first triplet contains the user profile that joined the conversation ($flag = 1$). In the second triplet, the users not joining the conversation ($flag = 0$) are selected randomly from elsewhere in the training data. The experiments use a $1 : 1$ ratio between positive ($flag = 1$), and negative ($flag = 0$) triples. Table 8.2 shows an example[6] in the training set (translated from Spanish to English language). The example shows the context of the conversation and two user profiles: the user profile with Flag set to 1 did join the conversation, whereas the other profile did not.

Conversation context	User Profiles	Flag
@KarlaMoralesR Any courier who supports us? There are more than 50 life straws in Atlanta, DHL charges regular fee for shipping, another way for them to arrive? Any courier that supports us?	@iChacha_Chacha Nutritionist, Dietitian and Esthetician. Now crossfit lover [...]	1
	@Sandinin lawyer, humanist, I still believe in the human being [...]	0

Table 8.2: Example of an instance in the train set.

To building the user profile, the experiments consider the tweets before the conversation in the dataset. Another strategy for building the profiles is to use a specific number of tweets posted by each user. However, due to the higher computational cost for training the models, this can be an issue for future research. In the validation and test sets, the distribution of the negative instances increases the difficulty of the recommendation task compared to the balanced distribution in the training set. The augmentation of negative instances resembles real scenarios where a large number of users with different preferences. In the dataset, for each positive instance (the correct user profile that joined the conversation), nine user profiles did not join the conversation as negative instances

[6] https://twitter.com/KarlaMoralesR/status/723235984662511616

or distractors. For evaluation purposes, the experimental settings use 20% of the conversations in the dataset to build the validation and test sets. For the task of recommending conversations, the experimental settings are the same to build both the training and evaluation sets.

8.3.3 Additional datasets

In addition to the corpus introduced, this chapter evaluates the models using the TREC dataset [159]. The experiments use the TREC dataset to perform benchmarking comparison of the models as detailed in Section 8.4. The TREC dataset contains approximately 10K conversations of diverse topics based on the tweets released by TREC 2011 microblog track[7].

8.3.4 Baseline methods

The experiments evaluate several two baseline methods (Random and naive TF-IDF) as well as the state of the art CTF method[8].

Random: This method ranks the items (users or conversations) for recommendation randomly and establish the quality of other models.

TF-IDF: The 'term-frequency' component (TF) is a count of the number of occurrences of a word in a given context, whereas the 'inverse document frequency' term (IDF) penalizes the occurrence of the word elsewhere in the corpus. The product of TF and IDF produces the TF-IDF score, as shown in subsection 8.3.4.

$$\text{TFIDF}(w, d, D) = f(w, d) \times \log \frac{N}{|\{d \in D : w \in d\}|},$$

where $f(w, d)$ denotes the number of times the word w appeared in document d, which represents the conversation context c or the user profile p. N defines the total number of documents, and the denominator represents the number of documents in which the word w appears. The term D represents the set of documents, i.e., the set of conversation contextss or user profiless. For each of the candidate users, the TF-IDF model calculates vectors for the conversation context and the user profile. The model considers the user profile with the highest dot product between the context vector and the profiles vectors (set of users candidates), as the most probable user that will join the conversation. For the proposed task, TF-IDF captures the importance of the words in the user profile for the conversation context.

[7] `https://trec.nist.gov/data/tweets/`
[8] To the best of the knowledge, the state-of-the-art before the method.

CFT: CFT method combines collaborative filtering with topic modeling for improving the recommendations [159]. The method optimizes objective function:

$$\min \mathcal{L} + \mu.NLL((C)|\Theta)$$

where $\mathcal{L}$ encodes the user reply preference similar to collaborative filtering (CF). The second term denotes the negative log-likelihood of a set of conversations C, with Θ containing parameters that capture the topical content and discourse structures by latent variables.

8.3.5 Evaluation metrics

The evaluation of the models does not require human annotation because the ground truth instances are available in the conversations itself, as described in subsection 8.2.1. The recommendation task uses ranking evaluation metrics [119]. Specifically, $Recall@k$ metrics to measure the relevance of the users recommended participating in the conversations.

For the recommendation tasks, the conversational recommender models learn to predict the ground truth instance from a set of candidate instances. The recommender models rank the users according to the prediction accuracy and evaluate if the instance predicted is among the top k candidates. That means the prediction is correct if the ranking of the ground truth instance is in the top K ranked instances specified by the K parameter of the $Recall@K$ metric. The value of the parameter k depends on the number of instances available for selection.

The evaluation uses nine negative instances (distractors) in addition to the ground truth instance. Therefore, the evaluation uses three recall metrics: $recall@(1, 10)$, $recall@(2, 10)$, and $recall@(5, 10)$. The number of candidate instances for each evaluation is 10, as specified by the second parameter.

Additionally, the evaluation includes two additional metrics ($Precision@1$ and $nDCG@5$) to compare with prior state-of-the-art method CFT [159].

8.3.6 Implementation details

The model implementation is on Tensorflow and uses a single worker instance with a single NVIDIA Tesla K80 GPU. The dataset and the code are available in Github[9] except for CFT requested to the original authors.

8.4 Results and Discussion

Table 8.3 shows the results for the baselines and the proposed model for the recommendation tasks. The RND column shows the accuracy of the random method and provides a reference for

[9] https://github.com/johnnytorres/recsys_twconv_s2s

the performance of other models. The results show that the Bi-LSTM outperforms other models on most of the evaluation metrics.

It is interesting to note that TF-IDF outperforms the CFT and RNN variant on Recall@2 and Recall@5. This result is due to the limited ability of the RNN to handle extended contexts but also depicts TF-IDF as a stable baseline for recall metrics. In the case of metric *precision*@1, TF-IDF fails considerably due to the metric's strict calculation for positive instances.

The LSTM variants of the model addressed the issues in the naive variant (RNN) and outperformed both TF-IDF and CFT.

Dataset	Metric	RND	TFIDF	CFT	RNN	S2S LSTM	BILSTM
TREC	precision@1	10.15	0.25	65.41	35.11	81.48	84.84
	nDCG@5	30.11	66.67	70.53	67.52	95.94	96.11
	recall@(1,10)	9.50	43.35	73.59	48.40	93.05	93.85
	recall@(2,10)	19.85	62.60	76.63	62.30	95.25	95.70
	recall@(5,10)	51.90	87.45	85.81	85.35	98.55	98.75
USERSEC	precision@1	10.07	-	43.28	16.43	49.51	70.74
	nDCG@5	29.90	52.95	71.06	47.03	78.32	81.89
	recall@(1,10)	11.20	36.15	60.76	25.47	65.38	78.03
	recall@(2,10)	21.11	48.18	66.41	39.32	75.64	81.11
	recall@(5,10)	49.23	68.55	81.85	67.95	89.91	85.38

Table 8.3: Results for the task of recommending conversations.

Table 8.4 shows the results of the models for the task of recommending users, which resulted in a more difficult task according to the scores. Again, the bi-LSTM variant of the model outperforms the other methods considerably. The CFT model does not support this type of recommendation and requires extensive modification in their implementation to support it.

Dataset	Metric	RND	TFIDF	CFT	RNN	S2S LSTM	BILSTM
TREC	precision@1	9.92	0.25	-	31.02	32.80	36.89
	nDCG@5	27.99	47.48	-	58.97	73.77	78.28
	recall@(1,10)	8.50	25.15	-	41.70	54.02	60.01
	recall@(2,10)	18.85	40.50	-	53.85	71.25	77.50
	recall@(5,10)	48.35	68.85	-	75.45	90.55	93.35
USERSEC	precision@1	10.01	-	-	47.80	56.58	63.02
	nDCG@5	29.88	61.66	-	76.82	82.32	83.12
	recall@(1,10)	11.62	42.39	-	65.56	75.13	76.32
	recall@(2,10)	21.02	57.35	-	73.93	79.91	81.54
	recall@(5,10)	49.06	79.31	-	87.35	89.40	89.49

Table 8.4: Results for the task of recommending users.

8.4.1 Embeddings

The experiments further evaluate the S2S model with two cases for the embedding layer. The first case is maintaining fixed the pre-trained fast text embeddings [20]. Figure 8.5 shows that the evaluation loss has a spike after $50k$ iterations, and this affects the metrics negatively, as shown in Figure 8.6. The second case allows the model to fine-tune the embeddings during training, and the loss is stable after $50k$ iterations. The pre-trained embeddings use input data from a different domain, thus fine-tuning the embeddings with different domains does improve the results.

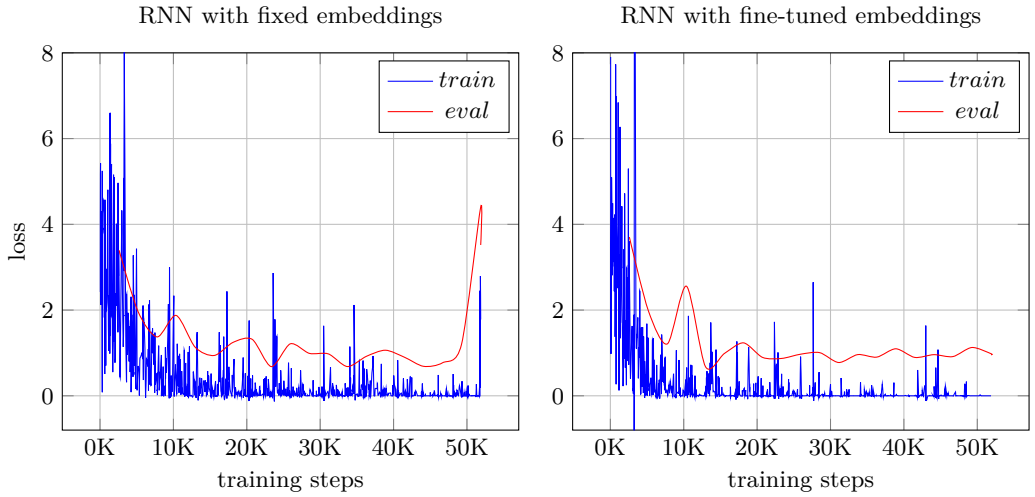


Figure 8.5: Training and evaluation set loss values for TREC dataset using fixed pre-trained embeddings (left) and fine-tuned pre-trained embeddings (right) using S2S model with RNN activation units.

Figure 8.7 shows the influence of the performance of the model as the number of training instances increase. The experiments found that using less than 60% of training instances, it causes the scores to drop significantly in all metrics for both recommendation tasks. The recall@1 metric shows that more training instances help the performance of the model. In real-world settings, the percentage of training size will depend on the availability of the tweets received or collected.

8.4.2 Discussion

8.4.3 Performance across domains

The experiments analyze the performance of the recommendation model across domains. Figure 8.8 shows the scores for the evaluation metric for each domain in the dataset: politics (@MashiRafael), sports (@aguschmer), and humanitarian activism (@KarlaMoralesR). For the task of recommending conversations, the results show that the model performs consistently better in the sport domain, mainly due to lower topics drift [3] compared to politics or sports domains. Humanitarian activism

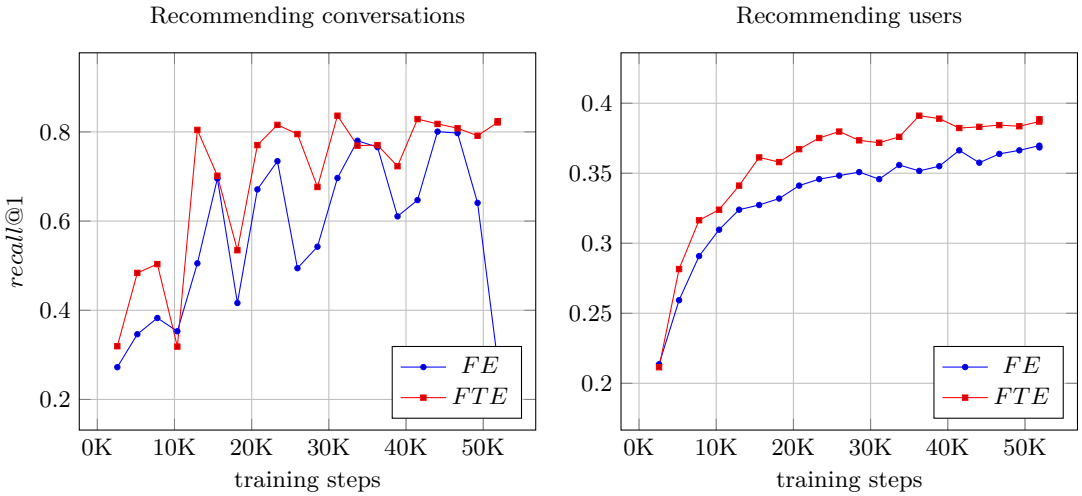


Figure 8.6: *Recall*@1 validation scores for recommending conversations (left) and users (right) in TREC dataset using fixed pre-trained embeddings (FE) and fine-tuned pre-trained embeddings (FTE).

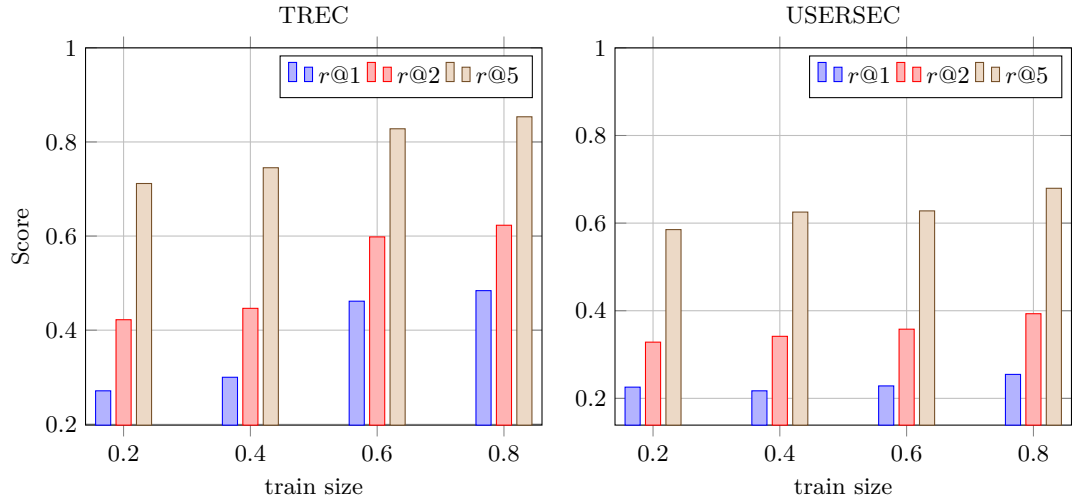


Figure 8.7: Influence of the training size in the recommendations of conversations and users. To the left, the *recall*@*k* scores for the TREC dataset, to the right for the USERSEC dataset. The x-axis represents the percentage of the conversations tweets used as training set.

is the second domain with higher scores for recommendations, while politics is the most challenging domain as well as the users that interact in several domains.

For the task of recommending users, the analysis shows the tweets related to humanitarian activism, often belong to standard categories like warnings, reports of people or infrastructure affected, relief or help requests, donations offer, emotional messages. Therefore, the predictability

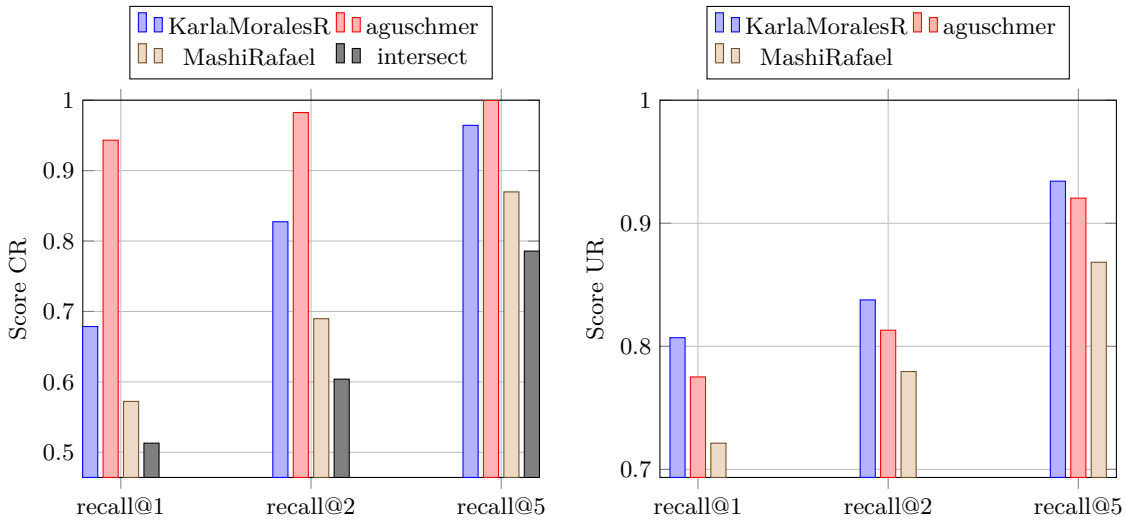


Figure 8.8: Performance of the recommendation tasks on across domains represented by the popular users. In the top, the scores for conversations recommendation (CR) task includes users that interact cross-domains (intersect). In the bottom, the scores for users recommendation (UR) task.

of the conversations is higher as the topics revolve around a small number of categories. In contrast, the domains of politics and sports change the topics rapidly in conversations according to the events unfolding in the temporal and spatial context. Thus, conversations in these domains are more difficult for the recommendation task. The results indicate that the model can perform well in specific domains and that collecting a large dataset for a specific domain for training the model could potentially further improve the performance of the recommendations.

8.4.4 Twitter conversations

Twitter allows the generation of hierarchical conversations similar to a tree structure. It means that multiple users can reply to the initial tweet, forming different branches. Each branch becomes a separate conversation thread, and deeper levels of the conversation can also have several branches. Hierarchical conversations cause the root and branching tweets to appear in multiple conversations. However, the number of such instances is small compared to the number of conversations in the dataset. Retweets can be a significant part of the streaming data, but only propagate content created by other users. For the case of retweets, the data collection includes only the original tweet (i.e., the original tweet in the propagation cascade). Another issue of conversations on Twitter is that users can include multimedia data in the tweets, such as photos and videos.

8.4.5 Task difficulty

The Recall@k metrics allow altering the difficulty of the recommendation task in a controlled manner. The difficulty of the task occurs by increasing the number of negative user profiles in the experiments, and the parameter k establishes the number of profiles the recommender system predicts as containing the correct user who joined the conversation. Future experimental settings could consider selecting negative instances that are similar to the ground truth profile, thus increasing the complexity of the task instead of choosing negative instances randomly or in different domains. Several metrics allow measuring the similarity of the user profiles based on the content found in their respective timelines. Recommender models that perform well on the more complex task should be able to capture a more fine-grained and deeper semantic meaning of utterances. This situation is evident when contrasted with models that naively select user profiles, maximizing the number of words in common with the context such as TF-IDF.

8.4.6 Model limitations

Although LSTM architecture helps to overcome the problem of vanishing gradients in RNNs and learn longer sequences [128], there is a disadvantage with recurrent neural networks. LSTM is limited to text sequences with length in the order of hundreds of words, not thousands compared to other methods that rely on topics modeling. The limitation in LSTM networks affects how much history the experiments can use to build user profiles. Ideally, the experiments should be able to feed the model with the entire timeline of each user to build his profile, which can be thousands of words long. However, LSTM fails to learn such long sequences, and it will take a long time to train, even with GPUs. Alternatives to solve the issue of very long text-content in user profiles or conversations context could use different approaches to summarize the entire user profile and obtain the latent topics of interest through TF-IDF, LDA [55], or extractive summarization [65]. Another research direction is to explore the compression of the content to a fixed hidden state using CNN. Then, the RNN model uses this fixed representation for modeling the recommendation task.

8.4.7 Model complexity

The proposed model uses the gradient descent algorithm to find the minimum of the loss function. The algorithm minimizes the error term by changing each weight in proportion to the derivative of the error for itself, given that non-linear activation functions are differentiable. The method applies back-propagation through time or BPTT and is a generalization of back-propagation for feed-forward networks [118]. For recursively computing the partial derivatives, BPTT has a time complexity of $O(n)$, where n is the number of weights per time step for computing the Jacobian matrices, at the cost of storing all forward activations within the given time horizon. For the model, the number of time steps is given by the length of the text content (conversation context or user

profile), as mentioned in Section 8.3.1. Nonetheless, the proposed method scales better compared to CF-based algorithms, such as CFT. For example, with millions of users, $O(M)$, and millions of items (conversations) $O(N)$, a CF algorithm with the complexity of n will suffer scalability problems [44].

8.5 Conclusions

The work presented in this chapter proposes a new recommendation task. It introduces a new conversational corpus on Spanish collected from Twitter, involving popular Ecuadorian users in the domains of politics (@MashiRafael), sports (@aguschmer), and humanitarian activism (@KarlaMorales). The model proposed employs Seq2Seq neural learning architecture for recommending unstructured multi-turn conversations to users that might be interested in participating. Our model is easily adaptable to a more generic task of recommending text-to-text settings, with the source and target content being interchangeable between conversation context and user profile. The experimental results demonstrate that the proposed neural architecture using variants such as LSTM and Bi-LSTM provide better results compared to baseline models as well as state of the art CFT. Moreover, the proposed model can generalize and improve the prediction performance by using pre-training embeddings compared to the baseline methods. However, there is a critical issue encoding large sequences of text through sequential networks (LSTMs). For example, full user profiles or conversation context can be computationally expensive. As stated in Section 8.4.6, the problem with gradient descent for RNN architectures is that error gradients vanish exponentially quickly with the number of time steps (i.e., number of words in the conversation context or user profile). Beside LSTM, recent work proposes to solve the issue with an independently recurrent neural network approach by reducing the context of a neuron to its former state and explore in the next layers the cross-neuron information [81]. It is promising to research direction for recommending conversations as the method help to learn memories of different range, including long-term memory, without the gradient vanishing and exploding problem.

Future research directions include improving the recommendation model to support significant conversational contexts that encapsulate global interests or preferences from users. By doing capturing global interest, it can reduce the model complexity as well as improve the precision of the recommendation tasks. To capture global user preferences, models can incorporate topics modeling into the recommending model, as stated in recent works [158, 43, 102]. This study also outlines several other aspects explored in recommending conversations on microblogs, such as incorporating multi-modal models [94] and multi-lingual pre-trained embeddings [35] that can help to deal with this kind of user interaction on social media. Finally, another aspect that is important to RSs is the ability to produce real-time recommendations. In domains such as crisis informatics, intelligent systems need to react immediately to online requests and provide recommendations to users regardless of the user history or conversation context.

Chapter 9

Conclusions

This thesis focused on six research chapters to address research problems concerning crisis-related conversation on Twitter, including several perspectives such as characterization, taxonomy, classification, and recommendation. Specifically, (1) in chapter 3, establishws a methodology to collected and built a corpus for crisis-related conversations and analyze the factor that ignites conversations; (2) in chapter 4, proposed a fine-grained taxonomy to categorize tweets crisis-related conversations; (3) in chapter 6, considered the task of context-aware categorization of crisis-related conversations; (4) in chapter 7, considered the task of semi-supervised multi-label classification of short-text on social media; (5) in chapter 5, considered the task of cross-lingual classification through contextual embeddings; (6) and, in chapter 8, proposed a model to recommend conversational content during crisis events.

This section summarizes the main findings and outlines future research directions. section 9.1 provide a detailed summary of the contributions of the research and provides an answer to the research questions enumerated in chapter 1. Then, it discusses directions for future work in section 9.2.

9.1 Main Findings

9.1.1 Characterizing crisis-related conversations

RQ1: What are the factors that ignite conversations? The analysis of Twitter data shows that 36% of the tweets are conversational in nature form threads with two or more tweets. Therefore, it investigated the factors that contribute to sparking a conversation on Twitter, i.e., identifying whether a tweet that will generate replies from other users. The goal was to predict f a tweet will evolve into a conversation rather the popularity as previous works. The predictive model depends on two main aspects: networking and activity level. The activity level is related to the number of favorites the user has given to other tweets or the number of tweets created.

The networking level is related to user attributes that indicate network relationships (followers, friends). Also, the content patterns of the tweets are a reliable indicator of tweets evolving into conversations, which is explained by a slight negative correlation with the feature mentions. In contrast, other content-related features (represented by the number of URLs, media, or hashtags) are slightly positive. The findings show that factors such as user mentions and hashtags are predominant in starting conversations.

RQ2: Does current taxonomies account for crisis-related conversations? Previous works have run into some issues to categorize tweets into different topical classes because: a) tweets can contain information that belongs to one or more classes; b) choosing the dominant category is difficult, even human scorers differ in their judgment about whether or not a tweet belongs to a specific category; and, c) the semantic ambiguity of the tweets, as well as the idiomatic phrases, sometimes makes it difficult to interpret them. Given these difficulties, this work proposed that in order to understand and analyze conversations in the context of crisis events, it is necessary to extend previous coarse-grained or generic taxonomies. For example a turn (tweets) categorized as an act of type Statement or Request based on generic taxonomies is not enough to extract useful information relevant to a crisis. This work proposed more detailed dialogue acts, such as Informative Statement, Complaint, Offers, or Requests to capture the intention of the participants. Similarly, turns often belongs to multiple overlapping conversational acts, so the proposed hierarchical multi-label approach provides more flexibility than single label approaches. With the proposed taxonomy, a classification task uses a CNN model to predict the fine-grained conversational acts for a conversation as well as the potential outcomes, such as prevention, situational awareness, and relief coordination. Although the taxonomy helps to understand the flow of conversations between users, the results show that as the number of labels increases, the classification scores decrease which indicated the necessity of a more substantial number of labeled tweets, especially for the tail classes or the use approaches such as zero or one-shot learning.

9.1.2 Modeling of crisis-related conversations

RQ3: How to deal with multiple languages that appear during a crisis event? This study introduced an annotated corpus of crisis-related tweets for Spanish and English language. As users worldwide can comment about the event, often in different languages about the same event, neural architectures allow identifying crisis-related tweets in a multi-lingual setting. The results show that deep contextual multi-lingual embeddings outperform strong baseline models. Analysis of the type of conversations that occur from the perspective of different languages identifies that certain types of conversations occur more in the native language and others in a foreign language. The findings show that conversations from foreign countries seek to gather situation awareness and give emotional support, while in the affected country, the conversations aim

mainly to humanitarian aid.

RQ4: Does the conversational context help downstream tweets classification tasks? Current classification approaches have focused on the analysis of individual tweets, which do not have enough context to disambiguate information. The analysis of conversations during several crisis events yields insights into the use of conversational context in NLP classification tasks. To that end, this study introduces a new corpus, the Conversational CrisisNLP dataset, and evaluate several non-neural and neural classifiers using contextual information from conversations. The empirical results show that the proposed approach slightly improves the performance of the classifiers in some cases for both binary and categorical classification tasks. Our findings indicate that using a more extended conversation context did improve the classification task from percentile 25 to 50 from the conversation history. The experiments validate the initial hypothesis through an interpretability model [54] for the binary classification task and applied it to the conversations, which extracts an interpretation of the classification model based on input sensitivity detected through an external mechanism. The results show that the interpretability results fit tightly with the initial hypothesis that classifiers can leverage additional data provided by conversation contexts to improve the understanding of the target tweet (labeled tweet). Although the gains are not significant, using context information could help to disambiguation or handling severe cases for the models in downstream NLP tasks beyond the Crisis Informatics domain.

RQ5: How to leverage the massive amount of unlabeled social media data for supervised tasks? This research question investigates how a semi-supervised approach can learn to categorize short texts in a multi-label taxonomy using a small set of labeled data and leveraging the availability of large amounts of unlabeled data on social media. To that end, this work proposes neural semi-supervised k-means clustering that modifies the normal objective function and adds a penalty term for labeled data [144]. The proposed model extended a neural semi-supervised clustering and applied it to multi-label settings. The results show that semi-supervised k-means outperform other baseline unsupervised models for multi-label classification tasks. The results are promising as the proposed model leverage the massive amount of data on social media and help to overcome the scarcity of labeled data.

RQ6: How to recommend users to join relevant conversations on social media? This research question analyzed Twitter conversations that may span a wide range of domains such as politics, religion, health, fitness, sports, food, among others. As conversational recommendations could be particularly useful during crisis events, such as earthquakes when people use social media platforms to look for relevant information, request for help, offer support, contact relatives, and other types of inquiries. This work proposes a novel recommendation task for OSNs, particularly recommending users to join relevant multi-agent and multi-turn conversations.

The recommendation tasks aim to recommend users to join a set of conversations on different topics they might be interested in, even when no link connections are available (i.e., users have no relationship between them). The proposed neural learning architecture uses a sequence-to-sequence model to tackle the task of recommending conversations using the conversation context and the user's history. The results show that recommending short-text conversations is feasible to train using sequence-to-sequence models without human-annotated data since the data itself provides the ground truth as in the case of the prediction of the next utterance. Empirical results show that the model outperforms several baseline models, such as TF-IDF and state of the art model based on collaborative filtering (CFT). The experiments highlight the performance of the bi-LSTM variant of the model, which outperforms other variants such as RNN or LSTM, and the potential for recommending content on social media.

9.2 Future Research Directions

9.2.1 Conversational chatbots

During crisis events, people turn to social media platforms to search for relevant information, ask advice, request help, offer support, among other things. On the other hand, humanitarian relief organizations, governmental agencies, or individual activists rely each time more on social media to gain situational awareness, learn urgent needs on affected zones, and coordinate relief efforts [100]. However, human resources are limited and cannot deal with the overwhelming volume of data generated during crisis events. Future work could explore the idea of building conversational agents that can interact with users in one-to-one or one-to-many, multi-turn conversations about several topics during natural disasters. The hypothesis is that multi-agent multi-turn conversations on social media, specifically in the context of natural disasters, can be modeled leveraging on deep neural architectures [83] trained over a large dataset collected from crisis-related conversations. The advantage of this task is that they do not require human annotation efforts because the ground truth is in the data itself.

9.2.2 Deep conversational reinforcement learning

Recently, several works proposed deep learning extensions of the classic reinforcement learning [145] such as Deep Q-Networks [92], Policy Networks [124], and Hierarchical Deep Reinforcement Learning [72]. Likewise, Deep Reinforcement Learning (DRL) have made inroads in several NLP tasks such as text games [97], social media [47], coreference resolution [31], knowledge graph reasoning [149], semi-supervised text classification [148], information extraction [98, 112], dialog systems [80, 160], and multimodal tasks [91, 143]. The focus is to predict the interest of users in

crisis-related conversations, similar to the work of predicting popularity prediction in Reddit forums [47]. The goal is to make recommendations based on the interest level of a broad group of users on a specific topic, in case conversations related to crisis events such as earthquakes. In this setting, a model tries to identify and track tweets in real-time, attempting to identify and recommend crisis-related tweets before they become popular conversations. The assumption is that the users capacity is limited, and models can make only a few recommendations out of the enormous space of possibilities on social media.

9.2.3 Image-grounded conversations

Another research direction could tackle the problem of understanding image grounded conversations during natural disasters that can provide further insights from data shared on social media. In this area, it is essential to know what type of information is published by users during a crisis event, where and when do image-grounded conversations arise during a crisis event, and how a model can categorize image-grounded conversations during a crisis event.

Appendix A

Appendix: Data Collection

A.1 Ecuador Earthquake Dataset

This appendix describe the data collection process using Twitter as the the main source of data for the analysis. This dataset initially use data from Artificial Intelligence for Disaster Response[1] (AIDR), which is an online platform to support disaster response that collects big crisis data from social networks [58]. Specifically, the data requested for AIDR is related to the Ecuadorian earthquake in 2016. The dataset contains randomly collected tweets from Twitter Streaming API.

In total, the collected corpus contains more than 150 million tweets, from January to April 2016 using the Twitter Streaming API[2]. Through this API, Twitter provides researchers with the 1% of its public data collected at a given time. To deal with the overwhelming amount of data obtained through Twitter Streaming API, the collection rely on Cassandra [74] as distributed storage. The choise Cassandra over other NoSQL databases base on its distributed architecture, scalability, and high availability without compromising performance as the database grows [4]. Figure A.1 shows the data capture and storage architecture used in this research work. It consists of four nodes, on each node running: a Twitter Capture Service that collects data from Twitter Streaming API, and a local Cassandra used to store the data.

The Twitter Streaming data collection processs use two types of filters: a geolocated bounding box, and tweets containing specific words. The bounding box filter allows to capture geolocated tweets in Ecuador. The data collection filter only in English or Spanish tweets or users who have specified those languages in their profiles, and exclude tweets in other languages. Also, considering that retweets can be a significant part of streaming data, and the fact that retweets only propagate content generated by other users, the process filter out retweets and only retain the original or root tweets. The reason is that the focus is the interactions in the form of conversations.

[1] `http://aidr.qcri.org/`
[2] `https://developer.twitter.com/en/docs/tweets/filter-realtime/overview`

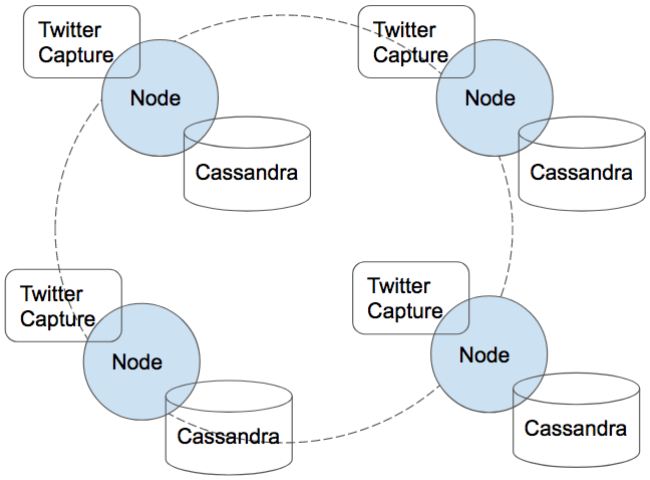


Figure A.1: Cassandra cluster to capture and store Twitter data.

The data augmentation step retrieves conversation trees (parents and replies tweets) for all the initial tweets in the dataset. To identify conversations, a preprocessing step uses tweet's field in-reply-to-status-id that specify if the tweet is a reply to another tweet, which allow to establish which tweet belongs to a specific conversation. If a parent tweet not present in the dataset, the data collection process obtain it through the Twitter REST API[3]. From the parent tweets (i.e., the tweets that initiate a conversation), the data augmentation crawls all the child tweets following the procedure in a previous work [133]. Next, the preprocessing step filters out isolated tweets and use only tweets that form part of conversations with at least one reply.

[3]https://dev.twitter.com/rest/public

Bibliography

[1] Earthquake facts and statistics.

[2] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In OSDI, volume 16, pages 265–283, 2016.

[3] Fabian Abel, Qi Gao, Geert-Jan Houben, and Ke Tao. Analyzing user modeling on twitter for personalized news recommendations. In International Conference on User Modeling, Adaptation, and Personalization, pages 1–12. Springer, 2011.

[4] Veronika Abramova and Jorge Bernardino. Nosql databases: Mongodb vs cassandra. In Proceedings of the international C* conference on computer science and software engineering, pages 14–22. ACM, 2013.

[5] Adam Acar and Yuya Muraki. Twitter for crisis communication: lessons learned from japan's tsunami disaster. International Journal of Web Based Communities, 7(3):392–402, 2011.

[6] Alan Akbik, Duncan Blythe, and Roland Vollgraf. Contextual string embeddings for sequence labeling. In COLING 2018, 27th International Conference on Computational Linguistics, pages 1638–1649, 2018.

[7] 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.

[8] Firoj Alam, Shafiq Joty, and Muhammad Imran. Graph based semi-supervised learning with convolution neural networks to classify crisis related tweets. In Twelfth International AAAI Conference on Web and Social Media, 2018.

[9] Noor Aldeen Alawad, Aris Anagnostopoulos, Stefano Leonardi, Ida Mele, and Fabrizio Silvestri. Network-aware recommendations of novel tweets. In Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, pages 913–916. ACM, 2016.

[10] James Allen, Nathanael Chambers, George Ferguson, Lucian Galescu, Hyuckchul Jung, Mary Swift, and William Taysom. Plow: A collaborative task learning agent. In AAAI, pages 1514–1519, 2007.

[11] Tim Althoff, Kevin Clark, and Jure Leskovec. Natural language processing for mental health: Large scale discourse analysis of counseling conversations. Transactions of the Association for Computational Linguistics, 2016.

[12] David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. In Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, pages 1027–1035. Society for Industrial and Applied Mathematics, 2007.

[13] Eric Bair. Semi-supervised clustering methods. Wiley Interdisciplinary Reviews: Computational Statistics, 5(5):349–361, 2013.

[14] Somnath Banerjee, Krishnan Ramanathan, and Ajay Gupta. Clustering short texts using wikipedia. In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, pages 787–788. ACM, 2007.

[15] Geoffrey Barbier, Reza Zafarani, Huiji Gao, Gabriel Fung, and Huan Liu. Maximizing benefits from crowdsourced data. Computational and Mathematical Organization Theory, 18(3):257–279, 2012.

[16] Susan A Bartels and Michael J VanRooyen. Medical complications associated with earthquakes. The Lancet, 379(9817):748–757, 2012.

[17] Moumita Basu, Saptarshi Ghosh, Arnab Jana, Somprakash Bandyopadhyay, and Ravikant Singh. Resource mapping during a natural disaster: a case study on the 2015 nepal earthquake. International Journal of Disaster Risk Reduction, 24:24–31, 2017.

[18] Mikhail Bilenko, Sugato Basu, and Raymond J Mooney. Integrating constraints and metric learning in semi-supervised clustering. In Proceedings of the twenty-first international conference on Machine learning, page 11. ACM, 2004.

[19] Steven Bird, Ewan Klein, and Edward Loper. Natural language processing with Python: analyzing text with the natural language toolkit. " O'Reilly Media, Inc.", 2009.

[20] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics, 5:135–146, 2017.

[21] Danah Boyd, Scott Golder, and Gilad Lotan. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In System Sciences (HICSS), 2010 43rd Hawaii International Conference on, pages 1–10. IEEE, 2010.

[22] Leo Breiman. Random forests. Machine learning, 45(1):5–32, 2001.

[23] Gregoire Burel and Harith Alani. Crisis event extraction service (crees)-automatic detection and classification of crisis-related content on social media. 2018.

[24] 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.

[25] 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.

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

[27] Chih-Chung Chang and Chih-Jen Lin. Libsvm: A library for support vector machines. ACM transactions on intelligent systems and technology (TIST), 2(3):27, 2011.

[28] Justin Cheng, Lada Adamic, P Alex Dow, Jon Michael Kleinberg, and Jure Leskovec. Can cascades be predicted? In Proceedings of the 23rd international conference on World wide web, pages 925–936. ACM, 2014.

[29] François Chollet et al. Keras. `https://keras.io`, 2015.

[30] D Manning Christopher, Raghavan Prabhakar, and Schutza Hinrich. Introduction to information retrieval. An Introduction To Information Retrieval, 151(177):5, 2008.

[31] Kevin Clark and Christopher D. Manning. Deep reinforcement learning for mention-ranking coreference models. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 2256–2262, Austin, Texas, November 2016. Association for Computational Linguistics.

[32] Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. Power-law distributions in empirical data. SIAM review, 51(4):661–703, 2009.

[33] Ian Davidson and Sugato Basu. A survey of clustering with instance level. Constraints, 1:2, 2007.

[34] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, pages 248–255. IEEE, 2009.

[35] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, 2019.

[36] Inderjit S Dhillon and Yuqiang Guan. Information theoretic clustering of sparse cooccurrence data. In Data Mining, 2003. ICDM 2003. Third IEEE International Conference on, pages 517–520. IEEE, 2003.

[37] Micha Elsner and Eugene Charniak. You talking to me? a corpus and algorithm for conversation disentanglement. In ACL, pages 834–842, 2008.

[38] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. Liblinear: A library for large linear classification. Journal of machine learning research, 9(Aug):1871–1874, 2008.

[39] Joseph L Fleiss. Measuring nominal scale agreement among many raters. Psychological bulletin, 76(5):378, 1971.

[40] Samah Fodeh, Bill Punch, and Pang-Ning Tan. On ontology-driven document clustering using core semantic features. Knowledge and information systems, 28(2):395–421, 2011.

[41] Michail Fragkias, Burak Güneralp, Karen C. Seto, and Julie Goodness. A Synthesis of Global Urbanization Projections, pages 409–435. Springer Netherlands, Dordrecht, 2013.

[42] Huiji Gao, Geoffrey Barbier, Rebecca Goolsby, and Daniel Zeng. Harnessing the crowdsourcing power of social media for disaster relief. Technical report, DTIC Document, 2011.

[43] Lin Gui, Jia Leng, Gabriele Pergola, Ruifeng Xu, Yulan He, et al. Neural topic model with reinforcement learning. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3469–3474, 2019.

[44] Pankaj Gupta, Ashish Goel, Jimmy Lin, Aneesh Sharma, Dong Wang, and Reza Zadeh. Wtf: The who to follow service at twitter. In Proceedings of the 22nd international conference on World Wide Web, pages 505–514. ACM, 2013.

[45] Ido Guy. Social recommender systems. In Recommender Systems Handbook, pages 511–543. Springer, 2015.

[46] Ji He, Mari Ostendorf, Xiaodong He, Jianshu Chen, Jianfeng Gao, Lihong Li, and Li Deng. Deep reinforcement learning with a combinatorial action space for predicting popular reddit threads. arXiv preprint arXiv:1606.03667, 2016.

[47] Ji He, Mari Ostendorf, Xiaodong He, Jianshu Chen, Jianfeng Gao, Lihong Li, and Li Deng. Deep reinforcement learning with a combinatorial action space for predicting popular Reddit threads. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 1838–1848, Austin, Texas, November 2016. Association for Computational Linguistics.

[48] Matthew Henderson, Blaise Thomson, and Steve Young. Word-based dialog state tracking with recurrent neural networks. In Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL), pages 292–299, 2014.

[49] Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and D Tikk. Session-based recommendations with recurrent neural networks. In 4th International Conference on Learning Representations, ICLR 2016, 2016.

[50] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. Neural computation, 9(8):1735–1780, 1997.

[51] Courtenay Honey and Susan C Herring. Beyond microblogging: Conversation and collaboration via twitter. In System Sciences, 2009. HICSS'09. 42nd Hawaii International Conference on, pages 1–10. IEEE, 2009.

[52] Liangjie Hong, Ovidiu Dan, and Brian D Davison. Predicting popular messages in twitter. In Proceedings of the 20th international conference companion on World wide web, pages 57–58. ACM, 2011.

[53] Yijue How and Min-Yen Kan. Optimizing predictive text entry for short message service on mobile phones. In Proceedings of HCII, volume 5, 2005.

[54] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 328–339, 2018.

[55] Yuening Hu, Jordan Boyd-Graber, Brianna Satinoff, and Alison Smith. Interactive topic modeling. Machine learning, 95(3):423–469, 2014.

[56] Ziniu Hu, Ting Chen, Kai-Wei Chang, and Yizhou Sun. Few-shot representation learning for out-of-vocabulary words. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 4102–4112, Florence, Italy, July 2019. Association for Computational Linguistics.

[57] 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.

[58] Muhammad Imran, Carlos Castillo, Ji Lucas, Patrick Meier, and Sarah Vieweg. Aidr: Artificial intelligence for disaster response. In Proceedings of the 23rd International Conference on World Wide Web, pages 159–162. ACM, 2014.

[59] 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.

[60] Muhammad Imran, Prasenjit Mitra, and Carlos Castillo. Twitter as a lifeline: Human-annotated twitter corpora for nlp of crisis-related messages. arXiv preprint arXiv:1605.05894, 2016.

[61] Muhammad Imran, Prasenjit Mitra, and Carlos Castillo. Twitter as a lifeline: Human-annotated twitter corpora for nlp of crisis-related messages. arXiv preprint arXiv:1605.05894, 2016.

[62] 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.

[63] Edward Ivanovic. Dialogue act tagging for instant messaging chat sessions. In Proceedings of the ACL Student Research Workshop, pages 79–84. Association for Computational Linguistics, 2005.

[64] Edward Ivanovic. Automatic instant messaging dialogue using statistical models and dialogue acts. 2008.

[65] Aishwarya Jadhav and Vaibhav Rajan. Extractive summarization with swap-net: Sentences and words from alternating pointer networks. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 142–151, 2018.

[66] Sina Jafarpour, Chris Burges, and Alan Ritter. Filter, rank, and transfer the knowledge: Learning to chat. Advances in Ranking, 10, 2010.

[67] Kokil Jaidka, Saran Mumick, Niyati Chhaya, and Lyle Ungar. The CL-Aff Happiness Shared Task: Results and Key Insights. In Proceedings of the 2nd Workshop on Affective Content Analysis @ AAAI (AffCon2019), Honolulu, Hawaii, January 2019.

[68] 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.

[69] Su Nam Kim, Lawrence Cavedon, and Timothy Baldwin. Classifying dialogue acts in multi-party live chats. In Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation, pages 463–472, 2012.

[70] Yoon Kim. Convolutional neural networks for sentence classification. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1746–1751, 2014.

[71] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

[72] Tejas D Kulkarni, Karthik Narasimhan, Ardavan Saeedi, and Josh Tenenbaum. Hierarchical deep reinforcement learning: Integrating temporal abstraction and intrinsic motivation. In Advances in neural information processing systems, pages 3675–3683, 2016.

[73] 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.

[74] Avinash Lakshman and Prashant Malik. Cassandra: a decentralized structured storage system. ACM SIGOPS Operating Systems Review, 44(2):35–40, 2010.

[75] Andrew K. Lampinen and James L. McClelland. One-shot and few-shot learning of word embeddings. CoRR, abs/1710.10280, 2017.

[76] J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. biometrics, pages 159–174, 1977.

[77] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. Nature, 521(7553):436–444, 2015.

[78] Kathy Lee, Ankit Agrawal, and Alok Choudhary. Real-time disease surveillance using twitter data: demonstration on flu and cancer. In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 1474–1477. ACM, 2013.

[79] Cheng Li, Yue Lu, Qiaozhu Mei, Dong Wang, and Sandeep Pandey. Click-through prediction for advertising in twitter timeline. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 1959–1968. ACM, 2015.

[80] Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. Deep reinforcement learning for dialogue generation. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 1192–1202, Austin, Texas, November 2016. Association for Computational Linguistics.

[81] Shuai Li, Wanqing Li, Chris Cook, Ce Zhu, and Yanbo Gao. Independently recurrent neural network (indrnn): Building a longer and deeper rnn. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 5457–5466, 2018.

[82] Gustavo López, Luis Quesada, and Luis A Guerrero. Alexa vs. siri vs. cortana vs. google assistant: A comparison of speech-based natural user interfaces. In International Conference on Applied Human Factors and Ergonomics, pages 241–250. Springer, 2017.

[83] Ryan Lowe, Nissan Pow, Iulian Vlad Serban, and Joelle Pineau. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 285–294, 2015.

[84] Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. Scoring, term weighting and the vector space model. Introduction to information retrieval, 100:2–4, 2008.

[85] Ramesh Manuvinakurike, Maike Paetzel, Cheng Qu, David Schlangen, and David DeVault. Toward incremental dialogue act segmentation in fast-paced interactive dialogue systems. In Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 252–262, 2016.

[86] LR Medsker and LC Jain. Recurrent neural networks. Design and Applications, 5, 2001.

[87] Stuart E Middleton, Lee Middleton, and Stefano Modafferi. Real-time crisis mapping of natural disasters using social media. IEEE Intelligent Systems, 29(2):9–17, 2014.

[88] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781, 2013.

[89] Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. Advances in pre-training distributed word representations. In Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018), 2018.

[90] 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.

[91] Dipendra Misra, John Langford, and Yoav Artzi. Mapping instructions and visual observations to actions with reinforcement learning. arXiv preprint arXiv:1704.08795, 2017.

[92] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. Nature, 518(7540):529, 2015.

[93] 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 Seventh international AAAI conference on weblogs and social media, 2013.

[94] Nasrin Mostafazadeh, Chris Brockett, Bill Dolan, Michel Galley, Jianfeng Gao, Georgios Spithourakis, and Lucy Vanderwende. Image-grounded conversations: Multimodal context for natural question and response generation. In Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 462–472, 2017.

[95] Jonas Mueller and Aditya Thyagarajan. Siamese recurrent architectures for learning sentence similarity. In thirtieth AAAI conference on artificial intelligence, 2016.

[96] James Munkres. Algorithms for the assignment and transportation problems. Journal of the society for industrial and applied mathematics, 5(1):32–38, 1957.

[97] Karthik Narasimhan, Tejas Kulkarni, and Regina Barzilay. Language understanding for text-based games using deep reinforcement learning. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 1–11, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[98] Karthik Narasimhan, Adam Yala, and Regina Barzilay. Improving information extraction by acquiring external evidence with reinforcement learning. arXiv preprint arXiv:1603.07954, 2016.

[99] Nasser M Nasrabadi. Pattern recognition and machine learning. Journal of electronic imaging, 16(4):049901, 2007.

[100] Dat Tien Nguyen, Kamla Al-Mannai, Shafiq R Joty, Hassan Sajjad, Muhammad Imran, and Prasenjit Mitra. Robust classification of crisis-related data on social networks using convolutional neural networks. In ICWSM, pages 632–635, 2017.

[101] Brendan O'Connor, Ramnath Balasubramanyan, Bryan R Routledge, Noah A Smith, et al. From tweets to polls: Linking text sentiment to public opinion time series. Icwsm, 11(122-129):1–2, 2010.

[102] Byungkook Oh, Seungmin Seo, Cheolheon Shin, Eunju Jo, and Kyong-Ho Lee. Topic-guided coherence modeling for sentence ordering by preserving global and local information. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 2273–2283, 2019.

[103] Shereen Oraby, Pritam Gundecha, Jalal Mahmud, Mansurul Bhuiyan, and Rama Akkiraju. How may i help you?: Modeling twitter customer serviceconversations using fine-grained dialogue acts. In Proceedings of the 22nd International Conference on Intelligent User Interfaces, pages 343–355. ACM, 2017.

[104] John D Owens, Mike Houston, David Luebke, Simon Green, John E Stone, and James C Phillips. Gpu computing. Proceedings of the IEEE, 96(5):879–899, 2008.

[105] Leysia Palen, Kenneth M Anderson, Gloria Mark, James Martin, Douglas Sicker, Martha Palmer, and Dirk Grunwald. A vision for technology-mediated support for public participation & assistance in mass emergencies & disasters. In Proceedings of the 2010 ACM-BCS visions of computer science conference, page 8. British Computer Society, 2010.

[106] Leysia Palen and Amanda L Hughes. Social media in disaster communication. In Handbook of Disaster Research, pages 497–518. Springer, 2018.

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

[108] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. Journal of machine learning research, 12(Oct):2825–2830, 2011.

[109] Jeffrey Pennington, Richard Socher, and Christopher 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.

[110] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), volume 1, pages 2227–2237, 2018.

[111] Hemant Purohit, Carlos Castillo, Fernando Diaz, Amit Sheth, and Patrick Meier. Emergency-relief coordination on social media: Automatically matching resource requests and offers. First Monday, 19(1), 2013.

[112] Pengda Qin, Weiran Xu, and William Yang Wang. Robust distant supervision relation extraction via deep reinforcement learning. arXiv preprint arXiv:1805.09927, 2018.

[113] Jason D Rennie, Lawrence Shih, Jaime Teevan, and David R Karger. Tackling the poor assumptions of naive bayes text classifiers. In Proceedings of the 20th international conference on machine learning (ICML-03), pages 616–623, 2003.

[114] Anthony Rios and Ramakanth Kavuluru. Few-shot and zero-shot multi-label learning for structured label spaces. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 3132–3142, 2018.

[115] Alan Ritter, Colin Cherry, and Bill Dolan. Unsupervised modeling of twitter conversations. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pages 172–180. Association for Computational Linguistics, 2010.

[116] Alan Ritter, Colin Cherry, and William B Dolan. Data-driven response generation in social media. In Proceedings of the conference on empirical methods in natural language processing, pages 583–593. Association for Computational Linguistics, 2011.

[117] Koustav Rudra, Siddhartha Banerjee, Niloy Ganguly, Pawan Goyal, Muhammad Imran, and Prasenjit Mitra. Summarizing situational tweets in crisis scenario. In Proceedings of the 27th ACM Conference on Hypertext and Social Media, pages 137–147. ACM, 2016.

[118] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science, 1985.

[119] Alan Said and Alejandro Bellogín. Comparative recommender system evaluation: benchmarking recommendation frameworks. In Proceedings of the 8th ACM Conference on Recommender systems, pages 129–136. ACM, 2014.

[120] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In Proceedings of the 19th international conference on World wide web, pages 851–860. ACM, 2010.

[121] AM Saxe, JL McClelland, and S Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. International Conference on Learning Represenatations 2014, 2014.

[122] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing, 45(11):2673–2681, 1997.

[123] Lifeng Shang, Zhengdong Lu, and Hang Li. Neural responding machine for short-text conversation. In Proceedings of the 53rd Annual Meeting of the Association for Computational

Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 1577–1586, 2015.

[124] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. nature, 529(7587):484, 2016.

[125] Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. A neural network approach to context-sensitive generation of conversational responses. In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 196–205, 2015.

[126] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research, 15(1):1929–1958, 2014.

[127] Bongwon Suh, Lichan Hong, Peter Pirolli, and Ed H Chi. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In Social computing (socialcom), 2010 ieee second international conference on, pages 177–184. IEEE, 2010.

[128] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In Advances in neural information processing systems, pages 3104–3112, 2014.

[129] Chenhao Tan, Lillian Lee, and Bo Pang. The effect of wording on message propagation: Topic-and author-controlled natural experiments on twitter. arXiv preprint arXiv:1405.1438, 2014.

[130] Andrea H Tapia, Kartikeya Bajpai, Bernard J Jansen, John Yen, and Lee Giles. Seeking the trustworthy tweet: Can microblogged data fit the information needs of disaster response and humanitarian relief organizations. In Proceedings of the 8th International ISCRAM Conference, pages 1–10, 2011.

[131] Luis Terán, Alvin Oti Mensah, and Arianna Estorelli. A literature review for recommender systems techniques used in microblogs. Expert Systems with Applications, 103:63–73, 2018.

[132] Johnny Torres and Carmen Vaca. Cl-aff deep semisupervised clustering. 2019.

[133] Johnny Torres, Carmen Vaca, and Cristina L Abad. What ignites a reply?: Characterizing conversations in microblogs. In Proceedings of the Fourth IEEE/ACM International Conference on Big Data Computing, Applications and Technologies, pages 149–156. ACM, 2017.

[134] Johnny Torres, Carmen Vaca, Luis Terán, and Cristina L Abad. Seq2seq models for recommending short text conversations. Expert Systems with Applications, 150:113270, 2020.

[135] Johnny Torres, Carmen Vaca. Cross-lingual perspectives about crisis-related conversations on twitter. In Companion Proceedings of The 2019 World Wide Web Conference, pages 255–261, 2019.

[136] István Varga, Motoki Sano, Kentaro Torisawa, Chikara Hashimoto, Kiyonori Ohtake, Takao Kawai, Jong-Hoon Oh, and Stijn De Saeger. Aid is out there: Looking for help from tweets during a large scale disaster. In ACL (1), pages 1619–1629, 2013.

[137] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Advances in Neural Information Processing Systems, pages 5998–6008, 2017.

[138] Sudha Verma, Sarah Vieweg, William J Corvey, Leysia Palen, James H Martin, Martha Palmer, Aaron Schram, and Kenneth Mark Anderson. Natural language processing to the rescue? extracting" situational awareness" tweets during mass emergency. In ICWSM, pages 385–392. Barcelona, 2011.

[139] Sarah Vieweg, Carlos Castillo, and Muhammad Imran. Integrating social media communications into the rapid assessment of sudden onset disasters. In International Conference on Social Informatics, pages 444–461. Springer, 2014.

[140] Soroush Vosoughi and Deb Roy. Tweet acts: A speech act classifier for twitter. In ICWSM, pages 711–715, 2016.

[141] Hao Wang, Naiyan Wang, and Dit-Yan Yeung. Collaborative deep learning for recommender systems. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 1235–1244. ACM, 2015.

[142] Lidan Wang and Douglas W Oard. Context-based message expansion for disentanglement of interleaved text conversations. In Proceedings of human language technologies: The 2009 annual conference of the North American chapter of the association for computational linguistics, pages 200–208. Association for Computational Linguistics, 2009.

[143] Xin Wang, Wenhu Chen, Jiawei Wu, Yuan-Fang Wang, and William Yang Wang. Video captioning via hierarchical reinforcement learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4213–4222, 2018.

[144] Zhiguo Wang, Haitao Mi, and Abraham Ittycheriah. Semi-supervised clustering for short text via deep representation learning. arXiv preprint arXiv:1602.06797, 2016.

[145] Christopher John Cornish Hellaby Watkins. Learning from delayed rewards. PhD thesis, King's College, Cambridge, 1989.

[146] Katrin Weller, Axel Bruns, Jean Burgess, Merja Mahrt, and Cornelius Puschmann. Twitter and society, volume 89. P. Lang, 2014.

[147] Yorick Wilks. Artificial companions as a new kind of interface to the future internet. 2006.

[148] Jiawei Wu, Lei Li, and William Yang Wang. Reinforced co-training. arXiv preprint arXiv:1804.06035, 2018.

[149] Wenhan Xiong, Thien Hoang, and William Yang Wang. Deeppath: A reinforcement learning method for knowledge graph reasoning. arXiv preprint arXiv:1707.06690, 2017.

[150] Jiaming Xu, Peng Wang, Guanhua Tian, Bo Xu, Jun Zhao, Fangyuan Wang, and Hongwei Hao. Short text clustering via convolutional neural networks. In VS@ HLT-NAACL, pages 62–69, 2015.

[151] Xiwang Yang, Yang Guo, Yong Liu, and Harald Steck. A survey of collaborative filtering based social recommender systems. Computer Communications, 41:1–10, 2014.

[152] Tae Yano, William W Cohen, and Noah A Smith. Predicting response to political blog posts with topic models. In Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pages 477–485. Association for Computational Linguistics, 2009.

[153] Shaozhi Ye and Shyhtsun Felix Wu. Measuring message propagation and social influence on twitter. com. SocInfo, 10:216–231, 2010.

[154] An Gie Yong and Sean Pearce. A beginner's guide to factor analysis: Focusing on exploratory factor analysis. Tutorials in quantitative methods for psychology, 9(2):79–94, 2013.

[155] Lei Yu, Karl Moritz Hermann, Phil Blunsom, and Stephen Pulman. Deep learning for answer sentence selection. In NIPS Deep Learning and Representation Learning Workshop, Montreal, 2014.

[156] Yang Yu, Xiaojun Wan, and Xinjie Zhou. User embedding for scholarly microblog recommendation. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 449–453, Berlin, Germany, August 2016. Association for Computational Linguistics.

[157] Elina Zarisheva and Tatjana Scheffler. Dialog act annotation for twitter conversations. In Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 114–123, 2015.

[158] Jichuan Zeng, Jing Li, Yan Song, Cuiyun Gao, Michael R Lyu, and Irwin King. Topic memory networks for short text classification. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 3120–3131, 2018.

[159] Xingshan Zeng, Jing Li, Lu Wang, Nicholas Beauchamp, Sarah Shugars, and Kam-Fai Wong. Microblog conversation recommendation via joint modeling of topics and discourse. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 375–385, 2018.

[160] Xianda Zhou and William Yang Wang. Mojitalk: Generating emotional responses at scale. arXiv preprint arXiv:1711.04090, 2017.

[161] Ciyou Zhu, Richard H Byrd, Peihuang Lu, and Jorge Nocedal. Algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound-constrained optimization. ACM Transactions on Mathematical Software (TOMS), 23(4):550–560, 1997.

[162] Yizhe Zhu, Mohamed Elhoseiny, Bingchen Liu, Xi Peng, and Ahmed Elgammal. A generative adversarial approach for zero-shot learning from noisy texts. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1004–1013, 2018.