Cross-lingual Perspectives about Crisis-Related Conversations on Twitter

Johnny Torres
Escuela Superior Politécnica del Litoral (ESPOL)
Guayaquil, Ecuador
jomatorr@espol.edu.ec

Carmen Vaca
Escuela Superior Politécnica del Litoral (ESPOL)
Guayaquil, Ecuador
cvaca@fiec.espol.edu.ec

ABSTRACT

The role of social networks during natural disasters is becoming crucial to share relevant information and coordinate relief actions. With the reach of the social networks, any user around the world has the possibility of interact in crisis-events as these unfold. A large part of the information posted during a disaster uses the native language where the disaster occurred. However, there are also users from other parts of the world who can comment about the event, often in another language. In this work, we conducted a study of crisis-related tweets about the earthquake that occurred in Ecuador in April 2016. To that end, we introduce a new annotated dataset in both Spanish and English languages with approximately 8K tweets; half of them belong to conversations. We evaluate several neural architectures to identify crisis-related tweets in a multi-lingual setting, and we found that deep contextual multi-lingual embeddings outperform other strong baseline models. We then explore the type of conversations that occur from the perspective of different languages. The results show that certain types of conversations occur more in the native language and others in a foreign language. 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.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in collaborative and social computing**; • **Information systems** → *Collaborative and social computing systems and tools*; *Social networks.*

KEYWORDS

Social Computing; NLP; Neural Networks

ACM Reference Format:
Johnny Torres and Carmen Vaca. 2019. Cross-lingual Perspectives about Crisis-Related Conversations on Twitter. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW'19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3308560.3316799

WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316799

Table 1: A conversation about the Earthquake in Ecuador initiated by an organization's Twitter account.

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

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 to respond quickly to the urgent needs of affected people, assess damages in the buildings and infrastructure, identify medical emergencies, or coordinate relief actions [8, 11]. With the worldwide reach of social media, users from different countries and languages can react and interact in any crisis event. In this paper, we aim to identify the crisis-related tweets in a multi-lingual scenario and characterize them in the context of conversations. To that end, we introduce a new annotated dataset about the Earthquake occurred in Ecuador on April 16, 2016. The corpus contains 8360 tweets annotated for English and Spanish language. The table 1 shows an example of a conversation between several users on Twitter. Based on the metadata (*in reply to* field) provided by Twitter, we can collect all the tweets that belong to conversations.

For the annotation task, we leverage on taxonomy proposed by Imran et al. [13] , in which 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, we 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.

Despite recent progress in natural language processing (NLP), the semantic interpretation of noisy short-texts remains a hard problem. A multi-lingual scenario difficult even more the interpretation task. To tackle these issues, we evaluate a neural architecture to identify crisis-related tweets across multiple languages, specifically in the Ecuadorian earthquake study case. The contribution of our work is as follow. a) we introduce an annotated corpus of crisis-related tweets for Spanish and English language, b) we evaluate deep contextual neural architectures for the multi-lingual classification task at hand, and c) we characterize the conversations from locals and foreigners about the study case earthquake. We made available the dataset and the code to the community[1].

The rest of the paper is structured as follows: we begin with a discussion of related work in section 2, followed the description of the dataset in section 3. Then, we describe the taxonomy and the annotation process in section 4. Next, we present the proposed cross-lingual classification model in section 5; and, we discuss the results and the limitations in section 6. Finally, we outline the conclusions and future work in section 8.

2 RELATED WORK

Previous works analyze the usefulness of large 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 [8, 11].

It can be crucial during natural disasters to gain insight into the situation as it unfolds for the relief efforts by organizations and activists [1, 24, 27]. Several systems have been implemented to classify, extract, and summarize crisis-relevant information from social media [11].

Imran et al. [12] implemented the AIDR platform to collect and classify data streams during crises on Twitter. The AIDR platform has been instrumental in creating public datasets[2] to advance the research in the area of crisis informatics. Although, there is a large number of crisis events including earthquakes, floods, hurricanes, and cyclones; to the best of our knowledge the Ecuadorian earthquake is missing, and this work contributes annotated data for this event.

Several learning methods have been proposed to classify and categorize crisis-related data. Verma et al. [29] used Naive Bayes and MaxEnt classifiers to learn situational awareness tweets from several crises. Cameron et al. [7] proposed a framework for emergency awareness using an SVM classifier to identify useful tweets during natural disasters. More recently, Imran et al. [11] uses a traditional approach that requires manually engineered features like cue words and TF-IDF vectors for learning representation and training the model.

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*). Imran et al. [14] tackle this problem and show the performance of some non-neural network classifiers trained on labeled data from past crisis events. Recent approaches use neural architectures to deal with the issue the *out-of-event data*, specifically Convolutional Neural Networks (CNN) [6, 20] and semi-supervised learning [3, 4].

Social media reach worldwide audiences; therefore learning methods have to deal with a myriad of languages for user-generated data during crisis events. To tackle this problem, Khare et al. [15] evaluated several scenarios of learning methods with cross-lingual data. They use 30 cross-lingual datasets of crisis events where the model is trained on one language and tested in another language. Their approach uses an SVM classifier with several semantic features in addition to the tweets' text. We leverage the experimental setup proposed by Khare et al. [15] and evaluate an end-to-end learning based multi-lingual deep contextual embeddings and neural architectures, focusing only on the Ecuadorian Earthquake.

Table 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.

	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

3 DATASET

The dataset presented in this paper contains tweets provided by AIDR Research Lab[3], a platform that collects crisis-related tweets [12]. The initial dataset contained tweets collected using the Twitter Streaming API for the Ecuadorian Earthquake in April 2016. The Streaming API allows anyone to retrieve at most a 1% sample of all the Twitter's data filtered by some parameters [19]. We then augmented the dataset by retrieving the entire conversation tree for the tweets in the initial dataset. To collect the complete conversations, we crawl the parent tweets using the field *[in reply to status id]* in the tweets' metadata. From the parent tweets (i.e., the tweets that initiate a conversation), we crawled all the child tweets following the procedure in [26].

The Table 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. We noticed that for *other* languages, the number of the average number of tweets is higher than English or Spanish.

In the second section, we filter out isolated tweets and use only tweets that form part of conversations with at least one reply. The

[1] https://github.com/johnnytorres/crisis_conv_crosslingual
[2] http://crisisnlp.qcri.org/
[3] http://aidr.qcri.org/

conversations row refers to the number of conversations by language. We identify each conversation by its initial tweet. Then, we have the number of *replies*, in total 55%, of the tweets in the dataset belong to a conversation. Although the average of the number of the number of replies and users vary across languages, the median resulted similar for all languages (num replies = 3 and num users = 2) due to the constraint in the preprocessing of the conversations. The average is far from the median for the number of *replies* and *users*, indicating some outliers (i.e., some popular conversations often initiated by influencers). Also, we noticed that the average of *other* languages is higher than Spanish or English due to outliers in the number of replies to a specific type of conversations (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 1 depicts the distribution of the number of replies on the conversations.

For further analysis, we perform several steps of pre-processing to the conversations. We filter out conversations non-alternating users' tweets (single tweet per user), as well as conversations with less than 3 tweets and more than 10 tweets[4]. We used 3 as the lower bound for the English dataset due to the limited number of conversations in that language. After the preprocessing, our dataset contains 518 Spanish and 172 English conversations.

4 ANNOTATION

We select the 518 Spanish and 172 English conversations including the replies for the annotation task, which account in total 2193 and 730 tweets Spanish and English respectively. Additionally, we randomly select tweets regardless if it is in a conversation until complete approximately 4000 tweets for each language. We rely on the multi-class taxonomy proposed by Imran et al. [13] for the annotation task. In the multi-class taxonomy, 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

[4]The lower bound was set to allow for at least two replies in the conversation, and we define the upper-bound after finding that 84% of the conversations had replies ≤ 10

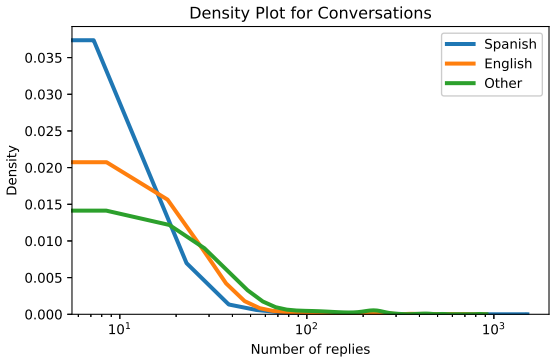


Figure 1: Distribution of the conversations by language.

We conduct our annotation process with two undergraduate students as human annotators and present them with the tweets selected for annotation. We provide to the annotators with instructions for the annotation task describing each of the categories and associated examples. For each tweet, we ask the annotators to select only one category based taxonomy. A third annotator resolves the discrepancies or disagreements between the two annotators by assigning the final category to the tweets without agreement.

Similar to the process in Imran et al. [13], we ask the annotator to make the judgments based on the text only, even if Twitter APIs truncated the text during the data collection. Also, we ask them not to open any link inside the text of the tweets as we seek to use only the available text for training the machine learning models. Since each tweet in our annotation process can have only one label, we calculate the percentage of agreement (p_o) between the two annotators. We also calculate the Fleiss' Kappa (κ) [9] agreement metric between the annotators. We calculate kappa metric 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.

As shown in Table 3, we find 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 an error in the annotation affects by a large margin the agreement score. We also observe that some tweets can belong to several categories. In these cases, the annotators might differ in their judgment, and this lowers the agreement score even in categories with a large number of instances.

5 MODEL

In this section, we evaluate different approaches for modeling crisis-related tweets on multi-lingual settings for our annotated corpus. First, we describe the learning representation for the tweets, then the neural architecture for classification used, and finally the experimental settings.

5.1 Learning Representation

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 [11]. In this work, we rely on the use of dense representation such as word embeddings Word2Vec [17]. Mikolov et al. [17] 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 based on the context words, and **b)** a *skip-gram* model that predicts surrounding words given the current word. The approach show that both models can learn high accuracy syntactic and semantic regularities and overcome the issues in sparse representation models.

Word2Vec [17] represents each word in the corpus like an atomic entity and generates a embeddings vector for each word. In this aspect, Word2Vec and Glove [21] are similar; both define *words*

Table 3: Agreement statistics for Spanish and English tweets.

	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

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 [18] 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 [22] and Flair [2], generate embeddings for a word based on the context, thus generating slightly different embeddings for each word depending on the context of its occurrence. We leverage on recent advances of the learning representation methods using the pre-trained embeddings and also fine-tuning the embeddings to our dataset.

5.2 Neural architecture

We are interested in identifying if a given tweet is related to a crisis event. We frame the problem of detecting crisis-related tweets as a multi-class classification task. To that end, we rely on both 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 [20].

We train the CNN and LSTM models by optimizing the binary cross entropy loss using the adaptive gradient-based learning algorithm [16]. We set the learning rate and parameters to the values as suggested by the authors. We set the number of epochs to 10 for the case of the random embeddings and 5 when using the stacked embeddings with Glove-Flair [2].

We use dropout [25] after embeddings and hidden units to avoid overfitting. We initialized the word vectors in L with random embeddings in the case of CNN. In the case of LSTM, we use a vanilla version with random embedding and a Stacked Glove-Flair pre-trained embedding with LSTM.

5.3 Experimental Settings

Data Preprocessing: We normalize all characters to their lowercased forms, truncate elongations to three characters, convert every digit to `D`, anonymize twitter usernames to `userID`, and all URLs to `HTTP`. We remove all punctuation marks except periods, semicolons, question and exclamation marks. We tokenize the tweets using the NLTK toolkit [5].

Label Grouping: Due to the imbalance of the labels, we group the labels into 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 4 shows the grouping labels that we use for the experiments. The grouping into two classes achieves a reasonable balance for both Spanish and English languages.

Table 4: Dataset for the binary classification task.

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

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. We randomly split the dataset into train and test sets to ensure that the classes distribution remains reasonably balanced in each set.

Classification Tasks: We define two set of experiments as described next. First, we train and test a model on the same language: train and predict on the Spanish dataset, and similarly train and predict on the English dataset. Second, we train a model on a language and predict in another language, e.g., train on Spanish dataset and predict on English dataset. We do not set up a third scenario that involves translating the dataset to a single language as described in Khare et al. [15] due to the cost constraint of using a translation API.

Classification Models: We implement a classifier based on Linear Regression (LR) as a baseline model and models based on neural architectures such as CNN and LSTM. Additionally, we evaluate an

Table 5: Performance of the models in single language and multi-lingual classification.

model	LR			LSTM			CNN			LSTM Stacked		
language	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**

LSTM model that uses Stacked Embeddings (Glove word embeddings + Flair deep contextual embeddings).

Evaluation Metrics: We use *Precision*, *Recall*, and F_1 macro metrics to evaluate the performance of classification models, and we report the results on the test set.

6 RESULTS

The Table 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 that means 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 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). Traditional approaches as LR using TF-IDF fail to generalize, and the performance fall in the case *es-en* and drop 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 generalize well in this setting and outperform other models, which is promising for tackling the multi-lingual scenarios in detecting crisis-related tweets.

7 DISCUSSION

This section details some of the findings related to the crisis-related conversations, cross-lingual analysis, and current limitations in this work.

7.1 Crisis-related conversations

We explore the conversational nature of the interactions on Twitter and how we can improve extracting insights during crisis events by analyzing entire conversations. In this analysis, we focus 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.

We look at the replies of root tweets not related to the crisis, and we found most of the replies (97%) are also not related to the earthquake, but 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. This small percentage indicates that we might lose some information even when the root tweet is not related to the crisis; thus we need to understand it in the context of the conversation.

A similar situation occurs in the replies to root tweets that are related to the crisis. In this case, there is a 24% of the replies that are annotated as not related to the crisis (in the case of English conversations up to 35%), when those tweets are part of conversations related to the crisis. In this case, these tweets could help to analyze the objective or outcome of the conversation semantically.

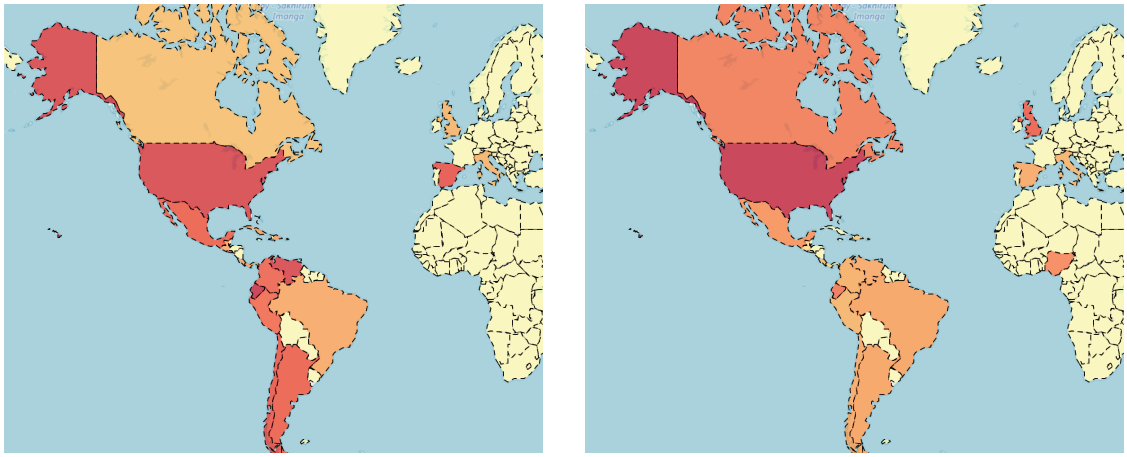
7.2 Cross-lingual analysis

Using the best model, we categorize the complete dataset to identify tweets related to the crisis event. We establish the country of the tweets' users by extracting that information from the *location* field in the metadata of the tweets. We preprocess the *location* field as it contains location information. Approximately 30% of the users do not provide location, and 15% contain noisy information in location. For our analysis, we consider only those with location information.

The Figure 2 shows the spatial distribution for each of the languages. In the Figure 2a, we show the crisis-related tweets in the Spanish language. For the Spanish language, the predominant country is where the earthquake occurred, Ecuador, followed by the United States, Venezuela. Also, there is a significant percentage in Spain, where there is a large number of Ecuadorian immigrants.

The Figure 2b shows that the Spanish-speaking countries decrease their participation in English dataset. The United States is the country where most tweets in English are posted, followed by England and Canada. An interesting issue would be to determine what percentage of the users are Ecuadorian immigrants living abroad.

Another aspect is the percentage of the annotated tweets associated to each category depending on the language. For the Spanish dataset, we identify the location information for approximately 49%



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

Figure 2: Spatial distribution of crisis-related tweets, darker color indicates larger percentage of tweets.

of the tweets. 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%.

For the English dataset, we identify the location information for approximately 77% of the tweets. 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%.

7.3 Limitations

We identify several issues while conducting experiments that are important to mention. Due to the limited size of our annotated dataset, we do not claim to make conclusive observations in the analysis of crisis-related conversations, local vs. foreigners, and 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 we treat each branch as an independent conversation, which could cause duplicated tweets in the dataset. In the dataset splitting strategy, we perform 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. Using the context for learning to predict a new tweet could improve the performance of the classification models, we consider a future research direction, especially using attention mechanisms [28].

Due to the ambiguity in some tweets, assigning labels can be confusing even for human annotators, and a limitation in our annotation process is the number of annotation per tweet compared to previous works. Similar to previous works on crisis-related datasets, our 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* [30] or *Few-shot* [23] 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 our dataset. Further evaluation of additional cross-lingual embeddings and language models are essential for transfer learning approaches to overcome the scarcity of annotated data in low-resource languages [10].

8 CONCLUSION

In this paper, we introduce an annotated corpus in Spanish and English language for the earthquake that occurred in Ecuador on April 16, 2016. We annotated not only isolated tweets but those that belong to conversations regarding the earthquake, which will enable future research for deeper understanding of that type of interactions and their outcome. We find that tweets often overlap semantically, an indication that a multi-label annotation would be more suitable for a more in-depth understanding of more complex interactions such as conversations. For our annotated corpus, we explore how we can identify crisis-related tweets on multi-lingual settings leveraging advances on multi-lingual deep contextual embeddings. We found that multi-lingual embedding outperforms other approaches based on sparse representation; however, for single language more simple model still perform better.

Future research directions include a comparative analysis of cross-lingual modeling of crisis-events using additional datasets publicly available. We also aim to improve the learning of labels with few instances, as in our corpus, through zero-shot or few show learning methods. Finally, we also seek to develop a fine-grained taxonomy as we detect approximately 15% of the tweets can belong to multiple labels.

REFERENCES

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

[2] Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual String Embeddings for Sequence Labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*. 1638–1649.

[3] Firoj Alam, Shafiq Joty, and Muhammad Imran. 2018. Domain Adaptation with Adversarial Training and Graph Embeddings. *arXiv preprint arXiv:1805.05151* (2018).

[4] Firoj Alam, Shafiq Joty, and Muhammad Imran. 2018. Graph Based Semi-Supervised Learning with Convolution Neural Networks to Classify Crisis Related Tweets. In *Twelfth International AAAI Conference on Web and Social Media*.

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

[6] Gregoire Burel and Harith Alani. 2018. Crisis Event Extraction Service (CREES)-Automatic Detection and Classification of Crisis-related Content on Social Media. (2018).

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

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

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

[10] Jeremy Howard and Sebastian Ruder. 2018. 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)*, Vol. 1. 328–339.

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

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

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

[14] Muhammad Imran, Prasenjit Mitra, and Jaideep Srivastava. 2016. 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.

[15] Prashant Khare, Grégoire Burel, Diana Maynard, and Harith Alani. 2018. Cross-Lingual Classification of Crisis Data. In *International Semantic Web Conference*. Springer, 617–633.

[16] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

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

[18] Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in Pre-Training Distributed Word Representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

[19] Fred Morstatter, Jürgen Pfeffer, Huan Liu, and Kathleen M Carley. 2013. 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*.

[20] Dat Tien Nguyen, Kamla Al-Mannai, Shafiq R Joty, Hassan Sajjad, Muhammad Imran, and Prasenjit Mitra. 2017. Robust Classification of Crisis-Related Data on Social Networks Using Convolutional Neural Networks.. In *ICWSM*. 632–635.

[21] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[22] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. 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)*, Vol. 1. 2227–2237.

[23] Anthony Rios and Ramakanth Kavuluru. 2018. 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*. 3132–3142.

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

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

[26] Johnny Torres, Carmen Vaca, and Cristina L Abad. 2017. What Ignites a Reply?: Characterizing Conversations in Microblogs. In *Proceedings of the Fourth IEEE/ACM International Conference on Big Data Computing, Applications and Technologies*. ACM, 149–156.

[27] István Varga, Motoki Sano, Kentaro Torisawa, Chikara Hashimoto, Kiyonori Ohtake, Takao Kawai, Jong-Hoon Oh, and Stijn De Saeger. 2013. Aid is out there: Looking for help from tweets during a large scale disaster. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1. 1619–1629.

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

[29] Sudha Verma, Sarah Vieweg, William J Corvey, Leysia Palen, James H Martin, Martha Palmer, Aaron Schram, and Kenneth Mark Anderson. 2011. Natural Language Processing to the Rescue? Extracting" Situational Awareness" Tweets During Mass Emergency.. In *ICWSM*. Barcelona, 385–392.

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