Multilingual Text Classification from Twitter during Emergencies

Multilingual Text Classification from Twitter during Emergencies

(Invited Paper)

Sara Piscitelli, Edoardo Arnaudo, Claudio Rossi
LINKS Foundation
<name>.<surname>@linksfoundation.com

***Abstract*—Social media such as Twitter are a valuable source of information due to their diffusion among citizens and to their speed in sharing data worldwide. However, it is challenging to automatically extract information from such data, given the huge amount of useless content. We propose a multilingual tool that automatically categorizes tweets according to their information content. To achieve real-time classification while supporting any language, we apply a deep learning classifier, using multilingual word embeddings. This allows our solution to be trained on one language and to apply it to any other language via zero-shot inference achieving acceptable performance loss.**

1. Introduction

In the last few decades, social networks have increasingly assumed a key role in communications and news sharing from all over the world, especially during emergency events such as earthquakes, hurricanes, wildfires, floods, extreme precipitations or temperatures. In fact, social media content can be useful to obtain a broader and more comprehensive view of ongoing emergencies, their progression and impacts, because they include communications from official agencies, press and also citizens, who often are the first witnesses and therefore can provide first hand information about the in-field situation.

Twitter is the most popular micro-blogging and news sharing platform counting an average of 500 million tweets per day [1] and 339.6 million active users [1], making it one of the main news channel during critical situations reporting time-critical information such as alerts or disrupted service. However, as pointed out in some works [2], [3], only a small portion of this large volume of data actually describes the ongoing events, making the retrieval of useful information a challenging tasks.

Another major problem affecting social media data is represented by the language diversity. While English is the most used language overall, relevant emergency management information also comes from the affected area, where the language varies greatly. This limits the use of supervised machine learning approaches for information extraction due to the lack of high-quality annotated data and training resources, which are usually available only in English.

The main contribution of this work is to propose a supervised machine learning approach able to perform information extraction and classification of emergency-related social media data able to cover any language, thus overcoming the limitations imposed by the aforementioned lack of labelled data. We exploit the availability of pre-trained multilingual word embeddings [4] [5] to filter informative tweets and classify them using a predefined taxonomy derived from the literature in order to facilitate *a posteriori* analysis and enable subsequent processing aimed at detecting events and to follow their evolution and impacts over time. Thanks to the ability of multilingual embeddings to represent words as semantic vectors, the classification model becomes almost independent from the input language. Moreover, the alignment in the same latent space provided by such embeddings allows to transfer the learned knowledge from a single training language (e.g. English) to other languages. In order to minimize the overall latency while maintaining robust classification performances, we exploit deep learning techniques based on Convolutional Neural Networks (CNN) because they allows for fast inference thanks to their high parallelization capabilities [6]. In summary, our social media classification approach can provide multiple advantages. First, we achieve a language-agnostic pipeline able to work seamlessly on any language. Second, we propose a multilingual text categorization method based on a deep learning technique capable of achieving low inference times and high accuracy in comparison with other natural language processing techniques. Moreover, this approach does not strictly require any training data belonging to the target language, thus achieving zero-shot inference with relatively low effort and limited degradation in terms of model accuracy. We test and validate our approach using open labelled data gathered from Twitter during real emergency events, showing that our approach overcomes state-of-art solutions based on language-dependent representations.
The remainder of this paper is structured as follows. After a brief review of the key related works In Section 2, we describe the dataset we use as well as our methodology implementation in Section 3 and Section 4, respectively. Next, we present the achieved performances in Section 5, while we draw conclusions in Section 6, outlining some directions for future works.

1. https://www.internetlivestats.com/twitter-statistics/

2. Related works

In recent years, many studies have experimented with the use of social media during emergency situations. Twitter serves this purpose extremely well, given its organization based on micro-posts, which are distributed all around the world in a matter of seconds.

Classical information extraction methods from social media entail the use of text classification techniques [7] such as Naïve Bayes classifiers (NBC) [8], [9], Support Vector Machines (SVM) [10], Random Forest or Logistic Regression [11]. These methods often required hand-engineered features such as Bag-of-word (BoW) or TF-IDF (Term Frequency – Inverse Document Frequency) [12], that may be further expanded with bi-grams or tri-grams, or character-level information. While these methods can usually serve as fast and reliable solutions, dimensionality explosion and extremely sparse representations remain recurring issues.

Previous works exploiting Twitter content during emergencies range from exploratory analyses of the information content in micro-blogs [13] [2] and the lexicon that can help maximizing the retrieval of informative data, to studies on the content volume and diversity during hazardous events [14] [15]. Many classification tasks for emergency management based on Twitter data have been defined, the most basic of which is the distinction between noise and useful content (i.e. relevant for the current context). For instance, in [16] a data ingestion pipeline is exploited to classify tweets based on their informativeness, relatively to flood events. In [17], the informativeness of tweets is evaluated using different machine learning models, such as Support Vector Machines (SVM) and Random Forests (RF) and handcrafted features extracted from text.

Other works analyzed Twitter data to: (i) classify the posts into specific categories based on the information content [18], [19] such as affected people or disrupted services; (ii) categorize the information source [20]; (iii) distinguish whether the tweet was originated from a news channel or a private user; (iv) rate the user credibility [21] giving a confidence score about the authenticity of each post. Common approaches for such tasks include standard machine learning models, which are trained on manually-defined features and Bag-of-Words representations [9]. More recently, deep learning techniques have been largely applied on social media analysis. For instance, in [19] a CNN is exploited in combination with word2vec embeddings and domain-specific vectors, highlighting how neural networks achieve better results than standard models.

However, despite the large number and variety of machine learning approaches exploiting Twitter content, none of them is able to perform well on a multi class classification tasks fully considering the multilingual nature of Twitter data. Moreover, previous results remain strongly limited by the lack of annotated data in many languages [16]. In this paper, we propose an information type classification technique that can be applied to Twitter data in multiple languages, taking advantage of multilingual word embeddings and deep learning classification. Exploiting the vector alignment of multilingual embeddings, we can define a language-agnostic pipeline that can be trained with any initial labeled data set, and then applied to Tweets written in different languages, thus achieving zero-shot inference. This allows us to train the algorithm using a limited amount of labeled data, which is often scarce and available in very few languages (e.g. English), and apply it at scale.

3. Data Preparation

For our supervised machine learning model we merged together and harmonized data from three different sources: CrisisLex [20], CrisisNLP [22] and I-REACT platform [23]. The CrisisLex repository, created in 2014, provides a set of fully-fledged datasets encompassing many different events during the last decade. The largest one, CrisisLexT26, includes a total of 26 different crises that took place from 2012 to 2013. Despite the large number of messages, only around 28,000 tweets were labelled via crowdsourcing using different criteria, including the information content. The latter includes the following categories: *affected individuals*, *infrastructure and utilities*, *donations and volunteering*, *sympathy and support*, *other useful information* and *not applicable*. CrisisNLP represents instead a collection of smaller datasets. These resources include data from 2011 to 2017 and comprise a large number of annotated documents, divided by hazardous event, with information about past emergencies such as earthquakes (e.g. Pakistan 2013, Chile and California 2014, Nepal 2015, Mexico 2017), hurricanes and typhoons (e.g. Sandy 2012, Joplin 2011), volcano eruptions (e.g. eruption in Iceland in 2014), floods and landslides (e.g. Pakistan and India 2014), pandemics (Ebola and MERS, 2014), vehicle accidents (e.g. Flight MH370, Malaysia 2014) and wildfires (e.g. California 2017). This data includes an information content taxonomy, comparable to the previous one. Specifically, it includes the following categories: *affected individuals*, *damage to infrastructure and utilities*, *injured or dead people*, *missing or found people*, *volunteering or donation effort*, *vehicle damage*, *other relevant information*, together with a specific category to identify irrelevant or not applicable content. Lastly, we included an additional set of annotations derived from the I-REACT platform (Improving Resilience to Emergencies through Advanced Cyber Technologies) [23], where a large sample of Italian tweets was gathered and annotated via crowdsourcing using the same information content taxonomy of CrisisLex, integrating more than 2200 additional tweets in the Italian language.

Because of the various sources, we carried out an extensive aggregation procedure over the datasets, merging common categories and eventually discarding redundant or needless information. As shown in Table 1, we aggregated the under-represented categories related to people into a single and more generic class named affected people; lastly, documents classified as sympathy and support were treated as not relevant due to their lack of actually useful content. We also discarded irrelevant data, focusing our attention on

Original label	Aggregations	Description	Unique counts
People injured or dead	People injured or dead, disease deaths	Information about casualties or injured	4501
People missing or found	People missing or found	Information about missing and found people	340
People affected	People affected, people evacuated or displaced, affected by the disease, help request	Information about people affected in other ways	5416
Infrastructures and utilities	Infrastructure and utility damage, vehicle damage	Information about damaged buildings, roads, services	4678
Caution and advice	Caution and advice, disease symptoms, transmission, prevention and treatment	Information about caution and advices from authorities	5320
Donation and volunteering	Rescue, volunteering or donation efforts, donations and volunteering	Information about donations, rescue and volunteering efforts	9635
Sympathy	Sympathy and support	Support, thoughts and prayers	7802
Other information	Other useful information, other relevant information	Information that does not fit into other categories	20648
Not relevant	Not relevant or can't judge, not applicable	Not relevant to the current crisis or simply useless	14470
		Total unique labelled tweets = 72810	

Table 1. Aggregation procedure, with description and frequency counts for each group.

Label	Description	Unique counts
People affected	Information about people affected, injured, found or missing	10257
Infrastructures and utilities	Information about damaged buildings, roads, services	4678
Caution and advice	Information about caution and advices from authorities	5320
Donation and volunteering	Information about donations, rescue and volunteering efforts	9635
Other information	Information that does not fit into other categories	20648
	Total amount of unique labelled tweets = 50538	

Language	Frequency
English	45105
Spanish	3898
Italian	1535
Other	1421

Table 2. Final dataset employed and language distribution.

the content classification of the useful categories. The resulting dataset contains more than 50000 tweets, distributed on five different classes. Despite the aggregation, a noticeable unbalance still remains as shown in Table 2, due to a majority of tweet belonging to the categories *other information* and *not relevant*. In terms of language distribution, around 90% of the aggregated dataset is composed of English tweets, with relevant frequencies recorded by Italian and Spanish documents. For this reason, we focused our tests on these two languages, using for test purposes, with a zero-shot approach, Spanish and Italian languages which are the with most annotated data after English

4. Methodology & Implementation

The main objective of this work is to obtain a machine learning model able to provide: (i) multilingual capabilities, (ii) good inference speed, and (iii) robust classification performances. The first feature can be directly obtained in the initial preprocessing step. Initially, we conduct a cleaning phase on the tweets, eliminating as much noisy data as possible at its source. We eliminate any unwanted symbols, such as mentions, emoticons or URLs, given their low significance in the context of text classification, then a simple rule-based tokenizer is applied, splitting words based on spaces and punctuation. Next, the resulting tokens are directly mapped to a word vector. This is typically achieved using word embeddings, dense vectors that represent each word capturing its semantic meaning, such as Word2vec [24], GloVe [25] or FastText [5]. However, word embeddings are strongly dependent on the language since they are usually generated from monolingual corpora, resulting in a specific multidimensional distribution based on the underlying documents. Various solutions have been proposed to solve this issue, ranging from transformations in latent space with Singular Value Decomposition (SVD) [26] or Canonical Correlation Analysis (CCA) [27], to fully-fledged trained multilingual embeddings such as ConceptNet [28] or MUSE [4]. While the former displays a stronger alignment among vectors, we opted for the latter because of the higher number of supported languages.

Given the shared embedding dimensions, this achieves a flexible, language-agnostic system by training a single model on the available data composed of English tweets, that can be directly applied later on testing data whose language was never seen before thus achieving zero-shot inference. The expectation is to obtain a degraded, albeit reasonable, performance even on those samples thanks to the similarity of word distributions in the embedding space.

Because of the large yet limited vocabulary size, it is likely that some tokens do not have the corresponding vector, such as numerical and other unknown words (e.g. uncommon or misspelled words). In the case of numerical words, we averaged all available numerical vectors thus forming a shared embedding able to cover such tokens in every language, while for unknown words we generated a generic random embedding, again shared by every language and employed only when every other lookup failed. Lastly, we also included a standard padding vector of zeros required to fill the missing tokens in shorter sentences, standardising the batch size at each iteration.

Once the resulting tokens have been converted into a standard representation, we define a multi-class task according to the taxonomy described in Table 2, associating each input data (tweet) to a single label corresponding to the most likely information content it contains.

The input obtained after the dynamic vector lookup by language comprises a 2D matrix of dimensions $B \times N \times D$, where D indicates the size of the pre-trained embeddings (300 in our specific case), while B identifies the batch dimension and N represents the maximum sequence length, whose value depends on the longest document in the batch. While there are physical limits on such length, imposed for instance by the maximum value of 280 tweet characters, the model supports variable sequence lengths through adaptive pooling, where the arbitrary sequence is constrained into a fixed-length vector of relevant activations, regardless of the input length. Such matrix is provided as input to a Convolutional Neural Network (CNN), as described in [29], using a variable window length. Considering deep learning approaches, CNNs provide in fact better inference timings thanks to their ability to scale in parallel, with very little loss on performance compared to recurrent solutions [6].

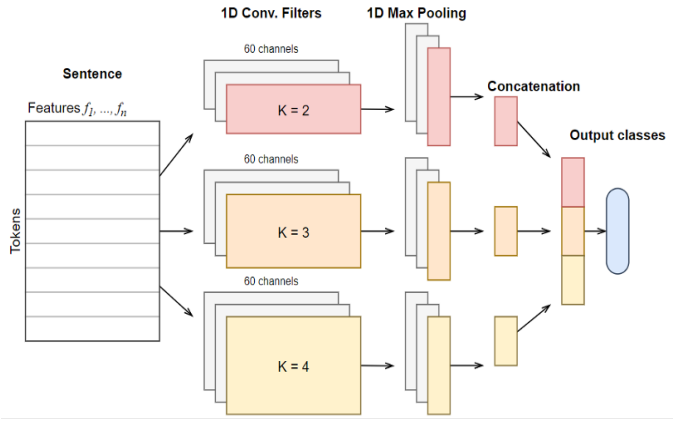

Figure 1. Tweet pre-processing and CNN model schema

We apply the same compact structure, namely 60 filters with kernel size 2 corresponding to bi-grams, 60 filters of size 3 observing the tri-grams and lastly 60 filters of size 4 for a wider context. Following the convolution, a max-pooling operation with windows size equal to 60 extracts the most representative features, moreover forcing the variable-length input into an output sequence of fixed length. Repeating the same parallel operation in every case, all three outputs are concatenated into a single 180-dimensional vector. The latter is then used as input for the classification layer, which consists of a fully connected layer followed by a standard Softmax activation.

5. Model performance evaluation

Concerning the CNN model, we encoded the cleaned and tokenized text using the MUSE embeddings, choosing the correct set at runtime based on the tweet language. We adopted an Adam optimizer with a learning rate set to $\lambda = 10^{-3}$, a dropout probability set to $p = 0.5$, a batch size equal to 64 and a Cross-Entropy Loss with smoothed class weights inversely proportional to the label frequencies, to maximize the recall. We iterated for 30 epochs on the English training corpus, subsequently evaluating the performance on the multilingual test set. To compare the performance of our multilingual solution based on the CNN model, we introduced a state-of-the art baseline, composed by a linear SVM with Bag-of-Words text encoding. In this case, we eliminated the stop-words and considered only those tokens with a document frequency greater than 10 to limit the resulting dimensionality and obtain cleaner representations. Furthermore, in order to assess the transfer learning capabilities of such baseline we resorted to word-to-word translation from English to other two target languages, namely Italian and Spanish, generating a bilingual dictionary using Microsoft's Cognitive Tools [2]. Therefore, using such look-up dictionaries, we were able to evaluate the model trained on the English tweets also on the other target languages, obtaining zero-shot inference.

2. https://azure.microsoft.com/en-gb/services/cognitive-services/

We set up two different sets of experiments. First, we performed a training and testing iteration using both English-only data and the whole set, evaluating first the zero-shot performance on the target languages (i.e., Italian, Spanish), then the improvement in terms of score when mixing together multilingual samples. Specifically, in the first experiment we randomly split the dataset in train, validation and test set as follows: 70% of the tweets for each class in the train set, 15% of the tweets for each label in the validation set and lastly 15% of the tweets of each class were included in the test set. For the first test, all Italian and Spanish tweets were included in the test set with the aim of assessing the transfer learning capabilities of the two solutions, without specific training data. In the second iteration, both foreign languages were also split using the same criteria, merging the splits with the English counterparts. Because of the noticeable class unbalance and relatively low number of documents, we trained the models using this subdivision, maintaining a significant test set to avoid excessive biases towards specific labels. We applied the validation data as an additional verification step to tune the hyperparameters and to determine when to stop the learning process. The overall results of the first experiment are shown in Table 3, from which it can be noted that the scores obtained from the baseline are much more degraded compared to the CNN, with the Italian language presenting the weakest results. Specifically, the CNN achieves a boost in the F1 micro-averaged of %5, %13, and %19 over the English, Italian and Spanish languages respectively. This demonstrates the added value given by the aligned vector space of the multilingual embeddings, which considerably enhance the transfer learning capabilities of our approach in comparison with the tested baseline. We also report the performance in terms of precision, recall and F1 score for each category in Table 4, where it can be observed that the CNN scores are more balanced and robust than the baseline in most cases. Comparing the zero-shot case in Table 3 (denoted with *English only*) with the standard mixed approach (named *Full dataset*), the monolingual solution suffers from a slight degradation in terms of performance, especially noticeable on the Italian language, with respect to the multilingual training counterpart. In the latter case however, the scores remain similar to the former, with again the exception of the Italian test set, suggesting a stronger alignment between English and Spanish.

In the second set of experiments, we adopted a leave-one-out approach carrying out a cross validation by event to verify the performances on different disaster types with different languages, contexts and label distributions. In this experiment we train both models (SVM and CNN) with the complete set of tweets in the three tested languages (i.e., English, Spanish, Italian) except those related to a specific event. Specifically, we chose to iterate on the CrisisLex T26 set, comprising 26 emergencies belonging to different hazard types, evaluating the performances for all categories having at least 10 labelled documents. The results of this evaluation are shown in Table 5, from which it is noticeable that the performance changes depending on

	BoW			CNN (English only)			CNN (Full dataset)		
F1	English	Italian	Spanish	English	Italian	Spanish	English	Italian	Spanish
micro-avg	0.72	0.39	0.51	0.77	0.52	0.70	0.74	0.61	0.71

Table 3. F1 micro-averaged for the BoW baseline and the CNN counterparts in zero-shot and multilingual scenarios.

		BoW			CNN (Zero-shot)		
Language	Categories	Precision	Recall	F1-score	Precision	Recall	F1-score
English	Caution and advice	0.71	0.46	0.56	0.61	0.83	0.70
	Donation and volunteering	0.80	0.83	0.82	0.80	0.89	0.84
	Infrastructures and utilities	0.64	0.44	0.52	0.60	0.83	0.70
	Other information	0.69	0.86	0.76	0.87	0.65	0.74
	People affected	0.76	0.59	0.66	0.77	0.78	0.78
Italian	Caution and advice	0.69	0.12	0.20	0.74	0.09	0.17
	Donation and volunteering	0.70	0.29	0.41	0.74	0.47	0.58
	Infrastructures and utilities	0.63	0.19	0.29	0.65	0.56	0.60
	Other information	0.35	0.84	0.49	0.41	0.83	0.55
	People affected	0.30	0.29	0.29	0.67	0.59	0.63
Spanish	Caution and advice	0.57	0.45	0.51	0.79	0.41	0.54
	Donation and volunteering	0.75	0.57	0.65	0.72	0.86	0.78
	Infrastructures and utilities	0.51	0.06	0.11	0.42	0.58	0.48
	Other information	0.46	0.93	0.61	0.73	0.72	0.72
	People affected	0.53	0.04	0.08	0.76	0.62	0.68

Table 4. Zero-shot results for the Bag-of-Words approach and the zero-shot CNN model.

	caution advice		donation volunteer		infrastructure utilities		other info		people affected		F1 micro-avg	
	CNN	BoW	CNN	Bow	CNN	BoW	CNN	BoW	CNN	BoW	CNN	BoW
Alberta Floods	0,44	0,22	0,87	0,82	0,58	0,42	0,54	0,54	0,51	0,52	0,63	0,54
Australia wildfires	0,58	0,17	0,63	0,52	0,47	0,38	0,68	0,67	0,27	0,07	0,59	0,44
Bohol earthquake	0,34	0,23	0,84	0,91	0,69	0,55	0,69	0,75	0,09	0,33	0,48	0,55
Boston Bombings	0,33	0,00	0,67	0,56	0,21	0,00	0,80	0,82	0,41	0,37	0,69	0,68
Brazil nightclub fire	-	-	0,57	0,40	-	-	0,71	0,74	0,59	0,11	0,62	0,28
Colorado Floods	0,49	0,10	0,81	0,80	0,66	0,41	0,63	0,59	0,44	0,32	0,61	0,48
Colorado wildfires	0,52	0,28	0,79	0,76	0,39	0,62	0,67	0,83	0,61	0,61	0,64	0,73
Costa Rica earthquake	0,83	0,73	0,73	0,22	0,66	0,48	0,82	0,78	0,05	0,08	0,76	0,69
Glasgow helicopter crash	0,36	0,00	0,60	0,54	0,49	0,06	0,63	0,70	0,65	0,47	0,61	0,51
Guatemala earthquake	0,39	0,43	0,75	0,61	0,56	0,21	0,83	0,81	0,08	0,23	0,51	0,53
Italy earthquakes	0,34	0,01	0,61	0,57	0,51	0,33	0,71	0,69	0,44	0,28	0,57	0,45
LA Airport Shootings	0,16	0,13	-	-	0,42	0,23	0,57	0,57	0,23	0,17	0,37	0,31
Lac-Megantic train crash	0,29	0,00	0,82	0,75	0,37	0,31	0,80	0,85	0,82	0,81	0,76	0,77
Manila Floods	0,69	0,30	0,83	0,83	0,40	0,39	0,32	0,34	0,36	0,33	0,67	0,52
NYC train crash	0,20	0,00	0,29	0,20	0,29	0,08	0,87	0,85	0,86	0,79	0,81	0,75
Philipinnes Floods	0,66	0,54	0,73	0,74	0,18	0,12	0,32	0,45	0,29	0,26	0,53	0,51
Queensland Floods	0,54	0,37	0,61	0,48	0,55	0,43	0,52	0,58	0,29	0,26	0,50	0,44
Russian meteor	0,40	0,17	0,40	0,00	0,34	0,29	0,91	0,87	0,51	0,53	0,76	0,73
Sardinia Floods	0,49	0,34	0,64	0,68	0,42	0,38	0,56	0,65	0,63	0,42	0,56	0,53
Savar building collapse	-	-	0,25	0,25	0,13	0,10	0,42	0,64	0,04	0,20	0,18	0,36
Singapore Haze	0,24	0,16	0,22	0,30	0,21	0,21	0,64	0,68	-	-	0,41	0,41
Spain train crash	-	-	0,83	0,79	-	-	0,81	0,89	0,60	0,77	0,72	0,83
Typhoon Pablo	0,63	0,26	0,93	0,91	0,57	0,41	0,37	0,51	0,58	0,66	0,62	0,53
Typhoon Yolanda	0,51	0,10	0,89	0,88	0,51	0,35	0,59	0,55	0,44	0,63	0,71	0,67
Venezuela refinery explosion	-	-	0,51	0,61	0,49	0,04	0,65	0,76	0,53	0,50	0,58	0,57
West Texas Explosion	0,38	0,13	0,75	0,79	0,41	0,34	0,64	0,78	0,48	0,56	0,56	0,64

Table 5. Results of leave one-out test, comparing BoW and zero-shot CNN approaches.

the event and on the category considered. This is partially due to the amount of labelled tweets for each event and category, which expresses a very high variability. Out of 26 events, the CNN model overcomes the baseline in 19 cases, improving the micro-averaged F1 score by 8% on average. Conversely, in the other 7 events, the CNN models decreases the performances again by 8%, on average. Although the aforementioned results highlight the prevalence of the CNN of the model, it should be noted that some events on specific categories perform rather poorly due to the peculiarity of the event in terms of affected area and temporal evolution and to the scarcity of tweets in some events and categories, such as the *Bohol earthquake* that includes only 366 tweets.

Lastly, we performed a simple inference test, assessing the speed of our devised solution when compared to an equivalent deep learning approach, namely a Bi-LSTM model composed of one hidden layer with 100 neurons. The test was carried out on a Intel Core i9-7940X CPU, computing the average number of tweets t processed per second. Through empirical evaluation, we obtained an average of $4500\ t/s$ for the CNN, against $200\ t/s$ for the Bi-LSTM model, confirming the efficiency of the former solution.

6. Conclusions and future works

Our work has proposed and validated an automatic multilingual tool able to categorize tweets by information content, allowing for a quick retrieval of useful data from

Twitter during emergency situations. This is achieved using multilingual word embeddings and a Convolutional Neural Network, which we trained with annotated data collected during past emergency situations. We exploited aligned multilingual word embeddings to overcome the need for specific training data on multiple languages, training our model using a single language (i.e., English) and achieving acceptable performances via zero-shot learning on a different set of languages (i.e., Spanish, Italian), which were never seen by the model. Future works will focus on improving the performances by minimizing the gap between the languages used in the training phases and the new ones. A possible solution, that might be applicable while maintaining the zero-shot approach, can be identified with recent state-of-the-art techniques based on the Transformer architecture, such as multilingual BERT [30].

7. Acknowledgements

The research leading to these results has received funding from the European Union Horizon 2020 Research and Innovation Programme under the following grant agreements: No 810812 (FASTER), No 821282 (SHELTER), and No 869353 (SAFERS).

References

[1] Twitter Inc. Q1 2019 - selected company metrics and financials. https://s22.q4cdn.com/826641620/files/doc_financials/2019/q1/Q1-2019-Selected-Company-Metrics-and-Financials.pdf, 2019.

[2] Farhad Laylavi, Abbas Rajabifard, and Mohsen Kalantari. Event relatedness assessment of twitter messages for emergency response, 2017.

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

[4] Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. *CoRR*, abs/1710.04087, 2017.

[5] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *CoRR*, abs/1607.04606, 2016.

[6] Wenpeng Yin, Katharina Kann, Mo Yu, and Hinrich Schütze. Comparative study of CNN and RNN for natural language processing. *CoRR*, abs/1702.01923, 2017.

[7] Kamran Kowsari, Kiana Jafari Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura E. Barnes, and Donald E. Brown. Text classification algorithms: A survey. *CoRR*, abs/1904.08067, 2019.

[8] Shuo Xu. Bayesian naïve bayes classifiers to text classification. 2016.

[9] Marc-André Kaufhold, Markus Bayer, and Christian Reuter. Rapid relevance classification of social media posts in disasters and emergencies: A system and evaluation featuring active, incremental and online learning. *Information Processing Management*, 57, 2020.

[10] S. M. H. Dadgar, M. S. Araghi, and M. M. Farahani. A novel text mining approach based on tf-idf and support vector machine for news classification. In *2016 IEEE International Conference on Engineering and Technology (ICETECH)*, pages 112–116, 2016.

[11] Tomas Pranckevicius and V. Marcinkevicius. Comparison of naive bayes, random forest, decision tree, support vector machines, and logistic regression classifiers for text reviews classification. *Balt. J. Mod. Comput.*, 5, 2017.

[12] K. Sparck-jones. A statistical interpretation of term specificity and its application in retrieval. *Journal on Documentation*, 28(1):11–21, 1972.

[13] Alexandra Olteanu, Sarah Vieweg, and Carlos Castillo. What to expect when the unexpected happens: Social media communications across crises. *CSCW '15: Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, page 994–1009, 2015.

[14] Nattiya Kanhabua and Wolfgang Nejdl. Understanding the diversity of tweets in the time of outbreaks. *the 22nd International Conference*, 2013.

[15] Rui Long, Haofen Wang, Yuqiang Chen, Ou Jin, and Yong Yu. Towards effective event detection, tracking and summarization on microblog data. In Haixun Wang, Shijun Li, Satoshi Oyama, Xiaohua Hu, and Tieyun Qian, editors, *Web-Age Information Management*, pages 652–663, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg.

[16] C. Rossi, F.S. Acerbob, K. Ylinenc, I. Jugac, P. Nurmic, A. Boscad, F. Tarasconid, M. Cristoforettie, and A. Alikadic. Early detection and information extraction for weather-induced floods using social media streams. *International journal of disaster risk reduction*, 2018.

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

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

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

[20] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. Crisislex: A lexicon for collecting and filtering microblogged communications in crises. *Int. Conf. on Weblogs and Social Media (ICWSM)*, 2014.

[21] Marcelo Mendoza, Barbara Poblete, and Carlos Castillo. Information credibility on twitter. *Conference: Proceedings of the 20th international conference on World wide web*, 2011.

[22] Muhammad Imran, Carlos Castillo, and Prasenjit Mitra. Twitter as a lifeline: Human-annotated twitter corpora for nlp of crisis-related messages. *In Proceedings of the 10th Language Resources and Evaluation Conference (LREC)*, 2016.

[23] European Union. Horizon 2020 research programme, improving resilience to emergencies through advanced cyber technologies, european union, 2016.

[24] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In Yoshua Bengio and Yann LeCun, editors, *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, 2013.

[25] 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, Doha, Qatar, October 2014. Association for Computational Linguistics.

[26] Samuel L. Smith, David H. P. Turban, Steven Hamblin, and Nils Y. Hammerla. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. *CoRR*, abs/1702.03859, 2017.

[27] Manaal Faruqui and Chris Dyer. Improving vector space word representations using multilingual correlation. 2014.

[28] Robyn Speer, Joshua Chin, and Catherine Havasi. ConceptNet 5.5: An open multilingual graph of general knowledge. pages 4444–4451, 2017.

[29] Yoon Kim. Convolutional neural networks for sentence classification. *CoRR*, abs/1408.5882, 2014.

[30] Telmo Pires, Eva Schlinger, and Dan Garrette. How multilingual is multilingual BERT? July 2019.