Understanding Fine-tuned BERT Models for Flood Location Extraction on Twitter Data

Jane Arleth dela Cruz*, Iris Hendrickx and Martha Larson

[1]*Centre for Language Studies, Centre for Language and Speech Technology, Radboud University, Netherlands*

Abstract

We address the two subtasks of the Disaster-Related Social Media Data (DisasterMM) Task at MediaEval 2022: flood relevance classification and location extraction for Italian tweets. We demonstrate the ability of BERT-based deep learning models to address these subtasks. We then carry out error analysis, with an emphasis on the location extraction task, and find that our models (1) perform well with limited training data, (2) are robust in handling text with punctuation and special characters, and (3) are capable of detecting phrasal locations that can be expressed in different ways and can be very challenging for conventional gazetteer-based approaches.

1. Introduction

The Multimedia Analysis of Disaster-Related Social Media Data (DisasterMM) at the 2022 Multimedia Evaluation (MediaEval) benchmark [1] asks researchers to develop artificial intelligence algorithms that can derive information from Italian Tweets that is relevant to flooding disasters.

The analysis of social media posts, and especially short posts, like those on Twitter, to support flood disaster management faces two basic difficulties, each addressed by a subtask of DisasterMM. First, even though posts contain words related to flooding, they might not actually be relevant to flood disaster. This challenge is addressed by the Relevance Classification of Twitter Posts (RCTP) subtask. Second, many posts do not contain geo-location information, which is important for disaster managers to determine where flood related events are happening. This challenge is addressed by the Location Extraction from Twitter Texts (LETT) subtask.

In this paper, we describe an approach using pre-trained multilingual BERT-base machine learning models to address both subtasks. These models are found (1) to perform well on text with limited training data, (2) to be robust against text with punctuation and special characters, and (3) to infer more specific phrase locations compared to traditional rule-based approaches. We also carry out an error analysis of output of our location extraction pipeline in order to identify patterns in the data that make location extraction difficult and formulate ways in which the task and approaches addressing the task can be improved in the future.

2. Related Work

With the advent of pre-trained language models such as BERT, once a sufficient amount of training data is available for specific NLP tasks, fine-tuning is often employed to address these tasks successfully [2]. Furthermore, irrelevant changes in punctuation are correctly ignored

MediaEval'22: Multimedia Evaluation Workshop, January 13–15, 2023, Bergen, Norway and Online

*Corresponding author.

janearleth.delacruz@ru.nl (J. A. dela Cruz); iris.hendrickx@ru.nl (I. Hendrickx); martha.larson@ru.nl (M. Larson)

by recent transformer models such as BERT while older RNN-based models were sensitive to them, making BERT attractive to use on tweet text data that is typically written in an informal and noisy style [3]. Zahera [4] employed a fine-tuned BERT model first to filter relevant tweets during disasters and then to assign these tweets to relevant information types. Also in the MediaEval task 2020 [5] on labeling Italian tweets related to floods, a pre-trained Italian BERT model was used as tokenizer [6] and one was used as their classifier [7] where using the classifier yielded the better performance so we follow this approach.

3. Approach

3.1. Relevance Classification of Twitter Posts

For RCTP, we employed different preprocessing techniques on the RCTP development dataset like stopword removal and stemming. However, best results were obtained with 'Clean text' consisting of lowercased tweet data, removing of whitespace, usernames, URL's and retweets. Upon removing duplicate tweets and retweets, 2,524 tweets remained, which were then split in 80%-20% training-validation via stratified sampling. We ran Logistic Regression, Support Vector Machine (SVM) and Naive Bayes (NB) algorithms with default parameters on the different preprocessed text data.

For the training of our BERT models, we used the Transformer Toolkit [8]. We used the BERT-base Multilingual cased model and BERT-based Italian cased model. The BERT-base Multilingual cased model contains 104 languages, 12-layer, 768-hidden, 12-heads, 110M parameters developed by Google Research [2]. The source data for the BERT-base Italian model consists of a recent Wikipedia dump and various texts from the OPUS corpora collection [9].

3.2. Location Extraction from Twitter Texts

As the automatic recognition of locations is one of the subtasks of Named Entity Recognition (NER), it is an obvious choice to evaluate the effectiveness of existing NER tools for this specific task of recognizing Italian locations mentioned in flood-related tweets. We did not employ any text preprocessing techniques on the LETT development dataset. We did a train-validation stratified sampling split of 80%-20%, resulting in 3,146 tweets as training set and 787 tweets as validation set. We fine-tuned 4 different BERT-base models: BERT-base Multilingual Cased, BERT-base Italian Cased and pretrained NER models called BERT-base NER and WikiNEuRal Mulitilingual NER, all using the same 3,146 tweets train set. We selected the cased models to better recognize locations that are proper nouns such as cities, towns, provinces, etc. BERT-base NER is ready to use for NER, trained to recognize four types of entities and fine-tuned on the English version of the standard CoNLL-2003 NER dataset [10] [11]. WikiNEuRal Multilingual NER is a BERT model fine-tuned on the recently-introduced WikiNEuRal dataset for Multilingual NER, it supports 9 languages, including Italian [12].

Table 1
Evaluation of RCTP Classifiers (F1-Score)

Text Input	Classifier	TF	TFIDF	BERT Tokenizer
Clean Text	Logistic Regression	0.902	0.911	
	SVM	0.901	0.902	
	Naive Bayes	0.882	0.895	
	BERT Multilingual Cased			0.922
	BERT Italian Cased			0.914

4. Experimental Results

Table 1 shows the average F1-score of the different classifiers on a 5-fold cross validation experiment on the development set of the RCTP task. We chose to implement the BERT-base Multilingual cased model with the clean text preprocessing on our RCTP Run 1 submission as it had the best performance amongst all.

Table 2 shows the models' results on the given test set in the different submitted runs. Two scores are measured each run at word level: the exact F1-score, where labels have to fully match, and the partial F1-score, where either "B-LOC" or "I-LOC" are considered as true. The big drop in the RCTP score on the test set compared to validation score can be attributed to missing data as we just missed out to classify 169 out of the 1,469 test tweets.

For the LETT task, we submitted the best classifier from our cross validation experiments on the development set as Run 1. For Runs 2 and 3, we used external data ,the Italian WikiNEuRal Dataset only to fine-tune our models. For our Run 2, we trained the BERT Multilingual Cased model while Run 3 was the BERT Italian Cased model. Both were fine-tuned with Italian WikiNEuRal and then with our training set of 3,146 tweets.

We see that for LETT runs, there was still drop in performance when compared to our models' performance on our validation set, with the WikiNEuRal Multilingual NER performing best at 0.835 F1-score. We consider this a good indicator that our models were able to perform well with the limited training data.

Table 2
LETT and RCTP Runs Results Development Set vs. Test Set (F1-Score)

Subtask	Run	Method	Val Set		Test Set	
RCTP	Run 1	BERT Multilingual Cased on Clean Text	0.922		0.742	
			Exact	**Partial**	**Exact**	**Partial**
LETT	Run 1	WikiNEuRal Mulitilingual NER	0.917	0.950	0.835	0.846
	Run 2	BERT Multilingual Cased	0.922	0.912	0.830	0.843
	Run 3	BERT Italian Cased	0.925	0.891	0.817	0.834

5. Error analysis on validation set

On Punctuations and Special Characters

We observed that all our BERT-base models are very robust against punctuations and special characters (periods, commas, hashtags, colons, ellipses, and others), which is useful as the tweet text format is unstructured. Although the BERT-base Italian Cased model had the lowest error rate at 0.038 when considering all special characters, it has a high error rate on other special characters (- ! " ()) at 0.142. Hence, we select the WikiNEuRal Multilingual NER having an error rate of 0.04 as the better choice when we want to catch all locations with special characters.

Learning from limited training

We can confirm that our models were able to generalize street names when we investigated into the training data. Not only can our models capture street name tokens such as "#SS340", "A51", and "SS36", it was also able to predict these streets as "B-LOC" of phrasal locations. Here we show an example:

(1)	*SS36*	*Del*	*Lago*	*Di*	*Como*	*E*	*Dello*	*Spluga*
	B-LOC	I-LOC	I-LOC	I-LOC	I-LOC	I-LOC	I-LOC	I-LOC

However, there is still room for improvement. "Faro#Fiumicino," was annotated as "B-LOC" in the dataset, but were annotated as "O" by all the models. "Faro#Fiumicino," seems to be a flight route from Faro in Portugal to Fiumicino in Italy, making this specific token ambiguous. Our models were also not able to detect a very specific long location phrase "Tronco Maestro del Bacchiglione" which is a branch of the river Bacchiglione. Instead, all models predicted only the word "Bacchiglione" as a location.

Going beyond the gazetteer

We want to confirm whether deep learning can detect locations above and beyond what can be detected by a simple, transparent, gazetteer-based approach. Our error analysis revealed that one place in which BERT has an advantage over gazetteer-based approaches is locations that consist of multiple words ("phrasal locations"), which are highly variable and would be difficult to find in a gazetteer or include in a fixed vocabulary. Take for example "Milano". In our validation set, it appears 61% of the time as "B-LOC" and 33% as "I-LOC" (inside a phrasal location) and 6% as "O". WikiNEuRaL NER and BERT-based Multilingual predicted all of these instances correctly.

6. Outlook

We are particularly interested in considering the fit between the ground truth of the DisasterMM task and possible real-world applications of the technology for disaster management. Here, we mention two aspects of the ground truth that we discovered while carrying out the task that would be interesting to address in the future.

First, we observed some inconsistencies in the annotations. For example, we observed in the ground truth that "Toscana" appeared 69 times in the validation set, 48 annotated as "B-LOC" (location) and 21 annotated as "O" (not location). Examples of the mismatches are:

> Both occurrences of Toscana were annotated as Not Location: *Allerta meteo Toscana - 05/04/2017 12:00 - Regione Toscana - https://t.co/NwsxgzAAiK https://t.co/6Rli3ttdfs*

> Both occurrences of Toscana were annotated as Location: *Allerta meteo Toscana - 14/06/2017 02:00 - Regione Toscana - https://t.co/NwsxgzAAiK https://t.co/VrQtOuYV7J*

We infer that these were annotation errors: "Toscana" should be location. All of our models tag "Toscana" as "B-LOC" in all instances, meaning that some of the errors our model makes might not be actual errors. Other possible inconsistencies are "Bacchiglione...", "vicentino:", "Roma." and "Incrocio" (Eng. "intersection").

Second, we observe that there are some location-related entities in the data set that are not annotated as locations, but would be important in the context of flood disaster management. Specifically, flooding events will often refer to rivers, and valleys crucial locations, but do not appear to be included in the definition of the location used to annotate the data. These events do not have a single location in terms of latitude and longitude. However, there is no principled reason for which they cannot be considered locations, since roads are annotated as locations in the dataset. The conventional datasets used to train NER tools are based on standard news articles that usually refer to cities and country names. Future work should focus on designing annotation protocols with experts that annotate all location- related entities that could possibly be relevant for flooding. Looking towards real world applications, we intend to improve our methodology to better detect specific phrasal locations for decision support in disaster management.

Acknowledgments This work is partly financed by the Dutch Research Council (NWO) with project number NWA.1292.19.399.

References

[1] S. Andreadis, A. Bozas, I. Gialampoukidis, A. Moumtzidou, R. Fiorin, F. Lombardo, T. Mavropoulos, D. Norbiato, S. Vrochidis, M. Ferri, I. Kompatsiaris, DisasterMM: Multimedia Analysis of Disaster-Related Social Media Data Task at MediaEval 2022, in: Proceedings of the MediaEval 2022 Workshop, Bergen, Norway and Online, 2023.

[2] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, CoRR abs/1810.04805 (2018). URL: http://arxiv.org/abs/1810.04805. arXiv:1810.04805.

[3] A. Ek, J.-P. Bernardy, S. Chatzikyriakidis, How does punctuation affect neural models in natural language inference, in: Proceedings of the Probability and Meaning Conference (PaM 2020), Association for Computational Linguistics, Gothenburg, 2020, pp. 109–116. URL: https://aclanthology.org/2020.pam-1.15.

[4] H. M. Zahera, Fine-tuned bert model for multi-label tweets classification, in: Text Retrieval Conference, 2019.

[5] S. Hicks, D. Jha, K. Pogorelov, A. G. S. D. Herrera, D. Bogdanov, P.-E. Martin, S. Andreadis, M.-S. Dao, Z. Liu, J. Vargas-Quirós, B. Kille, M. Larson (Eds.), Working Notes Proceedings of the MediaEval 2020 Workshop, CEUR Workshop Proceedings, 2020. URL: https://ceur-ws.org/Vol-2882/.

[6] N. Said, K. Ahmad, A. Gul, N. Ahmad, A. Al-Fuqaha, Floods detection in twitter text and images, in: [5], 2020. URL: https://ceur-ws.org/Vol-2882/paper34.pdf.

[7] F. Alam, Z. Hassan, K. Ahmad, A. Gul, M. A. Riegler, N. Conci, A. Al-Fuqaha, Flood detection via twitter streams using textual and visual features, in: [5], 2020. URL: https://ceur-ws.org/Vol-2882/paper35.pdf.

[8] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, A. Rush, Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 38–45. URL: https://aclanthology.org/2020.emnlp-demos.6. doi:10.18653/v1/2020.emnlp-demos.6.

[9] M. D. L. Team, bert-base-italian-cased, 2019. URL: https://huggingface.co/dbmdz/bert-base-italian-cased.

[10] D. S. Lim, 2020. URL: https://huggingface.co/dbmdz/bert-base-italian-cased.

[11] E. F. Tjong Kim Sang, F. De Meulder, Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition, in: Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003, 2003, pp. 142–147. URL: https://www.aclweb.org/anthology/W03-0419.

[12] S. Tedeschi, V. Maiorca, N. Campolungo, F. Cecconi, R. Navigli, WikiNEuRal: Combined neural and knowledge-based silver data creation for multilingual NER, in: Findings of the Association for Computational Linguistics: EMNLP 2021, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 2521–2533. URL: https://aclanthology.org/2021.findings-emnlp.215.