

How does a Pre-Trained Transformer Integrate Contextual Keywords? Application to Humanitarian Computing

Valentin Barriere*

European Commission

Joint Research Centre (JRC) - Ispra

valentin.barriere@ec.europa.eu

Guillaume Jacquet

European Commission

Joint Research Centre (JRC) - Ispra

guillaume.jacquet@ec.europa.eu

ABSTRACT

In a classification task, dealing with text snippets and metadata usually requires to deal with multimodal approaches. When those metadata are textual, it is tempting to use them intrinsically with a pre-trained transformer, in order to leverage the semantic information encoded inside the model. This paper describes how to improve a humanitarian classification task by adding the crisis event type to each tweet to be classified. Based on additional experiments of the model weights and behavior, it identifies how the proposed neural network approach is partially over-fitting the particularities of the Crisis Benchmark, to better highlight how the model is still undoubtedly learning to use and take advantage of the metadata's textual semantics.

Keywords

Transformers, contextual keywords, humanitarian computing.

INTRODUCTION

It is frequent to fuse information from different channels in order to help the classifier disambiguate some examples (Xu et al. 2013). Nonetheless, systems that mix together the different information fluxes inside the model are mainly used in multimodal data-processing (Baltrušaitis et al. 2017) and less frequent for a simple modality.

The rise of Transformer models (Vaswani et al. 2017) created a new set of possibilities in text processing. Nowadays, the new architectures are pushing the boundaries of the state-of-the-art (Raffel et al. 2019; Brown et al. 2020) in general tasks and no longer only in Machine Translation.

Those architectures allow us to efficiently process an input composed of 2 sentences, like for Natural Language Inference tasks (Williams et al. 2017), so that they can interact together and the fusion happens inside the model. This is interesting since the model is able to infer the meanings contained in both sentences in context.

In this paper, we propose the use of textual metadata semantics, by encoding it textually in the input of the classifier in a way the model can detect metadata as separate information and yet is able to use cross-input attention between the metadata and the text content. This can be seen as related to some works done on zero-shot learning using Natural Language Inference (NLI) where the semantic content of the label is encoded inside the classifier (Yin et al. 2020; Halder et al. 2020). (Clark et al. 2019) studied the attention mechanism of a BERT model and clustered the attention heads. (Rogers et al. 2020) reviewed the understanding of BERT-based transformers in the literature, highlighting that semantics information was spread across the entire model. From the best of our knowledge, no work has been done to analyze semantics between a contextual keyword and its associated sentence by retrieving the most important words with the help of the attention weights, then clustering the retrieved words embeddings.

*corresponding author

Table 1. Examples of text pre-processing for each model

Model	Example
BERT	[CLS] fire [SEP] After deadly Brazil nightclub fire, safety questions emerge. [SEP]
RoBERTa	<s>fire </s>After deadly Brazil nightclub fire, safety questions emerge. </s>
T5	cbmk context: fire sentence: After deadly Brazil nightclub fire, safety questions emerge.

In the context of a crisis caused by a natural or human source, it is important to be react quickly and to be aware of the population needs (Alam, Ofli, and Imran 2020). The resources created by social media users can contain crucial information (Olteanu et al. 2015; Imran et al. 2016) because of the poster witnessing the situation directly (Zahra et al. 2020). The type of event is highly variable, and the system needs to be relevant in the context of a natural disaster like a fire or a hurricane (Waqas and Imran 2019; Alam, Ofli, Imran, and Aupetit 2018), a train crash or a pandemic (Qazi et al. 2020). Moreover, it is known that out-of-domain prediction is harder (Alam, Joty, et al. 2018).

We chose to work on the Crisis Benchmark 14-event-types dataset (Alam, Sajjad, et al. 2021) on a 11-humanitarian-class supervised setting, in order to use the event type as metadata to enhance the quality of the predictions, comparing the results using BERT (Devlin et al. 2018), RoBERTa (Liu et al. 2019) and T5 (Raffel et al. 2019). It is important to note that in this type of task, the urge of retrieving information from social media arrives after the beginning of the disaster, hence, the type of disaster is always known and it is an information available for use. We found that in every configuration the event-aware models obtain better results than their Vanilla counterparts. We then investigated this improvement by an intensive study of the dataset, the learned model behaviors and predictions. In addition to the experiments using the official partition, we used a Leave-One-Event-Type-Out setting (LOETO) in order to prevent the model from overfitting some particularities of the dataset events. This work is directly in continuation with the work of (Alam, Sajjad, et al. 2021), in which they add the event type to the tweet content using a simple concatenation, leading to no improvement over the Vanilla model.

METHOD

The proposed approach consists of the integration of metadata information inside the input text representation. We chose a dataset where this was available and where the authors of the original dataset claimed that their method was not working. We decided to integrate the event type into the learning algorithm using a NLI configuration for 3 transformer models. After running several experiments, we observed an increase in all the models' performances, independently of the architecture. We ran some analyses in order to investigate whether this gain came from the textual content of the metadata, or from mechanical memorization of the label distribution conditioned by the metadata.

We chose to focus on the Crisis Benchmark's humanitarian classification task for space reasons. The reader should note that we also made experiments on the informativeness binary classification task described in (Alam, Sajjad, et al. 2021) that lead to positive results.

Incorporating Event Type

Instead of naively concatenating the event-type and the tweet together, we used the patterns offered by the different models to separate the two types of information. In this way, we are not breaking the syntax and the rhythm of the tweet sentence by adding a piece of text that does not belong in the initial sentence. For BERT and RoBERTa, we used the special tokens in order to separate the 2 sentences. For T5, we used a text pattern, with a new task name and the `sentence` and `context`. Examples of the different pre-processing configurations we used are visible in Table 1.

By calling y the humanitarian label, x the observed tweet and m the metadata, we can reformulate the prediction of the Vanilla model as $P(y|x)$. Instead of modeling $P(y|x)$, the event-aware configuration models a conditioned probability $P(y|x, m)$. Another way to integrate the metadata would be to model the joint probability of the class y and the metadata m in $P(y, m|x)$. We chose the conditioned model in order to easily use the semantic information contained inside the keyword, similarly to the way some zero-shot NLI models are handling the labels (Yin et al. 2020).

Model and dataset analysis

We ran some analysis of the event-aware model, in order to verify if the model was integrating semantic information regarding the textual content of the event. Transformers and more generally Neural Networks models are known to be black boxes, but some works have been done in the direction of interpreting them (Clark et al. 2019). In combination with analysis of the model and its behavior, we also analyze the dataset and its label distribution, allowing us to detect some patterns on which the neural network could rely on.

Table 2. Results on the humanitarian classification task. (Alam, Sajjad, et al. 2021) use concatenation.

Model	Event	Prec	Rec	u-F1	w-F1	Acc
BERT (Alam, Sajjad, et al. 2021)	✓	70.1	71.3	70.7	86.5	86.5
RoBERTa (Alam, Sajjad, et al. 2021)	✓	70.2	72.3	71.1	87.0	87.0
BERT	✗	73.5	71.9	72.5	87.5	87.5
	✓	75.3	72.5	73.7	88.3	88.1
RoBERTa	✗	74.2	73.6	73.7	87.9	88.0
	✓	74.1	74.5	74.1	88.5	88.5
T5	✗	75.0	74.4	74.6	88.3	88.4
	✓	76.7	73.8	75.1	88.8	88.9

We seek to answer the following questions:

- *Dataset label distribution*: What does the labels distribution look like for each event ?
- *Predicted label distribution*: What is the impact of conditioning over an event on the predictions distribution?
- *Out-of-domain learning*: Is the event-aware model still better on a Leave-One-Event-Type-Out setting?
- *Attention weights*: What words are influenced by the metadata event type token?

We ran the models with the `transformers` library (Wolf et al. 2019). For BERT and RoBERTa, we used a learning rate of 1e-6, Adam algorithm, for maximum 20 epochs, while we ran the T5 for 3 epochs, using a learning rate of 3e-4. For RoBERTa, we manually added and trained a new layer¹ allowing to specify the token types, which is normally not used during RoBERTa pre-training. On the LOETO setting, we created the dev by taking 25% of the test set, otherwise we used the official train/dev/test partition. We normalized the size of every sequence up to 128 tokens in total.

EXPERIMENTS AND ANALYSIS

Dataset

For our experiments, we used the Crisis Benchmark dataset (Alam, Sajjad, et al. 2021) composed of 87,557 tweets collected during several crisis events, that can be separated into 14 event types: bombing, collapse, crash, disease, earthquake, explosion, fire, flood, hazard, hurricane, landslide, shooting, volcano, or none. This dataset has been labeled into 11 humanitarian classes: *Affected individuals*, *Caution and advice*, *Displaced and evacuations*, *Donation and volunteering*, *Infrastructure and utilities damage*, *Injured or dead people*, *Missing and found people*, *Not humanitarian*, *Requests or needs*, *Response efforts*, *Sympathy and support*. We refer the readers to the original paper for more details on the dataset (Alam, Sajjad, et al. 2021).

Classifier results

The results of the experiments comparing the event-aware and event-unaware models are shown in Table 2. We used unweighted means of the Precision, Recall and F1, as well as global Accuracy and weighted F1. The best results among the transformers are obtained without surprise with the T5, which has 220M parameters, more than BERT and RoBERTa (110 and 125M). For each transformer, the event-aware model reaches higher performances compared to Vanilla model. The results event per event are available in Table 4, the only event that is worse in the event-aware setting is 'fire'. Finally, we believe the reason our Vanilla models obtain significantly better results than Alam, Sajjad, et al. 2021 remains in a better fine-tuning of the transformers, with a lower learning rate and a higher number of epochs.

Analysis

We run analyses on the dataset as well as BERT weights and behavior in order to interpret the good results of the event-aware model compared to the Vanilla model.

Dataset label distribution

It is interesting to look at the distribution of the labels regarding the event type, available in Figure 1 for the training set². We can see the type of labels appearing in each event type is highly imbalanced: for some events, like *Landslide* and *Volcano*, there is even only one label, which is *Not humanitarian*. These events, which are highly unrepresentative in term of labels distribution, only represent a small portion of the dataset (Figure 1).

¹Layer managing the `token_type_ids`

²The distributions are comparable on the test set

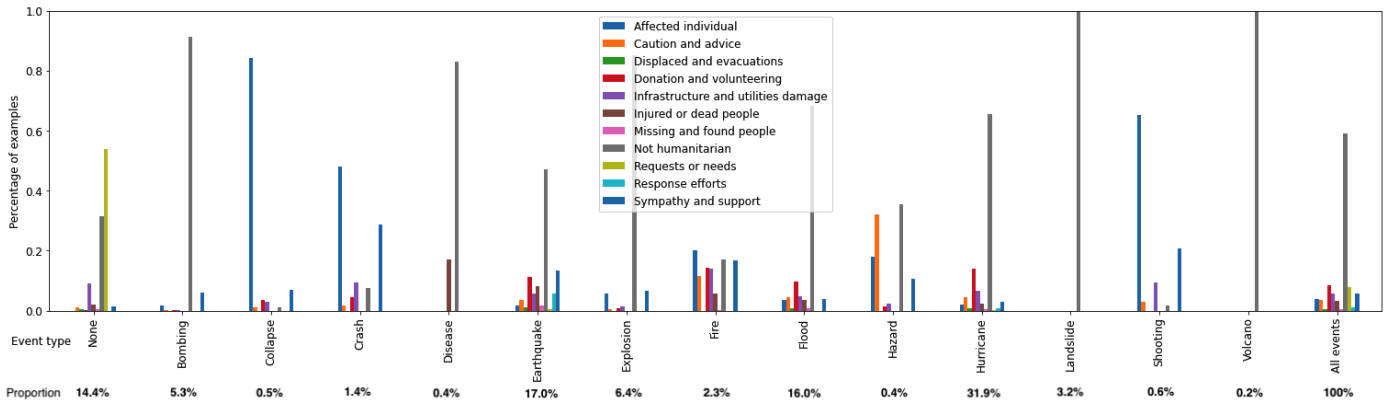


Figure 1. Distributions of labels regarding the event type in the train set, with the proportion of each event type

Table 3. Results of the BERT model on LOETO

Model type	Prec	Rec	F1	Acc
Vanilla	40.0	54.9	44.1	65.4
Event-aware	47.0	55.2	45.2	67.6

By comparing the distributions with the model’s performances, we can see that the only event type where the event-aware is worst is ‘fire’, where the distribution of labels is closer to uniform.

With such diverse distributions of labels regarding the event, it is obvious that memorizing the label distribution of each event will be highly beneficial for the model, but just an overfitting of the model on the dataset.

Predicted Label Distribution

In order to check if conditioning over an event was distorting the predictions, we calculated the Kullback-Leibler divergences between the test set distribution \mathcal{D}_t , the distributions of predictions over the test set conditioned by an event E , $\mathcal{D}_t(E)$, and the event’s label distributions $\mathcal{D}_e(E)$. We found that, when conditioned over an event, the model predicts on average over the test set, a distribution that is closer to its respective event distribution (0.62) than to the test set distribution (0.69) (Eq. 1). This seems to confirm the impact of the event type token in overfitting some dataset particularities.

$$\sum_E KL(\mathcal{D}_e(E) || \mathcal{D}_t(E)) < \sum_E KL(\mathcal{D}_t || \mathcal{D}_t(E)) \quad (1)$$

Leave One Event Type Out

We ran a LOETO cross-validation in addition to the experiments using the official partition of the dataset. This setting allows the comparison of results in a situation where the event-aware model is not able to infer the label distribution of the event type.

This method obtains necessarily worse results than when testing on samples from an event type that was used during learning, the goal is to compare the Vanilla BERT and the event-aware BERT. If the event-aware model reaches higher performances than the Vanilla, then adding textual metadata has a positive impact on the performances.

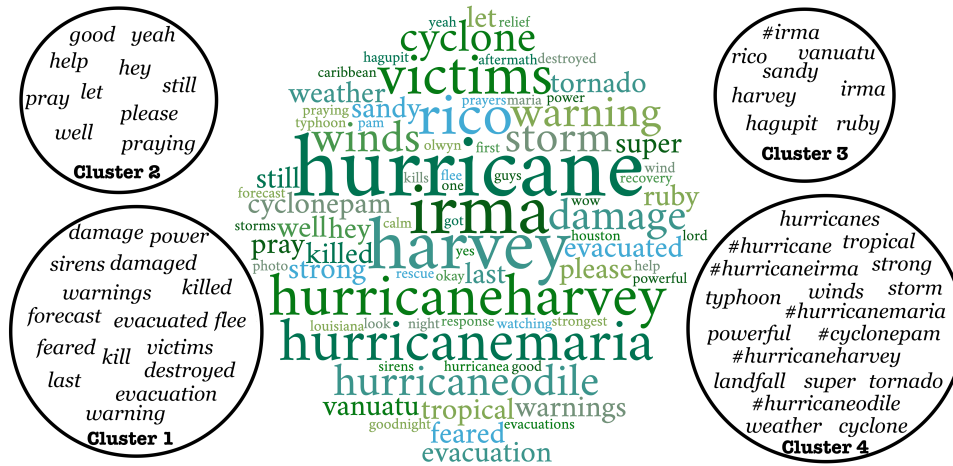
The results shown in Table 3 are explicit: adding the event type also improves the results when the system is faced with a new event it had not seen before. In Table 4 we can see on the results that the event-aware model is not homogeneously performing according to event types.

Attention weights

In order to understand how the event type is influencing the decision, we look at the interaction mechanism between the event type and the tweet by studying the attention weights of the BERT model. To avoid any model overfitting on specific words corresponding to specific events, we ran this study in a LOETO configuration, so that the model has never seen the event type token before and cannot make a correlation between this token and words appearing for this event type.

Table 4. Accuracies (differences with Vanilla) for each event type of the event-aware BERT on the humanitarian classification task, for official partition and LOETO

Partition	None	Bombing	Collapse	Crash	Disease	Earthquake	Explosion
Official	91.2 (1.2)	96.7 (0.4)	88.8 (0.0)	89.3 (1.1)	98.6 (2.9)	77.0 (1.2)	96.6 (0.3)
LOETO	34.3 (5.0)	89.7 (-4.3)	44.1 (19.7)	81.5 (-0.3)	59.4 (-11.3)	49.4 (-1.6)	93.1 (1.4)
Fire	Flood	Hazard	Hurricane	Landslide	Shooting	Volcano	Average
81.5 (-1.2)	90.7 (0.7)	52.8 (0.0)	88.0 (0.6)	100 (1.6)	87.5 (0.0)	97.1 (0.0)	88.3 (0.8)
67.6 (-4.2)	85.3 (1.7)	49.8 (1.4)	71.7 (5.0)	92.6 (-0.6)	77.8 (7.1)	72.0 (-2.8)	67.6 (2.2)

**Figure 2. Tokens interacting the most with the event type 'hurricane'. Clusters of the top-50 tokens.**

For each event, we counted the number of times every token from the tweet was linked to the event token with an attention weight bigger than an arbitrary threshold of 0.5. We discarded the weights between the punctuation and the stop-words³ tokens. Then we took the 50 tokens with the highest tf-idf, extracted their embeddings with a BERT model, and proceeded with a clustering.

A visualization of the words and clusters for *hurricane* has been made in Figure 2. We figured out that the tokens interacting directly with the event type were related to the type of disaster⁴, proper names⁵, and the classes of the task.⁶ This means that even for an event type not contained in the training set, the model is capable of using the semantics of this event in order to better infer the class of the tweet.

CONCLUSION

In this article, we studied the effect of adding textual metadata information inside three transformer models, using the classical ways to separate different sentences in an input.

We ran experiments on a Humanitarian Computing dataset, adding to each tweet its respective event type, which is contextual information always available, in order to better classify it into 11 classes. We discovered that this method improves the results, even when the event type has never been seen during training phase. It can be applied to different transformers and we obtained a new state-of-the-art on this dataset. Finally, we carried out an analysis of the dataset and the event-aware BERT model weights and behavior in order to shed light on the reasons for this increase in performance.

We show that the event-aware model is not only memorizing the unbalanced label distribution of the event type, but also learning semantics relation between the text and the event type token. We discovered that the tokens locally interacting with the event token were related to the lexical fields of the event type, and the classes of the classification task.

³from nltk toolbox

⁴*hurricane, cyclone, storm, tornado, typhoon...*

⁵*Irma, Sandy, Harvey, Vanuatu,...*

⁶*sympathy, material damages, human damages, warnings, evacuations,...*

REFERENCES

- Alam, F., Joty, S., and Imran, M. (2018). “Domain adaptation with adversarial training and graph embeddings”. In: *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)* 1, pp. 1077–1087. arXiv: [1805.05151](#).
- Alam, F., Ofli, F., and Imran, M. (2020). “Descriptive and visual summaries of disaster events using artificial intelligence techniques: case studies of Hurricanes Harvey, Irma, and Maria”. In: *Behaviour and Information Technology* 39.3, pp. 288–318.
- Alam, F., Ofli, F., Imran, M., and Aupetit, M. (2018). “A twitter tale of three hurricanes: Harvey, Irma, and Maria”. In: *Proceedings of the International ISCRAM Conference 2018-May*, pp. 553–572. arXiv: [1805.05144](#).
- Alam, F., Sajjad, H., Imran, M., and Ofli, F. (2021). “CrisisBench: Benchmarking Crisis-related Social Media Datasets for Humanitarian Information Processing”. In: *AAAI*. arXiv: [2004.06774](#).
- Baltrušaitis, T., Ahuja, C., and Morency, L.-P. (2017). “Multimodal Machine Learning: A Survey and Taxonomy”. In: pp. 1–20. arXiv: [1705.09406](#).
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). “Language Models are Few-Shot Learners”. In: arXiv: [2005.14165](#).
- Clark, K., Khandelwal, U., Levy, O., and Manning, C. D. (2019). “What does BERT look at? An analysis of BERT’s attention”. In: *arXiv*. arXiv: [1906.04341](#).
- Devlin, J., Chang, M.-w., Lee, K., and Toutanova, K. (2018). “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: arXiv: [arXiv:1810.04805v1](#).
- Halder, K., Akbik, A., Krapac, J., and Vollgraf, R. (2020). “Task-Aware Representation of Sentences for Generic Text Classification”. In: *COLING*.
- Imran, M., Mitra, P., and Castillo, C. (2016). “Twitter as a lifeline: Human-annotated Twitter corpora for NLP of crisis-related messages”. In: *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016*, pp. 1638–1643. arXiv: [1605.05894](#).
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). “RoBERTa: A Robustly Optimized BERT Pretraining Approach”. In: 1. arXiv: [1907.11692](#).
- Olteanu, A., Vieweg, S., and Castillo, C. (2015). “What to expect when the unexpected happens: Social media communications across crises”. In: *CSCW 2015 - Proceedings of the 2015 ACM International Conference on Computer-Supported Cooperative Work and Social Computing*, pp. 994–1009.
- Qazi, U., Imran, M., and Ofli, F. (2020). “GeoCoV19: A Dataset of Hundreds of Millions of Multilingual COVID-19 Tweets with Location Information”. In: pp. 1–3. arXiv: [2005.11177](#).
- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2019). “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer”. In: pp. 1–53. arXiv: [1910.10683](#).
- Rogers, A., Kovaleva, O., and Rumshisky, A. (2020). “A Primer in BERTology: What we know about how BERT works”. In: *arXiv* 8, pp. 842–866. arXiv: [2002.12327](#).
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). “Attention Is All You Need”. In: *arXiv:1706.03762 [cs]*.
- Waqas, H. and Imran, M. (2019). “#CampFiremissing: An analysis of tweets about missing and found people from California wildfires”. In: *Proceedings of the International ISCRAM Conference 2019-May*. May, pp. 848–856.
- Williams, A., Nangia, N., and Bowman, S. R. (2017). “A broad-coverage challenge corpus for sentence understanding through inference”. In: *arXiv*.
- Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. (2019). “HuggingFace’s Transformers: State-of-the-art Natural Language Processing”. In: arXiv: [1910.03771](#).
- Xu, C., Tao, D., and Xu, C. (2013). “A Survey on Multi-view Learning”. In: pp. 1–59. arXiv: [1304.5634](#).
- Yin, W., Hay, J., and Roth, D. (2020). “Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach”. In: *EMNLP-IJCNLP 2019 - Proceedings of the Conference*, pp. 3914–3923. arXiv: [1909.00161](#).
- Zahra, K., Imran, M., and Ostermann, F. O. (2020). “Automatic identification of eyewitness messages on twitter during disasters”. In: *Information Processing and Management* 57.1, p. 102107.