How does a Pre-Trained Transformer Integrate Contextual Keywords?

Application to Humanitarian Computing

Valentin Barriere, Guillaume Jacquet

European Commission's Joint Research Center

ISCRAM 21 – Oral presentation

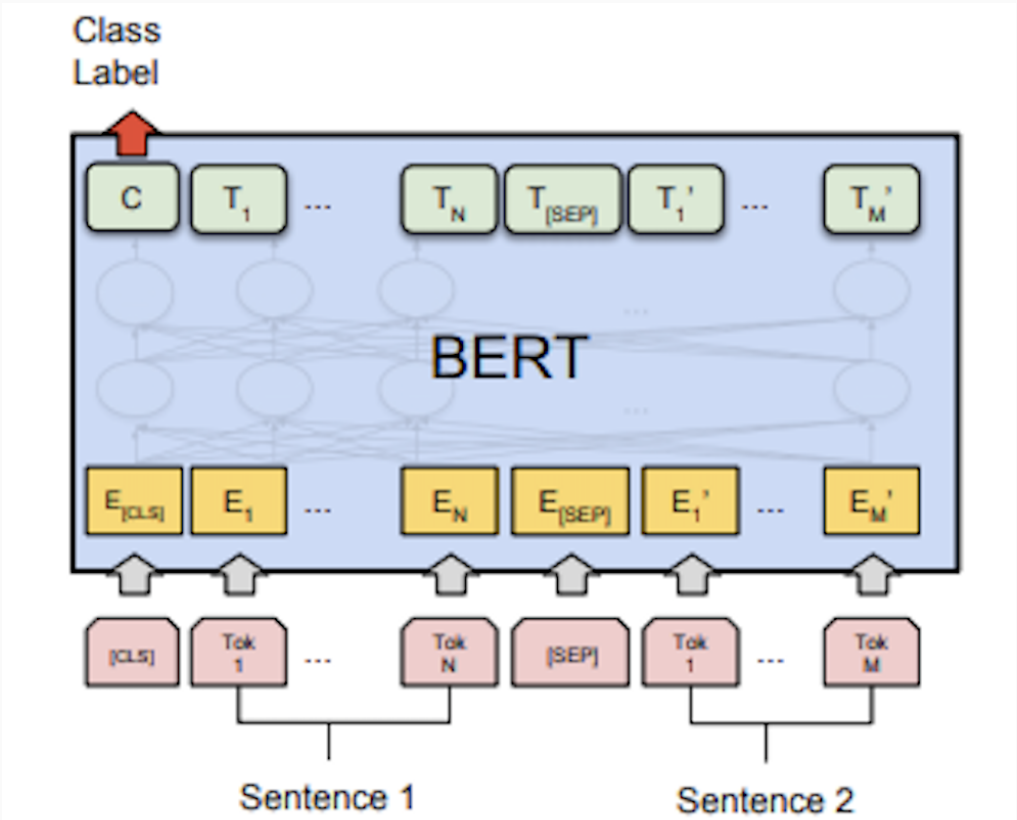
General Principle I

Transformers like BERT use general text like Wikipedia during pre-training, allowing to encode semantics knowledge [7].

It would be interesting to use the knowledge learned by the model regarding the word "*flood*", when classifying information from social media in the context of a flood.

General Principle II

Encode the event-type inside the model as a separate sentence, hence it does not interfere with the syntax of the text we want to classify.



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	`cmbk context: fire sentence: After deadly Brazil nightclub fire, safety questions emerge.`

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

Related Works

- [1] tackled a humanitarian classification task using pre-trained transformers, using simple concatenation to incorporate the event-type.
- [8, 4] encode the semantic content of the label inside the classifier.
- [2] studied the attention mechanism of a BERT model and clustered the attention heads

Research Questions

How to leverage the semantic information encoded inside a pre-trained model, in order to better classify a short text using textual metadata, and how to know it learns metadata-related patterns?

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?

Dataset : CrisisBench

We used the CrisisBench dataset from Alam et. al [1] composed of 87,557 tweets from several event types, labeled in 11 classes.

14 event types

Bombing, Collapse, Crash, Disease, Earthquake, Explosion, Fire, Flood, Hazard, Hurricane, Landslide, Shooting, Volcano, or none.

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 focus on the 11-humanitarian classification task, but also obtained good results on the binary relevance classification task.

Methodology

- 3 different transformers: BERT [3], RoBERTa [5], and T5 [6]
- Training over the official partition of the dataset
- Analysis of the label distribution of the dataset
- Training in a Leave-One-Event-Type-Out setting in order to make sure the models does not learn the label distributions of each event, overfitting over the dataset.
- Analysis of the word interacting the most with the event-type token, using the attention weights

Results – Official partition

Model	Event	Prec	Rec	u-F1	w-F1	Acc
BERT [1]	✓	70.1	71.3	70.7	86.5	86.5
RoBERTa [1]	✓	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**

Table 2: Results on the humanitarian classification task

Label distribution

The label distributions are very heterogeneous regarding the different events.

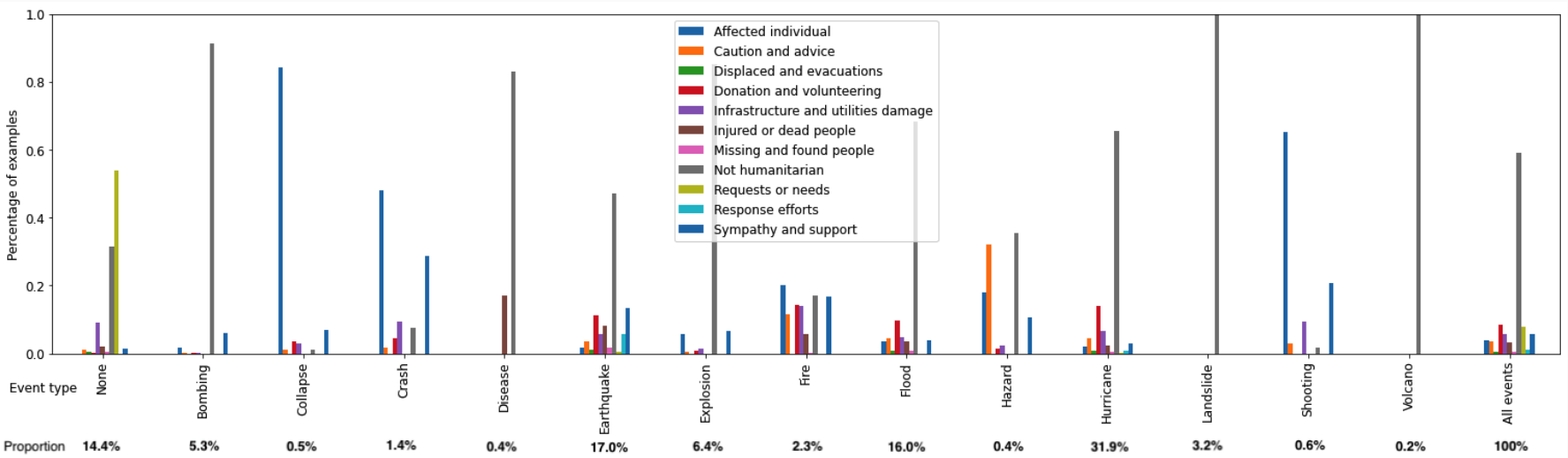


Figure 1: Distributions of labels regarding the event type in the train set, with the proportion of each event type

How to know that the model is not simply learning this pattern?

LOETO

Leave One Event Type Out Classification

In order to verify if the model was only learning the label distributions of each event, we proceeded to a LOETE. The event-aware model is still obtaining better results than the Vanilla one in this configuration.

14 trainings, every-time testing on a unknown event

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**

Table 3: Results of the BERT model on LOETE

Attention weights analysis

Clustering of the embedding of the 50 words having the highest attention weights *w.r.t.* the event-type word, **for an unknown event**.

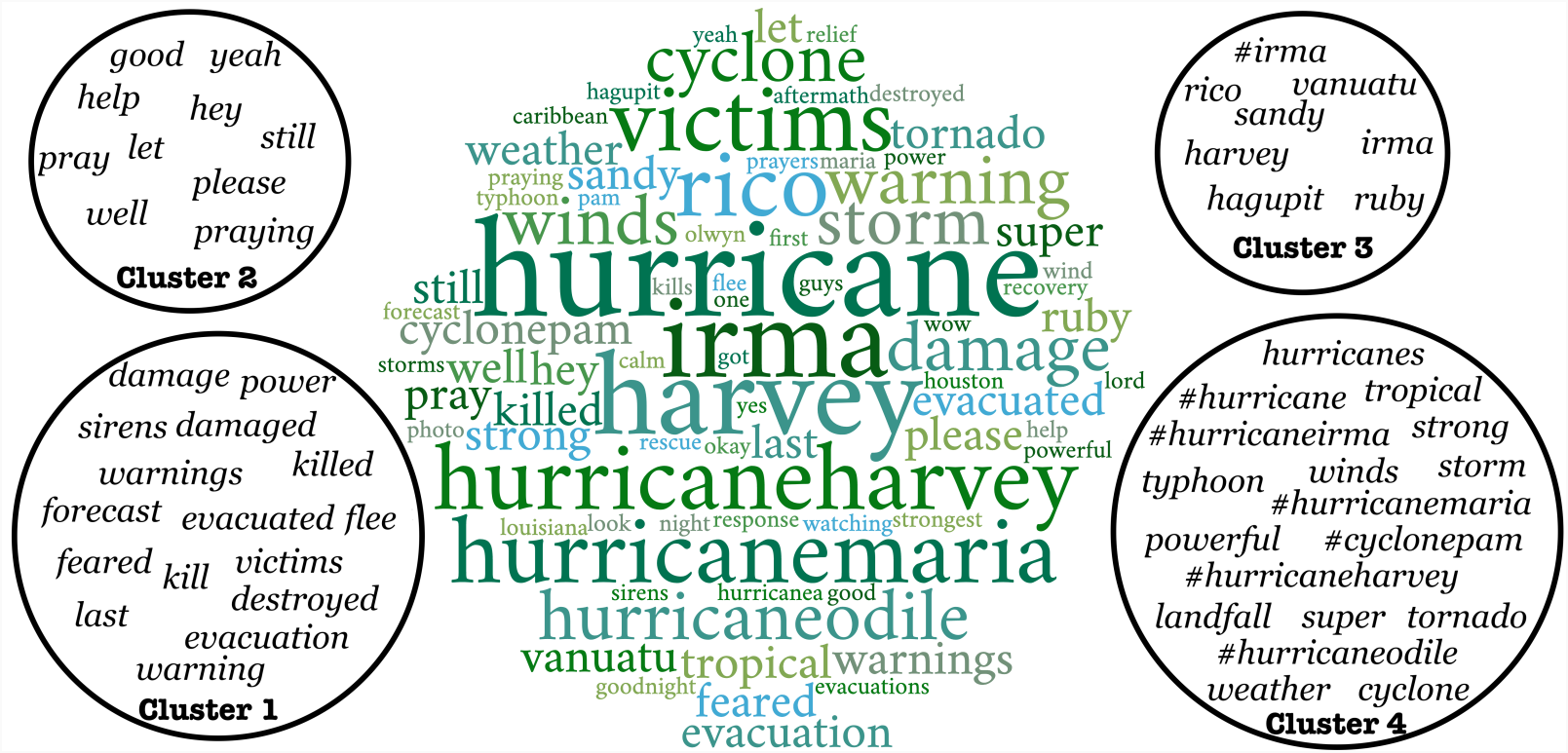


Figure 2: Tokens interacting the most with the event type '*hurricane*'. Clusters of the top-50 tokens.

Conclusion

- We studied the integration of a contextual information always available inside a pre-trained transformer model
- We made sure that the model is not only learning the label distributions of the event by training it with on a LOETE setting
- We looked at the interactions between the event-type and the other tokens of the tweet using the attention weights, and found meaningful clusters regarding the type of disaster, proper names, and events of the classification.

Questions?

Results Per Event

Partition	None	Bombing	Collapse	Crash
Official	91.2 (**1.2**)	96.7 (**0.4**)	88.8 (0.0)	89.3 (**1.1**)
LOETE	34.3 (**5.0**)	89.7 (**-4.3**)	44.1 (**19.7**)	81.5 (**-0.3**)

Disease	Earthquake	Explosion	Fire	Flood
98.6 (**2.9**)	77.0 (**1.2**)	96.6 (**0.3**)	81.5 (**-1.2**)	90.7 (**0.7**)
59.4 (**-11.3**)	49.4 (**-1.6**)	93.1 (**1.4**)	67.6 (**-4.2**)	85.3 (**1.7**)

Hazard	Hurricane	Lanslide	Shooting	Volcano
52.8 (0.0)	88.0 (**0.6**)	100 (**1.6**)	87.5 (0.0)	97.1 (0.0)
49.8 (**1.4**)	71.7 (**5.0**)	92.6 (**-0.6**)	77.8 (**7.1**)	72.0 (**-2.8**)

Table 4: Accuracies (differences with Vanilla) event by event of the event-aware BERT on the humanitarian classification task, for official partition and LOETE

References i


F. Alam, H. Sajjad, M. Imran, and F. Ofli.
CrisisBench: Benchmarking Crisis-related Social Media Datasets for Humanitarian Information Processing.
In *AAAI*, 2021.


K. Clark, U. Khandelwal, O. Levy, and C. D. Manning.
What does BERT look at? An analysis of BERT's attention.
arXiv, 2019.


J. Devlin, M.-w. Chang, K. Lee, and K. Toutanova.
BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.
2018.

References ii


K. Halder, A. Akbik, J. Krapac, and R. Vollgraf.
Task-Aware Representation of Sentences for Generic Text Classification.
In *COLING*, 2020.


Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov.
RoBERTa: A Robustly Optimized BERT Pretraining Approach.
(1), 2019.


C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu.
Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer.
pages 1–53, 2019.

References iii

A. Rogers, O. Kovaleva, and A. Rumshisky.
A Primer in BERTology: What we know about how BERT works.
arXiv, 8:842–866, 2020.

W. Yin, J. Hay, and D. Roth.
Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach.
EMNLP-IJCNLP 2019 - Proceedings of the Conference, pages 3914–3923, 2020.