Two improvements for mutli-lingual in-context classification over tweets using transformers

Application to Humanitarian Computing

Valentin Barriere

CENIA

RELELA – 01/25/23

Introduction

Tweet analysis

- Sentiment analysis

- Fake news detection

- Disaster response

Focus of today

Improving Sentiment Analysis over non-English Tweets using Multilingual Transformers and Automatic Translation for Data-Augmentation

Valentin Barriere
European Commission – DG-JRC
Via Enrico Fermi, 2749
21027 Ispra (VA), Italy
name.surname@ec.europa.eu

Alexandra Balahur
European Commission – DG-JRC
Via Enrico Fermi, 2749
21027 Ispra (VA), Italy
name.surname@ec.europa.eu

Figure 1: Paper from COLING 2020 [2]

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

Figure 2: Paper from ISCRAM 2021 [3]

Machine Translation-based Data-Augmentation

Main Principle

- We propose the use a multilingual pre-trained transformer instead of a monolingual one, so that it is possible to:
  - Adapt the model to the task by pre-training it over a huge annotated dataset of tweets in English
  - Adapt the model to other languages with a data-augmentation technique using automatic translation

Machine Translation for Data-Augmentation

- We proceed to data-augmentation by translating all the tweets from their native language to the 4 other languages used for testing.
- The translations from the source language to the 4 other languages were made by the automatic translation tool of the European Commission.

Lang.	Tweet
English	I'd rather dump gasoline all over myself and run into a burning building than use Excel.
French	Je préférerais jeter de l'essence partout et tomber dans un immeuble en feu plutôt que d'utiliser Excel.
German	Ich würde lieber Benzin auf mich werfen und in ein brennendes Gebäude laufen, als Excel zu benutzen.
Spanish	Prefiero tirar gasolina sobre mí mismo y correr hacia un edificio en llamas que usar Excel.
Italian	Preferirei buttarmi la benzina addosso e correre in un edificio in fiamme piuttosto che usare Excel.

Tweets Sentiment Analysis Datasets in 5 Languages

- We trained our models over 10 datasets and tested them over five different test sets in five languages: French, English, German, Spanish and Italian.
- This makes a total of 339,215 training examples when using data-augmentation with automatic translation.

Dataset	Language	Train	Dev	Test	All
SB-10k	German	4925	330	1315	6570
TASS-2019 TASS-2018	Spanish	2133	506	581	3220
DEFT-2015	French	6489	407	2938	9427
Sentipolc-16	Italian	6534	436	1964	8934
SemEval-2017 SemEval-2013 SemEval-2014 SemEval-2015 SemEval-2016	English	47762	2000	12284	62046

Base models

We use as classifiers:

Classifiers involved in this study

- XLM-RoBERTa [8] as a multilingual model
- Its monolingual counterparts CamemBERT [10] for French and RoBERTa [9] for English.
- AlBERTo [12] (BERT initialization) for Italian.

Results – Table

Language	Model	Using English	D-A	Rec_{avg}	$F1_{mac}$	$F1_{PN}$
English	[5] (winner SemEval-2017)	✓	✗	68.1	∅	68.5
	[11] (SOTA)	✓	✗	**73.2**	∅	**72.8**
	Monolingual	✓	✗	**72.8**	**71.7**	**72.3**
	Multilingual	✓	✗	71.9	70.0	70.3
		✓	✓	71.6	69.3	70.2
German	Multilingual	✗	✗	72.6	73.9	67.1
		✓	✗	74.1	**74.8**	**68.7**
		✓	✓	**74.2**	74.7	68.5
Spanish	Multilingual	✗	✗	63.5	63.2	72.7
		✓	✗	68.3	68.1	76.0
		✓	✓	**69.8**	**69.6**	**78.2**
French	Monolingual	✗	✗	72.9	72.8	71.6
	Multilingual	✗	✗	72.5	72.4	71.0
		✓	✗	73.8	73.7	72.2
		✓	✓	**74.4**	**74.5**	**72.8**
Italian	Monolingual	✗	✗	66.3	66.4	61.7
	Multilingual	✗	✗	63.0	60.7	55.3
		✓	✗	67.1	64.4	60.2
		✓	✓	**68.1**	**66.1**	**62.0**
All (non English)	Multilingual	✗	✗	68.0	67.6	66.6
		✓	✗	70.8	70.3	69.3
		✓	✓	**71.6**	**71.2**	**70.4**

Results – Comments

- Using English tweets to pre-train improves the results of the multilingual model.
- Data-Augmentation using Machine Translation allows once again to reach higher performances.
- The English monolingual model stays the most competitive.

Analysis

- Pre-training a multilingual model over English is a good option with a small target language training set (less than 6500).
- If there is enough of available data, it is better to use a monolingual model.
- Data-augmentation improves slightly the results for almost every language in different proportions. Our intuition is that the improvements follow the performances of the MT system.
- The utilization of English external data and data-augmentation allows to obtain better performances than the monolingual models for French and Italian.

Future Work

- Compare a zero-shot setting using English with/out data-augmentation.
- Extend to other European languages.

Integration of Textual Metadata into a Transformer

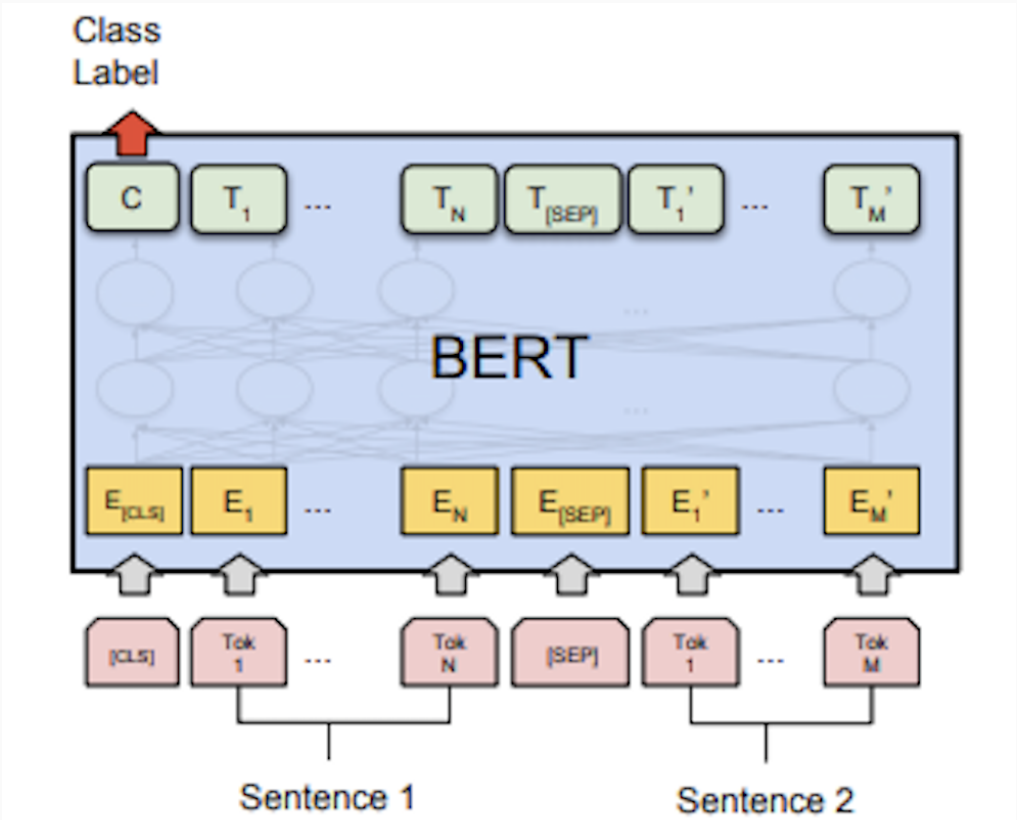
General Principle I

Transformers like BERT use general text like Wikipedia during pre-training, allowing to encode semantics knowledge [14].

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 4: 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.
- [16, 7] encode the semantic content of the label inside the classifier.
- [4] 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 [6], RoBERTa [9], and T5 [13]
- 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 5: Results on the humanitarian classification task

Label distribution

The label distributions are very heterogeneous regarding the different events.

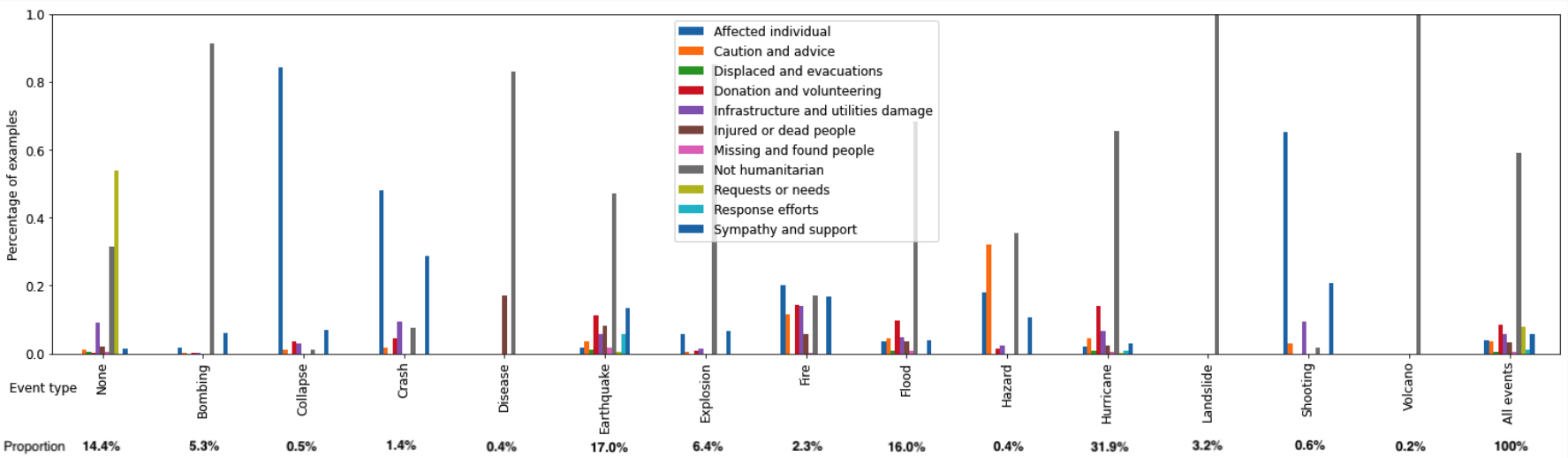


Figure 3: Distributions of labels regarding the event type in the train set, with the proportion of each event type

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 6: 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**.

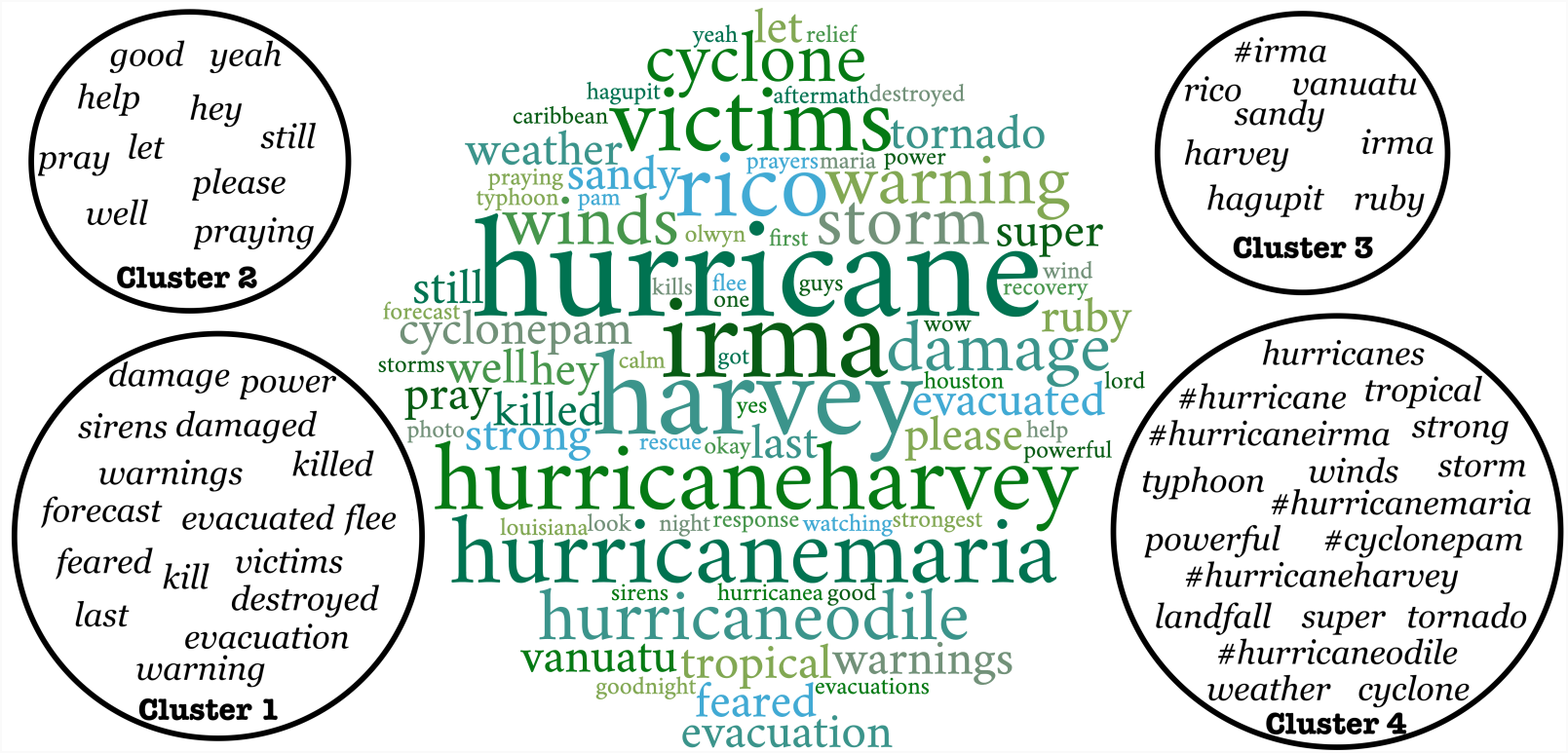


Figure 4: Tokens interacting the most with the event type '*hurricane*'. Clusters of the top-50 tokens.

Conclusion Metadata

- 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.

What about mixing them?

Multilingual Disaster-related tweets

Cross-Lingual and Cross-Domain Crisis Classification for Low-Resource Scenarios

Cinthia Sánchez,[1,2] **Hernan Sarmiento,**[1,2] **Andres Abeliuk,**[1,3] **Jorge Pérez,**[1,2] **Barbara Poblete,**[1,2,3]
[1]Department of Computer Science, University of Chile
[2]Millennium Institute for Foundational Research on Data (IMFD), Santiago, Chile
[3]National Center for Artificial Intelligence (CENIA), Santiago, Chile
{cisanche, hsarmien, aabeliuk, jperez, bpoblete}@dcc.uchile.cl

Small POC on Multi-domain Multi-lingual classification

We investigate the potential of our approaches on a mullti-lingual tweets for disaster response using the dataset from [15]

- Focusing on one event type: Earthquake – Spanish
- Focusing on XLM-R
- Not using any sampling method for train
- Not using any sampling method for test
- Weighting the examples to deal with class imbalance

Multilingual Disaster-related tweets

Model type	Prec	Rec	F1	Acc
Frozen XLM + RF	70.5	61.5	63.9	86.6
Frozen XLM + RF (train bal)	72.1	62.8	65.6	87.0
Vanilla XLM	77.0	69.8	72.5	88.7
DA	77.0	72.0	74.3	89.1
KW	77.8	72.9	75.0	89.3
Both	78.5	71.9	74.5	**89.4**

Table 7: Results of the BERT model on LOETE

Notes: One run only (time constraint), imbalance test set that does not represent the reality, all the data is not available anymore (Twitter account deleted/suspended)

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 8: 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.


V. Barriere and A. Balahur.
Improving Sentiment Analysis over non-English Tweets using Multilingual Transformers and Automatic Translation for Data-Augmentation.
In *COLING*, 2020.


V. Barriere and G. Jacquet.
How does a pre-trained transformer integrate contextual keywords? Application to humanitarian computing.
Proceedings of the International ISCRAM Conference, 2021-May(May):766–771, 2021.

References ii


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


M. Cliche.
BB_twtr at SemEval-2017 Task 4: Twitter Sentiment Analysis with CNNs and LSTMs.
SemEval-2017, (2014):573–580, 2017.


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

References iii


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


G. Lample and A. Conneau.
Cross-lingual Language Model Pretraining.
2019.


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.

References iv

L. Martin, B. Muller, O. S. P. Javier, Y. Dupont, L. Romary, E. Villemonte de la Clergerie, D. Seddah, and B. Sagot.
CamemBERT: a Tasty French Language Model.
In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

D. Q. Nguyen, T. Vu, and A. T. Nguyen.
BERTweet: A pre-trained language model for English Tweets.
2020.

M. Polignano, P. Basile, M. de Gemmis, G. Semeraro, and V. Basile.
AlBERTo: Italian BERT language understanding model for NLP challenging tasks based on tweets.
CEUR Workshop Proceedings, 2481, 2019.

References

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.

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

C. Sànchez, H. Sarmiento, J. Pérez, A. Abeliuk, and B. Poblete.
Cross-Lingual and Cross-Domain Crisis Classification for Low-Resource Scenarios.
Arxiv, 2022.


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