NTCIR 17 Conference: Proceedings of the 17th NTCIR Conference on Evaluation of Information Access Technologies, December 12-15, 2023, Tokyo, Japan
DOI: https://doi.org/10.20736/0002001279

FRAG at the NTCIR-17 MedNLP-SC Task

Anubhav Gupta
Yseop
France
agupta@yseop.com

Frédéric Rayar
Université de Tours
France
frederic.rayar@univ-tours.fr

ABSTRACT

The FRAG team participated in the Social Media (SM) subtask of the NTCIR-17 MedNLP-SC Task [13]. Our approach involved fine-tuning a multilingual transformer-based model on the train set. The team ranked 3^{rd} for English (SM-ADE-EN), German (SM-ADE-DE) and Japanese (SM-ADE-JA) based on Exact accuracy and Binary scores.

KEYWORDS

NLP, Multilingual, Transformers, Clinical texts

TEAM NAME

FRAG

SUBTASKS

SM-ADE-JA
SM-ADE-EN
SM-ADE-DE
SM-ADE-FR

1 INTRODUCTION

Prescribed by doctors, medicines are part of daily life for many people throughout the world. Yet, there is always a risk involved in taking medication. One of these risks is an Adverse Drug Event (ADE). It is *"an injury resulting from the use of a drug. Under this definition, the term ADE includes harm caused by the drug (adverse drug reactions and overdoses) and harm from the use of the drug (including dose reductions and discontinuations of drug therapy)"* [9].

An example of ADE mention, taken from the ADE Corpus v2 [6]: *"An 11-day-old infant became lethargic and apneic after a single drop of brimonidine"*. Since ADE can be dangerous to patients and a source of morbidity and mortality, hospitals and doctors need a system to support them in monitoring ADE occurrences in a fast and scalable way. To do so, Natural Language Processing (NLP) has been leveraged and has shown promising results [8].

1.1 Task description

The MedNLP-SC Social Media subtask addresses ADE detection from social media texts in four languages: Japanese (SM-ADE-JA), English (SM-ADE-EN), German (SM-ADE-DE) and French (SM-ADE-FR). There are 7965 synthetically generated tweets and we have to:

(1) identify the texts mentioning ADE,
(2) label the text with at least one of the 22 ADEs observed.

The training corpus is imbalanced: for each language 68.58% of the data did not have an ADE. Among the texts in the corpus that mention ADE, 53.19% had single label and among them 11 classes appeared in 80% of the messages with the ‘diarrhea’ label at 17.28%. One can find thorough details on the dataset in the task overview paper [13].

We approached the task as a multi-label classification task rather than an entity detection one.

2 RELATED WORK

Rawat et al. [11] used CNN to do binary classification (ADE vs. no ADE) on text extracted from MEDLINE. Zhang et al. [19] worked on binary classification. They trained a SVM on data scraped from DailyStrength and Twitter. Wunnava et al. [18] used a dual-attention network to perform joint task of ADE classification and NER. Wu et al. [17] created a tool composed of BERT, bi-LSTM-CRF [7] to identify ADE as named entities in *the unstructured section of Chinese ADR reports from the ADR monitoring center of Jiangsu Province in 2010-2016.* There are also works on ADR classification [1, 2, 5, 12, 14]. COLING hosted a similar shared task, SMM4H Task 1a [15], where the tweets in English containing ADE had to be identified. Most of the participants used BERT based models.

3 METHODS

First, we added a boolean column to the English dataset to state if there was at least 1 ADE or not. Then we separated 10% of the messages as validation set using scikit-learn’s [10] stratified train_test_split. We used the train-ids of English train and validation set to split the corpus for other languages. We combined train set of all the languages into a single dataset and we did the same with validation set. We fine-tuned a multi-lingual BERT [4], bert-base-multilingual-cased[1], and a multi-lingual RoBERTa [3], xlm-roberta-base[2] using Huggingface [16].

The parameters that have been used to fine-tune the two models are presented in Table 1. Except for the batch size, all the parameters are the huggingface defaults.

Table 1: Fine-tuning parameters.

	bert-base-multilingual-cased	xlm-roberta-base
max length[3]	128	128
learning rate	2e-5	2e-5
weight decay	0.01	0.01
epochs	10	10
batch size	16	32

At every epoch we computed the f1 for binary classification (ADE vs. no ADE). Around 4th epoch the eval loss for both models start rising and the eval f1 remains flat (See Figure 1 and Figure 2).

[1]https://huggingface.co/bert-base-multilingual-cased
[2]https://huggingface.co/xlm-roberta-base
[3]https://huggingface.co/docs/transformers/main_classes/tokenizer#transformers.PreTrainedTokenizer.__call__.max_length

For the first submission, we fine-tuned the multi-lingual BERT with the default parameters on all four languages combined without any preprocessing. We fine-tuned xlm-roberta in the same manner for the second. We chose the number of epochs arbitrarily as 6 for both submissions. Other parameters are those mentioned in Table 1.

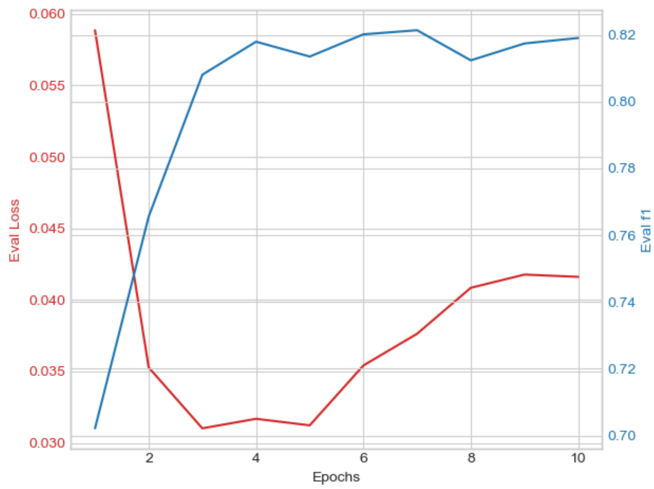

Figure 1: BERT Fine-tuning

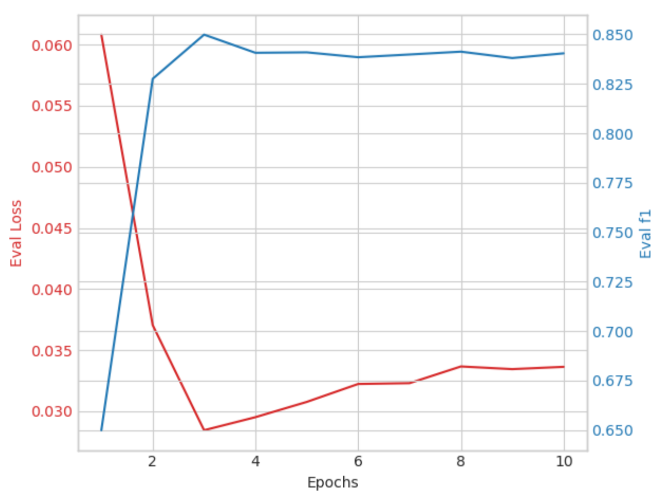

Figure 2: XLM-RoBERTa Fine-tuning

4 EXPERIMENTS

Unknown to us, the organization team had used fine-tuned xlm-roberta as their baseline. The only difference between these two models is the number of training epochs. The Tables 2 to 4 compare our model (Frag) with the winning team (Srcb) and the baseline (XLM-R_all).

Table 2: Binary Scores (ADE vs. no ADE) results.

Team	Japanese	English	German	French
Srcb	**0.881**	**0.872**	**0.873**	**0.869**
Frag (Submission 1)	0.83	0.82	0.82	0.82
Frag (Submission 2)	0.868	0.855	0.846	0.845
XLM-R_all	0.850	0.846	0.815	0.828

Table 3: (Full) Per Label Scores results.

Team	Japanese	English	German	French
Srcb	**0.910**	**0.905**	**0.908**	**0.902**
Frag (Submission 1)	0.87	0.87	0.86	0.86
Frag (Submission 2)	0.900	0.885	0.880	0.874
XLM-R_all	0.885	0.876	0.852	0.862

Table 4: Exact Match Accuracy results.

Team	Japanese	English	German	French
Srcb	**0.878**	**0.869**	**0.864**	**0.866**
Frag (Submission 1)	0.821	0.817	0.804	0.801
Frag (Submission 2)	0.858	0.841	0.833	0.828
XLM-R_all	0.837	0.828	0.803	0.806

5 CONCLUSIONS

We presented a finetuned multi-lingual RoBERTa model to identify tweets mentioning ADE. These social media texts were generated using a T5 model. Our model was an exact replica of the baseline (XLM-R_all) except it was finetuned to 6 epochs instead of 10. This small variation in the training made a big difference in the final ranking. The fact the models (of all the participants) scored the highest for Japanese and the lowest for French on all metrics is worth exploring. Lastly, it would be interesting to see how these models trained on synthetic data compare against the models trained on human generated social media texts.

REFERENCES

[1] Hassan Alhuzali and Sophia Ananiadou. 2019. Improving classification of Adverse Drug Reactions through Using Sentiment Analysis and Transfer Learning. In *Proceedings of the 18th BioNLP Workshop and Shared Task*. Association for Computational Linguistics, Florence, Italy, 339–347.

[2] I. S. Alimova and E. V. Tutubalina. 2019. Entity-Level Classification of Adverse Drug Reaction: A Comparative Analysis of Neural Network Models. *Programming and Computer Software* 45, 8 (Dec. 2019), 439–447.

[3] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised Cross-lingual Representation Learning at Scale. abs/1911.02116 (2019). http://arxiv.org/abs/1911.02116

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR* abs/1810.04805 (2018). http://arxiv.org/abs/1810.04805

[5] Vasiliki Foufi, Kuntheavy Ing Lorenzini, Jean-Philippe Goldman, Christophe Gaudet-Blavignac, Christian Lovis, and Caroline Samer. 2020. Automatic Classification of Discharge Letters to Detect Adverse Drug Reactions. *Studies in Health Technology and Informatics* 270 (June 2020), 48–52.

[6] Harsha Gurulingappa, Abdul Mateen Rajput, Angus Roberts, Juliane Fluck, Martin Hofmann-Apitius, and Luca Toldo. 2012. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *Journal of Biomedical Informatics* 45, 5 (2012), 885–892.

[7] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural Architectures for Named Entity Recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association*

for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, San Diego, California, 260–270.

[8] Rachel M. Murphy, Joanna E. Klopotowska, Nicolette F. de Keizer, Kitty J. Jager, Jan Hendrik Leopold, Dave A. Dongelmans, Ameen Abu-Hanna, and Martijn C. Schut. 2023. Adverse drug event detection using natural language processing: A scoping review of supervised learning methods. *PLOS ONE* 18, 1 (01 2023), 1–26.

[9] Jonathan R. Nebeker, Paul Barach, and Matthew H. Samore. 2004. Clarifying adverse drug events: a clinician's guide to terminology, documentation, and reporting. *Ann Intern Med.* 140(10) (2004), 795–801.

[10] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[11] Ashish Rawat, Mudasir Ahmad Wani, Mohammed ElAffendi, Ali Shariq Imran, Zenun Kastrati, and Sher Muhammad Daudpota. 2022. Drug Adverse Event Detection Using Text-Based Convolutional Neural Networks (TextCNN) Technique. *Electronics* 11, 20 (2022).

[12] Andrey Sakhovskiy and Elena Tutubalina. 2022. Multimodal model with text and drug embeddings for adverse drug reaction classification. *Journal of Biomedical Informatics* 135 (Nov. 2022), 104182.

[13] Shoko Wakamiya, Lis Kanashiro Pereira, Lisa Raithel, Hui-Syuan Yeh, Peitao Han, Seiji Shimizu, Tomohiro Nishiyama, Gabriel Herman Bernardim Andrade, Noriki Nishida, Hiroki Teranishi, Narumi Tokunaga, Philippe Thomas, Roland Roller, Pierre Zweigenbaum, Yuji Matsumoto, Akiko Aizawa, Sebastian Möller, Cyril Grouin, Thomas Lavergne, Aurélie Névéol, Patrick Paroubek, Shuntaro Yada, and Eiji Aramaki. 2023. NTCIR-17 MedNLP-SC Social Media Adverse Drug Event Detection: Subtask Overview. In *Proceedings of the 17th NTCIR Conference on Evaluation of Information Access Technologies, NTCIR-17*. National Institute of Informatics (NII).

[14] Chen-Kai Wang, Hong-Jie Dai, Feng-Duo Wang, and Emily Chia-Yu Su. 2018. Adverse Drug Reaction Post Classification with Imbalanced Classification Techniques. In *2018 Conference on Technologies and Applications of Artificial Intelligence (TAAI)*. IEEE, Taichung, 5–9.

[15] Davy Weissenbacher, Juan Banda, Vera Davydova, Darryl Estrada Zavala, Luis Gasco Sánchez, Yao Ge, Yuting Guo, Ari Klein, Martin Krallinger, Mathias Leddin, Arjun Magge, Raul Rodriguez-Esteban, Abeed Sarker, Lucia Schmidt, Elena Tutubalina, and Graciela Gonzalez-Hernandez. 2022. Overview of the Seventh Social Media Mining for Health Applications (#SMM4H) Shared Tasks at COLING 2022. In *Proceedings of The Seventh Workshop on Social Media Mining for Health Applications, Workshop & Shared Task*, Graciela Gonzalez-Hernandez and Davy Weissenbacher (Eds.). Association for Computational Linguistics, Gyeongju, Republic of Korea, 221–241.

[16] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. 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, 38–45.

[17] Hong Wu, Jiatong Ji, Haimei Tian, Yao Chen, Weihong Ge, Haixia Zhang, Feng Yu, Jianjun Zou, Mitsuhiro Nakamura, and Jun Liao. 2021. Chinese-Named Entity Recognition From Adverse Drug Event Records: Radical Embedding-Combined Dynamic Embedding–Based BERT in a Bidirectional Long Short-term Conditional Random Field (Bi-LSTM-CRF) Model. *JMIR Med Inform* 9, 12 (1 Dec 2021), e26407.

[18] Susmitha Wunnava, Xiao Qin, Tabassum Kakar, Xiangnan Kong, and Elke Rundensteiner. 2020. A Dual-Attention Network for Joint Named Entity Recognition and Sentence Classification of Adverse Drug Events. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Online, 3414–3423.

[19] Ying Zhang, Shaoze Cui, and Huiying Gao. 2020. Adverse drug reaction detection on social media with deep linguistic features. *Journal of Biomedical Informatics* 106 (2020), 103437.