CompLx@SMM4H'22: In-domain pretrained language models for detection of adverse drug reaction mentions in English tweets

Orest Xherija
Independent Researcher
xherija.orest@gmail.com

Hojoon Choi
Nielsen
hojoon.choi@nielsen.com

Abstract

The paper describes the system that team CompLx developed for sub-task 1a of the Social Media Mining for Health 2022 (#SMM4H) Shared Task. We finetune a RoBERTa model, a pretrained, transformer-based language model, on a provided dataset to classify English tweets for mentions of Adverse Drug Reactions (ADRs), i.e. negative side effects related to medication intake. With only a simple finetuning, our approach achieves competitive results, significantly outperforming the average score across submitted systems. We make the model checkpoints[1] and code[2] publicly available. We also create a web application[3] to provide a user-friendly, readily accessible interface for anyone interested in exploring the model's capabilities.

1 Introduction

The Shared Task (Weissenbacher et al., 2022) of the 2022 Social Media Mining for Health Applications (#SMM4H) workshop proposed ten sub-tasks in the domain of social media mining for health monitoring and surveillance. From the perspective of Natural Language Processing (NLP), these tasks present a considerable challenge since the nature of social media posts requires dealing with both a significant level of language variation (informal and colloquial expressions, ambiguity, multilingual posts) and data sparsity, as well as a widespread presence of noise such as misspellings of clinical concepts and syntactic errors.

In the 2022 instantiation of the #SMM4H Shared Task, our team participated in: (i) sub-task 1a, the classification of English tweets containing mentions of Adverse Drug Reactions (ADRs) (Magge et al., 2021), (ii) sub-task 3, the classification of English tweets (3a) and WebMD reviews (3b) containing mentions of changes in medication treatments, and (iii) sub-task 8, the classification of English tweets self-reporting chronic stress. In this paper we primarily describe our approach for task 1a, as that constituted the major focus of our efforts.

To address these challenges, we finetune a variant of a RoBERTa (Liu et al., 2019) model, a transformer-based (Vaswani et al., 2017) language model pretrained on approximately 128 million tweets (Loureiro et al., 2022) on each sub-task's provided dataset. Without any domain adaptation efforts (apart from standard finetuning on the downstream task) or hyperparameter optimizations, the model outperforms the average of all submissions for sub-task 1a by a 9% absolute difference in F1-score.

In the following sections, we introduce the sub-tasks' datasets, describe the model architecture and training setup, report our results, and conclude with a discussion of related research and potential avenues for future work.

2 Datasets

In Section [1] we provided a brief summary of each sub-task in which we participated. For each of them, participants were given access to a labeled training and validation set, as well as an unlabeled evaluation set that was used to determined the final performance of the submitted systems. Table [1] summarizes the number of samples per dataset per task. Additionally, Table [2] provides representative samples from sub-task 1a. As can be noted upon quick inspection, merely depending

	1a	3a	3b	8
training	17174	5898	10378	2936
validation	909	1572	1297	420
evaluation	10969	15360	13132	839

Table 1: Number of samples per split per task.

[1] https://huggingface.co/orestxherija/roberta-base-adr-smm4h2022

[2] https://github.com/orestxherija/CompLx-SMM4H2022

[3] https://huggingface.co/spaces/orestxherija/adr-mention-classifier

Proceedings of the 29th International Conference on Computational Linguistic, pages 176–181
October 12–17, 2022.

Sentence	Label
vyvanse make me so hyper and creative and i think of so many tweets	ADR
feed an ocd vyvanse and cover him in crayons	No ADR
trazodone has screwed up my sleep schedule. its helping tho.	ADR

Table 2: Selection of samples from training set of sub-task 1a.

on medication-related keywords for label assignment is going to be problematic: both the first and the second example contain the medication term "vyvanse" but they have been assigned different labels, "ADR" and "No ADR" respectively. This motivates the use of a modeling approach that leverages the overall semantic content of the sentence, rather than keyword matching with individual constituents.

3 Modeling Approach

3.1 Model Architecture

The establishment of language modeling as the pretraining step in the transfer learning pipeline revolutionized modern NLP with models such as ULMFiT (Howard and Ruder, 2018), ELMo (Peters et al., 2018) and, most notably, transformer-based language models such as GPT (Radford et al., 2018) and BERT (Devlin et al., 2019). In recent years, there have been intensive efforts in the research community to produce ever-larger transformer-based pretrained language models that are trained using a variety of datasets, transformer-model architectures, training objectives and optimization techniques. This should come as no surprise, since such language models have dominated virtually all NLP leaderboards, most notably GLUE (Wang et al., 2018) and SuperGLUE (Wang et al., 2019).

Considering this overwhelming success, we opt for a RoBERTa (Liu et al., 2019) model[4] that has been trained on approximately 128 million tweets (Loureiro et al., 2022). Our exact modeling approach is depicted in Figure [1]. We opt for a model that has been trained on an in-domain corpus, namely tweets, as transfer learning has been shown to yield improved results when there is in-domain pretraining (Gururangan et al., 2020). We do not use any text normalization steps.

[4]https://huggingface.co/cardiffnlp/twitter-roberta-base-mar2022

3.2 Training Regime

We train the model to minimize the negative log-likelihood loss using back-propagation with stochastic gradient descent and a mini-batch size of 16. To monitor model performance, we use the train/validation split provided by the organizers. For optimization, we use the AdamW optimizer (Loshchilov and Hutter, 2019) with gradient clipping (Pascanu et al., 2013) and a linear scheduler with no warm-up. We use FP-16 mixed precision (Micikevicius et al., 2018) training (and inference) in order to afford a larger batch size and increased training speed. To optimize GPU use by minimizing the amount of memory allocated for padding tokens, we use dynamic padding and length-based batching in the sense of (Skinner, 2018). Finally, we employ label smoothing (Szegedy et al., 2016) with a smoothing factor of 0.1.

3.3 Hyperparameters

As mentioned in Section [1], we do not experiment with hyperparameter tuning but rather keep the default parameters of the `Trainer` API in the Hugging Face `transformers` library. More specifically, we use $\beta_1 = 0.9, \beta_2 = 0.999$ and $\epsilon = 10^{-8}$ for the AdamW optimizer parameter values and a learning rate of 0.00005. We train the models for a maximum of 25 epochs with an early stopping patience level set to 0.001 for 3 epochs. Finally, we set a maximum sequence length of 128 since input sentences are generally short and we would like to avoid consuming GPU memory for padding tokens.

4 Experiments and Results

In this section, we give a brief description of the system we used to conduct our experiments, share our results and provide a brief discussion.

4.1 Setup

The model was developed using the PyTorch (Paszke et al., 2019) implementation of the Hugging Face `transformers` (Wolf et al., 2020) library. The experiments were executed on a

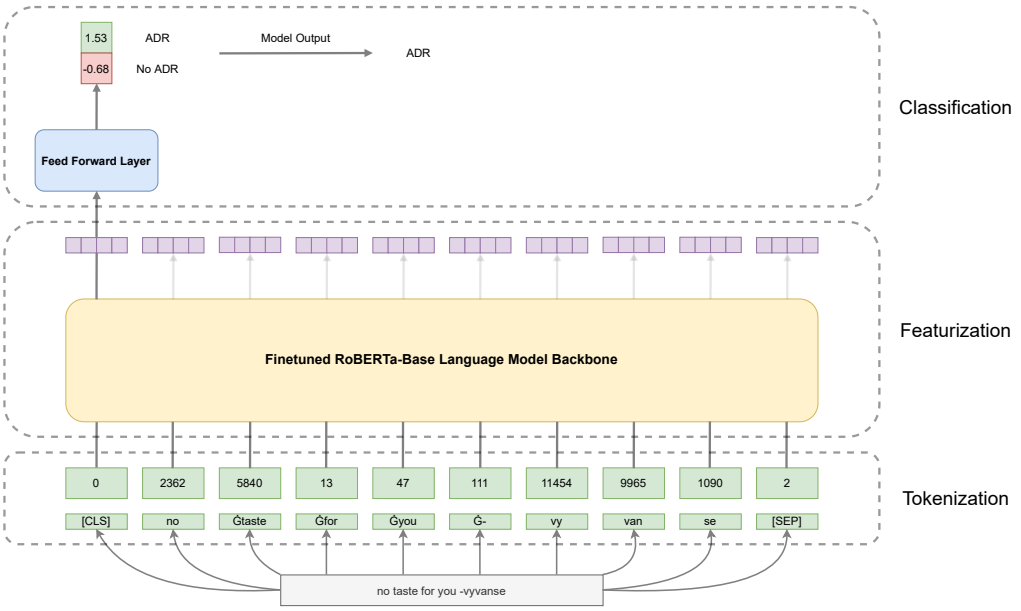

Figure 1: Illustration of modeling approach (inference step): the string input is tokenized and the tokens are passed through the language model backbone so as to obtain contextualized (vector) representations of the tokens. The vector associated with the [CLS] token is passed through a feed-forward layer and the logit outputs are used to decide the sample's class label, "ADR" or "No ADR".

machine with an Intel Core i9-9820X CPU @ 3.30GHz and a NVIDIA GeForce RTX 2080 Ti GPU with 11GB of memory.

4.2 Results

Table [3] summarizes the performance of our approach in the validation set for each sub-task. Note that in this set of experiments, the validation set was used both during training (e.g. for early stopping or selection of batch size) as well as for the reporting of the systems' performance. Table [4] summarizes the performance of our approach in the evaluation set for each sub-task. The organizers chose to disclose to each team only their respective score along with the average score of all submitted systems. Our system performed considerably better than the average in sub-task 1a and surpassed the existing state-of-the-art F1-score of 0.63 reported in (Magge et al., 2021). Performance was considerably poorer for sub-tasks 3 and 8. As mentioned in Section [1], our main efforts were dedicated to sub-task 1a and the system developed did not transfer well to the remaining sub-tasks.

	P	R	F1
1a	0.769	0.769	0.769
3a	0.030	0.312	0.055
3b	0.571	0.995	0.725
8	0.372	1.0	0.543

Table 3: Validation set results for sub-tasks 1a, 3 and 8.

	P	R	F1
Subtask-1a	**0.737** (0.646)	**0.585** (0.497)	**0.652** (0.562)
Subtask-3a	0.034 (0.535)	0.341 (0.458)	0.061 (0.456)
Subtask-3b	0.567 (0.778)	**1.0** (0.888)	0.723 (0.818)
Subtask-8	0.372 (0.720)	**1.0** (0.760)	0.542 (0.750)

Table 4: Results on the evaluation set for sub-tasks 1a, 3 and 8. Average score of all participating systems in parentheses. Metric is F1-score for class 1.

5 Conclusion and Future Directions

We demonstrated that a RoBERTa model (Liu et al., 2019) pretrained on approximately 128 million tweets performs very competitively when finetuned on English tweet classification for ADRs. Using only a standard finetuning approach, our model obtained competitive results, outperforming the average of all submissions for sub-task 1 by a 9% absolute difference in F1-score. This constitutes yet another testament of the fact that large pre-

trained language models have rightfully become the default approach in virtually all NLP tasks.

With respect to potential future work, there is a large collection of available options. Text classification and, more generally, binary classification is one of the oldest and most widely researched topics in NLP. Most approaches aiming to improve performance of classification models can be broadly categorized into three groups, depending on the segment of the machine learning workflow that they are targeting. Data augmentation methods typically target the initial part of the workflow, the data, aiming to increase the quantity, quality and diversity of the training dataset to ensure that model performance is robust to small syntactic or semantic perturbations in the inputs. Transformations acting directly on strings, such random token insertions or deletions, synonym/antonym replacements and related techniques (Wei and Zou, 2019; Karimi et al., 2021, inter alia) have shown significant performance improvements, especially in low-resource scenarios much like the one in this shared task.

A second approach, evidently a natural extension of the previous technique, would be to target vector encodings of the tokens and/or documents that are produced by the various layers of the neural networks. We can distinguish two different approaches here: (i) improve the language model backbone during the pretraining phase, or (ii) improve the weights of the language model backbone during finetuning. The research community has devoted intensive efforts in the former approach, as can be observed by the ever-increasing list of transformer-based pretrained language models (Devlin et al., 2019; Joshi et al., 2020; Kitaev et al., 2020; Raffel et al., 2020; Brown et al., 2020, inter multi alia) released. Model size, in terms of total number of trainable parameters, has been consistently shown to correlate strongly with downstream performance, so opting for a larger pretrained model would be a reasonable first steps towards more transferable vector representations (and hence improved performance) in the downstream task. The latter approach would include domain adaptation techniques, such as continued self-supervised pretraining followed by supervised finetuning, which has been shown (Gururangan et al., 2020) to consistently lead to superior results relative to direct finetuning.

Finally, one could aim to improve performance by modifying aspects of the objective function. (Hui and Belkin, 2021), in an extensive series of experiments, show that the established practice of using a cross-entropy loss for classification is not well-founded and show through a variety of diverse experiments that a square loss can, in many cases, significantly improve performance.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Jeremy Howard and Sebastian Ruder. 2018. Universal Language Model Fine-tuning for Text Classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

Like Hui and Mikhail Belkin. 2021. Evaluation of neural architectures trained with square loss vs cross-entropy in classification. In *International Conference on Learning Representations*.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving Pre-training by Representing and Predicting Spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Akbar Karimi, Leonardo Rossi, and Andrea Prati. 2021. AEDA: An Easier Data Augmentation Technique for

Text Classification. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2748–2754, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. 2020. Reformer: The Efficient Transformer. In *International Conference on Learning Representations*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR*, abs/1907.11692.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

Daniel Loureiro, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-collados. 2022. TimeLMs: Diachronic Language Models from Twitter. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 251–260, Dublin, Ireland. Association for Computational Linguistics.

Arjun Magge, Elena Tutubalina, Zulfat Miftahutdinov, Ilseyar Alimova, Anne Dirkson, Suzan Verberne, Davy Weissenbacher, and Graciela Gonzalez-Hernandez. 2021. DeepADEMiner: a deep learning pharmacovigilance pipeline for extraction and normalization of adverse drug event mentions on twitter. *Journal of the American Medical Informatics Association*, 28(10):2184–2192.

Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. 2018. Mixed Precision Training. In *International Conference on Learning Representations*.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. 2013. On the Difficulty of Training Recurrent Neural Networks. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1310–1318, Atlanta, GA, USA. PMLR.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Alec Radford, Karthik Narasimha, Tim Salimans, and Ilya Sutskever. 2018. Improving Language Understanding by Generative Pre-Training. Blog post, OpenAI.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Michael Skinner. 2018. Product Categorization with LSTMs and Balanced Pooling Views. In *SIGIR 2018 Workshop on eCommerce (ECOM 18)*, SIGIR '18, Ann Arbor, MI, USA. ACM.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the Inception Architecture for Computer Vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Jason Wei and Kai Zou. 2019. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Davy Weissenbacher, Ari Z. Klein, Luis Gascó, Darryl Estrada-Zavala, Martin Krallinger, Yuting Guo, Yao Ge, Abeed Sarker, Ana Lucia Schmidt, Raul Rodriguez-Esteban, Mathias Leddin, Arjun Magge,

Juan M. Banda, Vera Davydova, 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 Social Media Mining for Health (#SMM4H) Workshop and Shared Task*, Gyeongju, Republic of Korea. Association for Computational Linguistics.

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*, pages 38–45, Online. Association for Computational Linguistics.