How far can we go with just out-of-the-box BERT models?

Lucie M. Gattepaille
Uppsala Monitoring Centre, Uppsala, Sweden
lucie.gattepaille@who-umc.org[1]

Abstract

Social media have been seen as a promising data source for pharmacovigilance. However, methods for automatic extraction of Adverse Drug Reactions from social media platforms such as Twitter still need further development before they can be included reliably in routine pharmacovigilance practices. As the Bidirectional Encoder Representations from Transformer (BERT) models have shown great performance in many major NLP tasks recently, we decided to test its performance on the SMM4H Shared Tasks 1 to 3, by submitting results of pretrained and fine-tuned BERT models without more added knowledge than the one carried in the training datasets and additional datasets. Our three submissions all ended up above average over all teams' submissions: 0.766 F_1 for task 1 (15% above the average of 0.665), 0.47 F1 for task 2 (2% above the average of 0.46) and 0.380 F1 score for task 3 (30% above the average of 0.292). Used in many of the high-ranking submissions in the 2019 edition of the SMM4H Shared Task, BERT continues to be state-of-the-art in ADR extraction for Twitter data.

1 Introduction

Any medicinal product that has an effect, has the potential of causing harmful effects, in some individuals under some circumstances (Lindquist, 2008). Defined as 'a response to a drug that is noxious and unintended and occurs at doses normally used in man for the prophylaxis, diagnosis or therapy of disease, or for modification of physiological function' (WHO Meeting on International Drug Monitoring: the Role of National Centres, 1972), Adverse Drug Reactions (ADRs) are a public health concern and have been identified as the 5th leading cause of deaths within the European Union, with an estimated rate of 197,000 deaths per year and a cost of 79 billion euros within EU per year (Bouvy et al., 2015). Monitoring the safety of medicinal products over time, as they enter the market and get used in much more heterogenous populations than the clinical trials in which their safety was primarily assessed, is therefore essential for public health and is the goal of pharmacovigilance.

Traditionally, pharmacovigilance has relied on spontaneous reporting systems such as FAERS in the US and VigiBase, the WHO database if individual case safety reports, gathering reports of suspected ADRs from more than 130 national spontaneous reporting systems, including FAERS (Lindquist, 2003). Although these systems are particularly effective for detecting rare and serious ADRs, they suffer from limitations, notably under-reporting (Hazell and Shakir, 2006).

The appearance of social media platforms such as Twitter has provided pharmacovigilance with new data sources which could potentially complement spontaneous reports by the breadth of coverage of the populations (Sloane et al., 2015). Twitter alone was boasting 321 million active users as of February 2019 and thus could partially address the under-reporting problem spontaneous report systems suffer from. Nevertheless, it is yet unclear how social media could be integrated meaningfully into pharmacovigilance activities. The 2015-launched research consortium WEB-RADR (Ghosh and Lewis, 2015) has concluded that, under the current performance of methods for detection and extractions of ADRs, Twitter has limited value for pharmacovigilance (van Stekelenborg et al., 2019; Caster et al., 2018). Although, the methods they have employed are not today's state-of-the-art NLP methods (Gattepaille et

Proceedings of the 5th Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task, pages 95–100
Barcelona, Spain (Online), December 12, 2020.

al., 2020), it shows there is a need for developing performant algorithms for automatic detection, extraction and characterization of ADRs and their associated medicinal products.

To stimulate the development of such methods, the Social Media Mining for Health (SMM4H) has launched several Shared tasks over the years, with focus on specific aspects of the ADR extraction problem. This year, we participated in task 1 (binary classification for tweets containing a drug name or a dietary supplement), task 2 (binary classification for tweets containing an ADR mention) and task 3 (NER for ADR mentions and normalization to MedDRA® the Medical Dictionary for Regulatory Activities terminology is the international medical terminology developed under the auspices of the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH)). As the Bidirectional Encoder Representations from Transformer (BERT) (Devlin et al., 2019) was proven to be powerful in the 2019 SMM4H Shared Task (Weissenbacher et al., 2019; Miftahutdinov et al., 2019), we decided to apply pre-trained BERT models everywhere, with some fine-tuning to the tasks data, to see how this simple approach, requiring no domain expertise, would lead us. In this document, we report on the results of this experiment.

2 Methods

2.1 Preprocessing

All tweets and text extracts across the different datasets and tasks were preprocessed in the same manner. We first lowercased the text, then converted URLs, user tags as well as numbers to the special URL, USER and NUMBER tags respectively. We separated but kept the hash from the hashtags, collapsed characters appearing at least 3 times in a row into one character. Finally, we separated non-alpha characters from all other characters by a white space and collapsed all multiple white spaces into one, so the final token list became white-space-separated. For use in BERT models, all the preprocessed tweets were then passed to the BERT-base-uncased tokenizer, and consequently padded or truncated to a single length which was task-dependent.

2.2 Task 1 submission

For this task, we added the SMM4H 2018 tweets annotated for presence or absence of drug names to the training set (Weissenbacher et al., 2018), resulting in a training set of 65,041 tweets. During preprocessing, we padded/truncated all tweets to a length of 88 'tokens' (tokens include words, punctuation and non-alpha characters, and the smaller word chunks created by the BERT tokenizer). We fine-tuned the pre-trained BERT-base-uncased model found in the *Transformers* Python library from HuggingFace (Wolf et al., 2019) on the extended training set for 4 epochs, using the binary cross-entropy loss, with a batch size of 12, 0.1 dropout, the Adam optimizer and a linearly decreasing learning rate starting at 2e-5. A tweet was classified as a drug tweet if its post-softmax score exceeded 0.99. Very little fine-tuning of the hyperparameters was done. We submitted only one system run on the test set.

2.3 Task 2 submission

For this task, we also used a simple pre-trained BERT-base-uncased model and fine-tuned it on Task 2 training data, as well as Task 3's training and validation data, where the tweet labels were computed based on the presence or absence of an ADR mention, leading to a training set of 23,350 tweets. All tweets were padded/truncated to a length of 88 BERT tokens. The parameters and settings used for the model were the following: batch size of 12, dropout of 0.05, binary cross-entropy loss, Adam optimizer with a linearly decreasing learning rate starting at 2e-5 and 4 training epochs. A tweet was classified as an ADR tweet if its post-softmax score exceeded 0.99. Very little fine-tuning of the hyperparameters was done. We submitted only one system run on the test set.

2.4 Task 3 submission

For this task, we also applied BERT, both for the NER part and the normalization part of the task. We used the BIO-labelling scheme. BERT representations for all tokens were obtained via a pre-trained BERT-base-uncased model and passed to a softmax layer to classify each token as B, I or O. We did not include additional training data and only used the tweets with at least one ADR mention. We padded/truncated the tweets to a length of 50 BERT tokens. We trained the model for 16 epochs, with a

batch size of 12, the Adam optimizer with a linearly decreasing learning rate starting at 2e-5, 0.1 dropout, unweighted cross-entropy loss. All tokens predicted as I but preceded by a token predicted as O were converted to O. Separately, we trained a multi-class classifier based on another pre-trained BERT-base-uncased for the normalization part of the task. We combined the training ADR extracts with the CADEC (Karimi et al., 2015) and the SMM4H 2017 task 3 (Sarker and Gonzalez-Hernandez, 2017) datasets, leading to a total of 40,162 ADR-text and MedDRA PT code pairs for training, spanning over 674 unique PT codes. Text extracts were preprocessed and padded/truncated to a length of 40 BERT tokens. We topped the BERT output layer with a softmax layer on the 674 PT classes, and trained the entire model with cross-entropy loss, for 16 epochs, with an Adam optimizer and a fixed learning rate of 2e-5, a batch size of 12 and a dropout of 0.1. We submitted only one system run on the test set.

3 Results

3.1 Task 1

Our system performed above average on all metrics (F1, precision and recall), and was particularly performant in recalling the tweets with drug names or dietary supplements (Table 1). Although we did not apply a full grid search on the different hyperparameters of the model, we can still see some clear over-fitting to the validation set, especially regarding precision. As precision is the metric most influenced by prevalence of the positive class, a lower prevalence in the test set could partially explain a drop as well, but nothing in the task description indicated a lower prevalence of drug tweets in the test set. With only 35 positive examples in the validation set (thus presenting a high degree of imbalance), the chase for 'that extra positive example' can quickly have a strong effect on the generalization capabilities of a given architecture, leading us to believe that, that a particular 'improvement' to the model is nothing less than additional overfitting to the validation set.

	F1	**Precision**	**Recall**
Validation	0.82	0.7895	0.8571
Test	0.7665	0.7111	0.8312
Average Test	0.6646	0.7007	0.7039

Table 1: Comparison of performance for Task 1. The average test represents the performance metrics for the test dataset averaged across all Task 1 submissions.

3.2 Task 2

Our system performed above average in terms of F1-score but only slightly so (Table 2). It seemed clearly more geared towards precision, most likely owing to the very high threshold applied for ADR classification of a given tweet (0.99 on the post-softmax score). The large drop in performance between the validation and test performance came as a big surprise, considering that wedid very little hyperparameter tuning to the validation set, to avoid overfitting, that we kept the BERT approach rather similar across tasks and that the class imbalance was lower than in task 1 (at least on the validation set, with 0.2% drug tweets in task 1 against 9% ADR tweets in task 2). This seems to indicate systematic differences between the validation and test sets. A big difference in prevalence could explain in part the performance drop, although the biggest effect was observed on the recall metric which should, in theory, be more robust to prevalence effects. Without the ability to perform a proper analysis of the different types of errors made by the system on the test set, any explanation is pure speculation.

	F1	**Precision**	**Recall**
Validation	0.81	0.7412	0.8882
Test	0.47	0.58	0.40
Average Test	0.46	0.42	0.59

Table 2: Comparison of performance for Task 2. The average test represents the performance metrics for the test dataset averaged across all best submissions made by teams in Task 2.

3.3 Task 3

Our system performed above average on all relaxed metrics (F1, precision and recall), but only marginally so for the precision metric. The recall, on the other hand, exceeded the average by 51% (Table 3). The drop in performance between validation results and test results is less extensive than in task 2 but clearly more substantial than in task 1, maybe owing to the fact that overfitting happens both for the NER subtask and the normalization subtask. Our results on the strict metrics (when a true positive is obtained by matching the span of the ADR exactly and normalizing to the appropriate PT) were basically 0, revealing a likely error in indexing the start and end characters of the extractions.

	Metric type	F1	Precision	Recall
Validation	Relaxed	0.42	0.359	0.510
Test	Relaxed	0.380	0.335	0.439
Average Test	Relaxed	0.292	0.312	0.29

Table 3: Comparison of performance for Task 3. The average test represents the performance metrics for the test dataset averaged across all best submissions made by teams in Task 3.

Results of the NER subtask were also provided separately (Table 4). The simple BERT classifier on BIO labels was quite performant. The best performing system in the ADR NER task in the SMM4H 2019 (task 2) was using BioBERT (Lee et al., 2019) topped with a CRF and obtained and relaxed F1 score of 0.658 (0.554 precision and 0.81 recall) (Miftahutdinov et al., 2019), which is thus slightly lower than the performance of our system on this year's dataset. Although performance comparisons across different datasets should always be made with great caution, as one dataset may not be representative of the other, this shows that a simple pre-trained and fine-tuned BERT model can be powerful for this NER task, with relatively small amounts of data.

	Metric type	F1	Precision	Recall
Test	Relaxed	0.730	0.652	0.830
Average Test	Relaxed	0.564	0.607	0.557
Best system 2019	Relaxed	0.658	0.554	0.81

Table 4: Comparison of performance for the NER subtask of Task 3. The average test represents the performance metrics for the test dataset averaged across all best submissions made by teams in the NER subtask of Task 3.

4 Conclusion

"If all you have is a hammer, everything looks like a nail" (*Abraham Maslow*). As out-of-the-box pre-trained and fine-tuned BERT models have shown great performance in many kinds of NLP problems, we decided to pick up the BERT hammer and apply it to all tasks we registered for, to test its performance against the current efforts of the community for automated ADR extraction. Although we do not have the final rankings at the time of this writing, unfortunately, we see that our simple approach performed either slightly above (task 2) or largely above average (tasks 1 and 3). BERT has already been identified in the 2019 edition of the SMM4H Shared task as a contributor to good performance (Weissenbacher et al., 2019), as most high-ranking submissions in all tasks were using BERT or BioBERT in one way or another. We believe the picture is likely to be similar this year, although it will be interesting to see how the community integrated domain knowledge into their approaches and how such approaches fared against this submission, for which the domain knowledge is only included in the fine-tuning of the BERT algorithms.

Acknowledgements

MedDRA® trademark is registered by IFPMA on behalf of ICH.

References

Jacoline C. Bouvy, Marie L. De Bruin, and Marc A. Koopmanschap. 2015. Epidemiology of Adverse Drug Reactions in Europe: A Review of Recent Observational Studies. *Drug Safety*, 38(5):437–453, May.

Ola Caster, Juergen Dietrich, Marie-Laure Kürzinger, Magnus Lerch, Simon Maskell, G. Niklas Norén, Stéphanie Tcherny-Lessenot, Benoit Vroman, Antoni Wisniewski, and John van Stekelenborg. 2018. Assessment of the Utility of Social Media for Broad-Ranging Statistical Signal Detection in Pharmacovigilance: Results from the WEB-RADR Project. *Drug Safety*, 41(12):1355–1369, December.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*, May. arXiv: 1810.04805.

Lucie M. Gattepaille, Sara Hedfors Vidlin, Tomas Bergvall, Carrie E. Pierce, and Johan Ellenius. 2020. Prospective Evaluation of Adverse Event Recognition Systems in Twitter: Results from the Web-RADR Project. *Drug Safety*, May.

Rajesh Ghosh and David Lewis. 2015. Aims and approaches of Web-RADR: a consortium ensuring reliable ADR reporting via mobile devices and new insights from social media. *Expert Opinion on Drug Safety*, 14(12):1845–1853, December.

Lorna Hazell and Saad A W Shakir. 2006. Under-Reporting of Adverse Drug Reactions: A Systematic Review. *Drug Safety*, 29(5):385–396.

Sarvnaz Karimi, Alejandro Metke-Jimenez, Madonna Kemp, and Chen Wang. 2015. Cadec: A corpus of adverse drug event annotations. *Journal of biomedical informatics*, 55:73–81.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*:btz682, September.

Anna Marie Lindquist. 2003. *Seeing and observing in international pharmacovigilance: achievements and prospects in worldwide drug safety*. Uppsala Monitoring Centre, Uppsala.

Marie Lindquist. 2008. VigiBase, the WHO Global ICSR Database System: Basic Facts: *Drug Information Journal*, September.

Zulfat Miftahutdinov, Ilseyar Alimova, and Elena Tutubalina. 2019. KFU NLP team at SMM4H 2019 tasks: Want to extract adverse drugs reactions from tweets? BERT to the rescue. In *Proceedings of the Fourth Social Media Mining for Health Applications (# SMM4H) Workshop & Shared Task*, pages 52–57.

Abeed Sarker and Graciela Gonzalez-Hernandez. 2017. Overview of the Second Social Media Mining for Health (SMM4H) shared tasks at AMIA 2017. *Training*, 1(10,822):1239.

Richard Sloane, Orod Osanlou, David Lewis, Danushka Bollegala, Simon Maskell, and Munir Pirmohamed. 2015. Social media and pharmacovigilance: A review of the opportunities and challenges: Social media and pharmacovigilance. *British Journal of Clinical Pharmacology*, 80(4):910–920, October.

John van Stekelenborg, Johan Ellenius, Simon Maskell, Tomas Bergvall, Ola Caster, Nabarun Dasgupta, Juergen Dietrich, Sara Gama, David Lewis, Victoria Newbould, Sabine Brosch, Carrie E. Pierce, Gregory Powell, Alicia Ptaszyńska-Neophytou, Antoni F. Z. Wiśniewski, Phil Tregunno, G. Niklas Norén, and Munir Pirmohamed. 2019. Recommendations for the Use of Social Media in Pharmacovigilance: Lessons from IMI WEB-RADR. *Drug Safety*, 42(12):1393–1407, December.

Davy Weissenbacher, Abeed Sarker, Arjun Magge, Ashlynn Daughton, Karen O'Connor, Michael Paul, and Graciela Gonzalez. 2019. Overview of the fourth social media mining for health (SMM4H) shared tasks at ACL 2019. In *Proceedings of the Fourth Social Media Mining for Health Applications (# SMM4H) Workshop & Shared Task*, pages 21–30.

Davy Weissenbacher, Abeed Sarker, Michael Paul, and Graciela Gonzalez. 2018. Overview of the third social media mining for health (SMM4H) shared tasks at EMNLP 2018. In *Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task*, pages 13–16.

WHO Meeting on International Drug Monitoring: the Role of National Centres (1971: Geneva and World Health Organization). 1972. *International drug monitoring: the role of national centres , report of a WHO meeting [held in Geneva from 20 to 25 September 1971]*.World Health Organization technical report series ; no. 498. World Health Organization.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *ArXiv*, abs/1910.03771.