OFU@SMM4H'22: Mining Advent Drug Events Using Pretrained Language Models

Omar Adjali
Université Paris-Saclay, CEA
Laboratoire Analyse
Sémantique Texte et Image,
Gif-sur-Yvette, F-91191 France
omar.adjali@cea.fr

Fréjus A. A. Laleye,
Opscidia, Paris, France
frejus.laleye@
opscidia.com

Umang Aggarwal
Eco-Compteur, France
umang.aggarwal
@eco-counter.com

Abstract

We describe in this paper our proposed systems for the Social Media Mining for Health 2022 shared task 1. In particular, we participated in the three sub-tasks, tasks that aim at extracting and processing Adverse Drug Events. We investigate different transformer-based pretrained models we fine-tuned on each task and proposed some improvement on the task of entity normalization.

1 Introduction

Social media platforms became a compulsory source of information thanks to its immediacy and accessibility. The recent COVID-19 global crisis has demonstrated the necessity to increase the research effort toward processing health-related information with the aim of infoveillance, pharmacovigilance (Chew and Eysenbach, 2010) and fighting missinformation (Zarocostas, 2020). Indeed, the abundance of data provided by social media poses new challenges which encouraged the Natural Language Processing (NLP) community to design and adapt suitable text processing approaches. The Social Media Mining for Health (SMM4H) shared tasks (Weissenbacher et al., 2022) follow this initiative and propose several NLP tasks for health information monitoring. In particular, SMM4H task 1 proposes to tackle the problem of mining Adverse Drug Events (ADEs) in tweets using the annotated dataset proposed in (Magge et al., 2021). We describe in this paper the systems that solve the three subtasks we participated in: 1) Binary ADEs tweets classification. 2) ADE span detection in tweets and 3) ADE mention normalization.

2 Motivation

Active research effort have been spent on learning language representation models from large general-purpose corpora. In particular, Transformers-based models such as BERT (Devlin et al., 2018), XL-Net (Yang et al., 2019) and RoBERTa (Liu et al., 2019) greatly contributed to boosting the performance of many natural language processing (NLP) applications. Although, such models have strong generalization abilities, they need an additional domain-adaptation training step in order to be employed in supervised settings for downstream tasks. For example, in the biomedical domain, pretrained language models (LMs) such as clinicalBert (Huang et al., 2019), BioBERT (Lee et al., 2020) have been proposed. While these domain-adapted models unarguably outperform general ones, the performance gain may vary depending on a number of factors, including the nature of the downstream task, the proximity of the task domain to the domain-adaptation training corpus and the task dataset size (Gururangan et al., 2020). Furthermore, the characteristics of social media posts may adversely affect the general performance, as they are often in the form of short noisy texts with an informal writing style that involves abbreviations and misspelled words. In this work, we evaluate different pretrained language models on the three aforementioned tasks, investigating how well they transfer to different information extraction tasks in the context of social media and pharmacovigilance.

3 Task1a: Tweet Classification

In Task1a, the objective is to classify tweets as ADE or noADE. The training dataset is imbalanced and contains 1711 tweets reporting ADE and 15674 NoADE examples. Previous work such as (Vydiswaran et al., 2019; Cortes-Tejada et al., 2019), proposed a text processing pipeline and bidirectional LSTM to classify ADE tweets. (Rawal et al., 2019) proposed a lexicon-based approach that identified keywords used as markers of whether a tweet contains an ADE. (Miftahutdinov et al., 2019) relied on a bert architecture for tweet classification, while (Mahata et al., 2019) additionally showed that ULMFiT(Howard and Ruder, 2018) produces descent performance.

Proceedings of the 29th International Conference on Computational Linguistic, pages 171–175
October 12–17, 2022.

Following (Miftahutdinov et al., 2019), we first preprocessed tweets using the tweet-preprocessor package[1] to remove re-tweets, Hashtags, emojis and URLs. We investigate several Transformer-based pretrained language models (Vaswani et al., 2017) which we fine-tuned on ADE tweet classification. For each model, we take the pooled output ([CLS] token representation) followed by a linear layer and a softmax function for classification. For all tasks, we used the following models:

BERT: We used the "base-uncased" BERT pretrained on general-domain text corpora.

BioBERT (Lee et al., 2020): It is pretrained on biomedical domain corpora (PubMed abstracts + PMC full-text articles). We experimented with the BioBERT-Large v1.1 (PubMed 1M), a version based on BERT-large-Cased with a custom 30k vocabulary.

XLNet: With a different architecture, training objectives and without sequence length limit, XLNet performs autoregressive (AR) and permutation-based Language Modelling (PLM) to encode bidirectional contexts. (Wang et al., 2018) showed XLNet performance improvements over BERT on several benchmark datasets. We used the XLNet-Base pretrained model version.

Roberta (Liu et al., 2019): similar to BERT, RoBERTa is trained on larger general-domain corpora without the next sentence prediction objective. It has been shown that it outperforms BERT on several downstream tasks.

3.1 Experiments Setup

For all experiments, our model implementations are based on the huggingface library (Wolf et al., 2019) and pytorch (Paszke et al., 2019). Texts are tokenized with respect to each architecture tokenization scheme and special tokens. We fine-tuned the pretrained models on the down stream tasks for a maximum of 100 epochs with a batch size of 32 for BERT and RoBERTa models, and 64 for XLnet models. Early stopping is applied to stop learning after no improvement on the validation loss for 10 epochs. Model selection is based on the best validation performance on the development set, which represents 20% of the training set. Optimization is done using Adam optimizer with a learning rate of 3e-5 and a weight decay of 0.1. The performance was measured on the F1-score of the ADE classe.

[1]https://pypi.org/project/tweet-preprocessor/

Meth.	F1
BERT	0.692
BioBERT	0.634
XLNet	0.683
Roberta	**0.715**

Table 1: Classification results on the validation set

3.2 Discussion

Table 1 shows the classification results on the validation set. Results show that RoBERTa is the best performing model, while BERT performs slightly better than XLNet. Without surprise, RoBERTa being trained on larger corpora and with longer sequences, it confirms superiority over BERT and the XLNet architecture. Althought, BioBERT is a domain-specific model pretrained on biomedical corpora, it achieved the worse performance. We suggest that general-domain pretrained models have better generalization capabilities on social media posts even when their texts involve biomedical-related terms. Obviously, extensive experiments are necessary to support such an assumption. We conducted additional experiments by oversampling the minority class (ADE) in the training set to overcome the imbalance nature of the dataset, however experiments with BERT showed no improvements.

4 Task1b: ADE span detection in Tweets

Given a tweet, the goal is to detect a span that refers to an expressed ADE. Similar to (Rawal et al., 2019; Mahata et al., 2019; Miftahutdinov et al., 2019), we formulated the span detection task as a sequence labeling problem. We automatically annotated the dataset with labels following the BIO tagging scheme. Hence, the beginning of ADE mention is tagged with B-ADE, the inside tokens of an ADE mention are tagged with I-ADE, and the rest of tokens are tagged with O (outside ADE mention). Our model relies mainly on a sequence encoder (BERT, RoBERTa, BioBERT), followed by a Conditional Random Field (CRF) layer to predict output labels. We additionally implement a BiLSTM + CRF as a strong baseline for sequence labeling. We used the same hyper-parameter settings as for the classification task, and computed the F1 score of the correct detected spans. In this task, we did not experiment with XLNet because of implementation incompatibilities with the CRF layer.

Meth.	F1
BiLSTM + CRF	0.455
BERT + CRF	0.574
RoBERTa + CRF	**0.591**
BioBERT + CRF	0.526

Table 2: Span detection results on the validation set

4.1 Discussion

Table 2 shows the results obtained with different sequence encoders it terms of strict F1 score. The results are in line with the those obtained in task1a. Obviously, Transformers-based models outperformed the BiLSTM+CRF basline. RoBERTa produced the best performance while BioBERT failed to produce satisfying results. We suggest that pretraining on well formatted domain-specific documents poorly transfers on downstream tasks that involve short noisy texts like tweets. In particular, the variability of ADE span forms in tweets makes the detection very challenging.

5 Task1c: ADE mention normalization

In task1c, we map ADE mentions to their standard concept IDs in the Medical Dictionary for Regulatory Activities (MedDRA) vocabulary. Like (Vydiswaran et al., 2019), we used MedDRA dictionary from SIDER, the database of drugs and side effects (Kuhn et al., 2016).

We followed a metric learning approach to address the entity normalization task, where mentions are encoded using a neural network, and entity labels are represented using pretrained continuous vector representations. During training, these representations are optimized with the objective of minimizing the embedding distance between mention representations and their normalized entity representation according to a given metric.

In this work, span mention representations are obtained by encoding their texts using one of the previously used encoders (BERT, BioBERT, XLNet, RoBERTa), while entities are encoded using Word2Vec (Mikolov et al., 2013) representations obtained after training word embeddings on PubMed and PMC texts (Chiu et al., 2016). We therefore average the word embeddings of the words composing entity labels to build vector representations of all entities in the MedDRA dictionary.

During inference, we computed the cosine similarity between the vector representations of each detected span mention and all the entity vector representations in the MedDRA dictionary. Finally, we normalize mentions following two retrieval methods: 1) the nearest neighbor (NN) retrieval method which consists in selecting the closest entity representation in the embedding space, according to the cosine similarity. 2) the Cross-Domain Similarity Local Scaling (CSLS) (Conneau et al., 2017), which is a similarity measure that computes the cosine similarity normalized with the mean cosine similarity of each entity vector representation with its K entity neighbors (see (Conneau et al., 2017) for more details).

5.1 Distant supervision

The dateset being small sized, we offset by following a distant supervision strategy to synthetically augment the training data. We therefore used the the MedDRA dictionary labels as training examples of ADE mentions, where each entity label is normalized with its entity ID resulting in a training set of relatively large size (90k examples).

5.2 Experiments

We run several experiments to evaluate the two retrieval methods and the impact of distant supervision. We trained the encoders used in the previous tasks with the same hyper-parameter settings. Evaluation is done using the strict F1 metric. We assume that all span were correctly detected, as we used the ground truth span in the validation set to compute F1 scores.

5.3 Discussion

Table 4 shows the different configuration results. We first observe that the CSLS metric slightly improves the performance over the NN retrieval except for XLNet models. The CSLS metric achieved significant improvement in word translation retrieval by alleviating the problem of hubness in high-dimensional vector spaces (Conneau et al., 2017). We note a substantial performance gain for all distant supervision settings, with a 4 points performance gain which demonstrates the benefit of data augmentation on such small datasets. Moreover, we note that the performance are close when comparing the different encoders. BERT slightly outperforms RoBERTa, While XLNet has the worse performance on this task.

Task	$Ol-P$	$Ol-R$	$Ol-F1$	$Strict-P$	$Strict-R$	$Strict-F1$
1a	-	-	-	0.235	0.409	0.299
1b	0.178	0.288	0.22	0.097	0.158	0.12
1c	0.094	0.152	0.116	0.067	0.108	0.082

Table 3: Final results on the Test set. (OL:Overlapping, P:precision, R:recall, F1: F1 score).

Meth.	F1
BERT-NN	0.609
BERT-CSLS	0.616
BERT-CSLS+ Distant superv.	**0.656**
XLNet-NN	0.594
XLNet-CSLS	0.592
XLNet-CSLS + Distant superv.	0.638
RoBERTa-NN	0.612
RoBERTa-CSLS	0.614
RoBERTa-CSLS+ Distant superv.	0.651
BioBERT-NN	0.608
BioBERT-CSLS	0.612
BioBERT-CSLS+ Distant superv.	0.626

Table 4: Normalization results on the validation set

6 Conclusion

In this work, we explored different transformer-based pretrained models in a transfer learning setting for several information extraction tasks. Our approaches proposed to solve the three subtasks of SMM4H Shared Task 1, by exploring the best transfer pretrained model. The results we obtained are different from our expectations since domain-specific pretrained models such as BioBERT did not outperform general-domain on domain-specific downstream tasks. We suggested that general-domain pretrained models have better generalization abilities to extract ADE on social media texts. Further experiment and analysis are planned in order to draw more meaningful conclusions. Our results on the test set are showed in table 3. For each task, the run corresponds to the best system we obtained on the validation set. Unfortunately, results on the test are not in line with those obtained on the validation set. Only our entity normalization results reached the mean score of all task1c submissions. As Task 1 requires an end-to-end pipeline, the results on each subtask are affected by the performance of its upstram subtasks. We will further investigate our implementations used for test runs.

References

Cynthia Chew and Gunther Eysenbach. 2010. Pandemics in the age of twitter: content analysis of tweets during the 2009 h1n1 outbreak. *PloS one*, 5(11):e14118.

Billy Chiu, Gamal Crichton, Anna Korhonen, and Sampo Pyysalo. 2016. How to train good word embeddings for biomedical nlp. In *Proceedings of the 15th workshop on biomedical natural language processing*, pages 166–174.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Javier Cortes-Tejada, Juan Martinez-Romo, and Lourdes Araujo. 2019. Nlp@ uned at smm4h 2019: neural networks applied to automatic classifications of adverse effects mentions in tweets. In *Proceedings of the fourth social media mining for health applications (# SMM4H) workshop & shared task*, pages 93–95.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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. *arXiv preprint arXiv:2004.10964*.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*.

Michael Kuhn, Ivica Letunic, Lars Juhl Jensen, and Peer Bork. 2016. The sider database of drugs and side effects. *Nucleic acids research*, 44(D1):D1075–D1079.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

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. *arXiv preprint arXiv:1907.11692*.

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.

Debanjan Mahata, Sarthak Anand, Haimin Zhang, Simra Shahid, Laiba Mehnaz, Yaman Kumar, and Rajiv Shah. 2019. Midas@ smm4h-2019: identifying adverse drug reactions and personal health experience mentions from twitter. In *Proceedings of the fourth social media mining for health applications (# SMM4H) workshop & shared task*, pages 127–132.

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.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

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 32*, pages 8024–8035. Curran Associates, Inc.

Samarth Rawal, Siddharth Rawal, Saadat Anwar, and Chitta Baral. 2019. Identification of adverse drug reaction mentions in tweets–smm4h shared task 2019. In *Proceedings of the Fourth Social Media Mining for Health Applications (# SMM4H) Workshop & Shared Task*, pages 136–137.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

VG Vinod Vydiswaran, Grace Ganzel, Bryan Romas, Deahan Yu, Amy Austin, Neha Bhomia, Socheatha Chan, Stephanie Hall, Van Le, Aaron Miller, et al. 2019. Towards text processing pipelines to identify adverse drug events-related tweets: university of michigan@ smm4h 2019 task 1. In *Proceedings of the fourth social media mining for health applications (# SMM4H) workshop & shared task*, pages 107–109.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

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 *In Proceedings of the Seventh Social Media Mining for Health (#SMM4H) Workshop and Shared Task*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

John Zarocostas. 2020. How to fight an infodemic. *The lancet*, 395(10225):676.