BERT Prescriptions to Avoid Unwanted Headaches: A Comparison of Transformer Architectures for Adverse Drug Event Detection

Beatrice Portelli[1] **Edoardo Lenzi**[1] **Emmanuele Chersoni**[2]
Giuseppe Serra[1] **Enrico Santus**[3]

[1]AILAB UniUd - University of Udine, Italy
[2]The Hong Kong Polytechnic University
[3]Decision Science and Advanced Analytics for MAPV & RA, Bayer
{portelli.beatrice,lenzi.edoardo}@spes.uniud.it
emmanuele.chersoni@polyu.edu.hk
giuseppe.serra@uniud.it, enrico.santus@bayer.com

Abstract

Pretrained transformer-based models, such as BERT and its variants, have become a common choice to obtain state-of-the-art performances in NLP tasks. In the identification of Adverse Drug Events (ADE) from social media texts, for example, BERT architectures rank first in the leaderboard. However, a systematic comparison between these models has not yet been done. In this paper, we aim at shedding light on the differences between their performance analyzing the results of 12 models, tested on two standard benchmarks.

SpanBERT and PubMedBERT emerged as the best models in our evaluation: this result clearly shows that span-based pretraining gives a decisive advantage in the precise recognition of ADEs, and that in-domain language pretraining is particularly useful when the transformer model is trained just on biomedical text from scratch.

1 Introduction

The identification of Adverse Drug Events (ADEs) from text recently attracted a lot of attention in the NLP community. On the one hand, it represents a challenge even for the most advanced NLP technologies, since mentions of ADEs can be found in different varieties of online text and present unconventional linguistic features (they may involve specialized language, or consist of discontinuous spans of tokens etc.) (Dai, 2018). On the other hand, the task has an industrial application of primary importance in the field of digital pharmacovigilance (Sarker et al., 2015; Karimi et al., 2015b).

This raising interest is attested, for example, by the ACL workshop series on Social Media Health Mining (SMM4H), in which shared tasks on ADE detection have been regularly organized since 2016 (Paul et al., 2016; Sarker and Gonzalez-Hernandez, 2017; Weissenbacher et al., 2018, 2019). With the recent introduction of Transformers architectures and their impressive achievements in NLP (Vaswani et al., 2017; Devlin et al., 2019), it is not surprising that these tools have become a common choice for the researchers working in the area.

The contribution of this paper is a comparison between different Transformers on ADE detection, in order to understand which one is the most appropriate for tackling the task. Shared tasks are not the best scenario for addressing this question, since the wide range of differences in the architectures (which could include, for example, ensembles of Transformers and other types of networks) does not allow a comparison on the same grounds. In our view, two key questions deserve a particular attention in this evaluation. First, whether there is an advantage in using a model with some form of *in-domain language pretraining*, given the wide availability of Transformers for the biomedical domain (Lee et al., 2020; Gu et al., 2020). Second, whether a model trained *to predict coherent spans of text* instead of single words can achieve a better performance (Joshi et al., 2019), since our goal is to identify the groups of tokens corresponding to ADEs as precisely as possible.

Two models that we introduce for the first time in this task, SpanBERT and PubMedBERT, achieved the top performance. The former takes advantage of a span-based pretraining objective, while the latter shows that in-domain language data are better used for training the model from scratch, without any general-domain pretraining.

2 Related Work

2.1 ADE Detection

Automatic extraction of ADE in social media started receiving more attention in the last few years, given the increasing number of users that discuss their drug-related experiences on Twitter and similar platforms. Studies like Sarker and Gonzalez (2015); Nikfarjam et al. (2015); Daniulaityte et al. (2016) were among the first to propose machine learning systems for the detection of ADE in social media texts, using traditional feature engineering and word embeddings-based approaches.

With the introduction of the SMM4H shared task, methods based on neural networks became a more and more common choice for tackling the task (Wu et al., 2018; Nikhil and Mundra, 2018), and finally, it was the turn of Transformer-based models such as BERT (Devlin et al., 2019) and BioBERT (Lee et al., 2020), which are the building blocks of most of the top performing systems in the recent competitions (Chen et al., 2019; Mahata et al., 2019; Miftahutdinov et al., 2019).

At the same time, the task has been independently tackled also by researchers in Named Entity Recognition, since ADE detection represents a classical case of a challenging task where the entities can be composed by discontinuous spans of text (Stanovsky et al., 2017; Dai et al., 2020; Wunnava et al., 2020).

2.2 Transformers Architectures in NLP

There is little doubt that Transformers (Vaswani et al., 2017) have been the dominant class of NLP systems in the last few years. The "golden child" of this revolution is BERT (Devlin et al., 2019), which was the first system to apply the bidirectional training of a Transformer to a language modeling task. More specifically, BERT is trained with a Masked Language Modeling objective: random words in the input sentences are replaced by a [MASK] token and the model attempts to predict the masked token based on the surrounding context.

Following BERT's success, several similar architectures have been introduced in biomedical NLP, proposing different forms of in-domain training or using different corpora (Beltagy et al., 2019; Alsentzer et al., 2019; Lee et al., 2020; Gu et al., 2020). Some of them already proved to be efficient for ADE detection: for example, the top system of the SMM4H shared task 2019 is based on an ensemble of BioBERTs (Weissenbacher et al., 2019).

Another potentially interesting addition to the library of BERTs for ADE detection is SpanBERT (Joshi et al., 2019). During the training of SpanBERT, random contiguous spans of tokens are masked, rather than individual words, forcing the model to predict the full span from the tokens at its boundaries. We decided to introduce SpanBERT in our experiments because longer spans and relations between multiple spans of text are a key factor in ADE detection, and thus encoding such information is potentially an advantage.

3 Experimental Settings

3.1 Datasets

The datasets chosen for the experiments are two widely used benchmarks. They are annotated for the presence of ADEs at character level: each document is accompanied by list of start and end indices for the ADEs contained in it. We convert these annotations using the IOB annotation scheme for the tokens: B marks the start of a mention, I and O the tokens inside and outside a mention respectively.

CADEC (Karimi et al., 2015a) contains 1250 posts from the health-related forum "AskaPatient", annotated for the presence of ADEs. We use the splits made publicly available by Dai et al. (2020).

SMM4H is the training dataset for Task 2 of the SMM4H shared task 2019 (Weissenbacher et al., 2019). It contains 2276 tweets which mention at least one drug name, 1300 of which are positive for the presence of ADEs while the other 976 are negative samples. The competition includes a blind test set, but in order to perform a deeper analysis on the results, we use the training set only. As far as we know there is no official split for the training set alone, so we partitioned it into training, validation and test sets (60:20:20), maintaining the proportions of positive and negative samples. This split and the code for all the experiments are available at https://github.com/AilabUdineGit/ADE.

The datasets correspond to different text genres: the tweets of SMM4H are mostly short messages, containing informal language, while the texts of CADEC are longer and structured descriptions. To verify this point, we used the TEXTSTAT Python package to extract some statistics from the texts of the two datasets (see Appendix A).

3.2 Metrics

As evaluation metrics we use the Strict F1 score, which is commonly adopted for this task (Segura-Bedmar et al., 2013). It is computed at the entity level, and assigns a hit only in case of perfect match between the labels assigned by the model and the labels in the gold annotation.

In CADEC around 10% of mentions are discontinuous (Dai et al., 2020) and it is possible to have overlaps and intersections of discontinuous spans. We performed data tidying by merging overlapping ADE mentions, keeping only the longer span (as it is customary in the literature) and splitting discontinuous spans in multiple continuous spans.

3.3 Overview of the Models

3.3.1 Pretrained BERT Variants

Apart from the original BERT, we experimented with SpanBERT, for its peculiar pretraining procedure which focuses on predicting and encoding spans instead of single words, and with four BERT variants with in-domain knowledge, which differ from each other both for the corpus they were trained on and for the kind of pretraining.

BERT Standard model, pretrained on general purpose texts (Wikipedia and BookCorpus).

SpanBERT This model is pretrained using the same corpus as the original BERT, so it comes with no in-domain knowledge. But the pretraining procedure makes its embeddings more appropriate for NER-like tasks. as it introduces an additional loss called Span Boundary Objective (SBO), alongside the traditional Masked Language Modelling (MLM) used for BERT.
Let us consider a sentence $S = [w_1, w_2, \dots, w_k]$ and its substring $S_{m:n} = [w_m, \dots, w_n]$. w_{m-1} and w_{n+1} are the boundaries of $S_{m:n}$ (the words immediately preceding and following it). We *mask* S by replacing all the words in $S_{m:n}$ with the `[MASK]` token. SpanBERT reads the masked version of S and returns an embedding for each word. The MLM loss measures if it is possible to reconstruct each original word $w_i \in S_{m:n}$ from the corresponding embedding. The SBO loss measures if it is possible to reconstruct each $w_i \in S_{m:n}$ using the embeddings of the boundary words w_{m-1} and w_{n+1}.

BioBERT (Lee et al., 2020), pretrained from a BERT checkpoint, on PubMed abstracts.
The authors of BioBERT provide different versions of the model, pretrained on different corpora. We selected the version which seemed to have the greatest advantage on this task, according to the results by Lee et al. (2020). We chose BioBERT v1.1 (+PubMed), which outperformed other BioBERT v1.0 versions (including the ones trained on full texts) in NER tasks involving Diseases and Drugs. Preliminary experiments against BioBERT v.1.0 (+PubMed+PMC) confirmed this behaviour (see Appendix D).

BioClinicalBERT (Alsentzer et al., 2019), pretrained from a BioBERT checkpoint, on clinical texts from the MIMIC-III database.

SciBERT (Beltagy et al., 2019), pretrained *from scratch*, on papers retrieved from Semantic Scholar (82% of medical domain).

PubMedBERT (Gu et al., 2020), pretrained *from scratch*, on PubMed abstracts and full text articles from PubMed Central. This model was created to prove that pretraining from scratch on a *single domain* produces substantial gains on in-domain downstream tasks. Gu et al. (2020) compared it with various other models pretrained on either general texts, mixed-domain texts or in-domain texts starting from a general-purpose checkpoint (e.g. BioBERT), showing that PubMedBERT outperforms them on several tasks based on medical language. The vocabulary of PubMedBERT contains more in-domain medical words than any other model under consideration. However, it should be kept in mind that ADE detection requires an understanding of *both* medical terms and colloquial language, as both can occur in social media text.

Notice that two in-domain architectures were pretrained from scratch (SciBERT and PubMedBERT), meaning that they have a unique vocabulary tailored on their pretraining corpus, and include specific embeddings for in-domain words. BioBERT and BioClinicalBERT were instead pretrained starting from a BERT and BioBERT checkpoint, respectively. This means that the vocabularies are built from general-domain texts (similarly to BERT) and the embeddings are initialized likewise.

3.3.2 Simple and CRF Architecture

For all of the BERT variants, we take into account two versions. The first one simply uses the model to generate a sequence of embeddings (one for each sub-word token), which are then passed to a Linear Layer + Softmax to project them to the output space (one value for each output label) and turn them into a probability distribution over the labels.

The second version combines the Transformer-based model with a Conditional Random Field (CRF) classifier (Lafferty et al., 2001; Papay et al., 2020). The outputs generated by the first version become the input of a CRF module, producing another sequence of subword-level IOB labels. This step aims at denoising the output labels produced by the previous components.

The output labels are calculated for sub-word tokens, then we aggregate each set of sub-word labels $\{\ell_i\}$ into a word label $\mathcal{L}$ using the first rule that applies: (i) if $\ell_i = \text{O}$ for all i, then $\mathcal{L} = \text{O}$; (ii) if $\ell_i = \text{B}$ for any i, then $\mathcal{L} = \text{B}$; (iii) if $\ell_i = \text{I}$ for any i, then $\mathcal{L} = \text{I}$. The aggregated output is a sequence of word-level IOB labels.

3.3.3 Baseline

As a strong baseline, we used the **TMRLeiden** architecture (Dirkson and Verberne, 2019), which achieved the 2nd best Strict F1-Score in the latest SMM4H shared task (Weissenbacher et al., 2019) and is composed of a BiLSTM taking as input a concatenation of BERT and Flair embeddings (Akbik et al., 2019). We chose this baseline since the TMRLeiden code is publicly available.

3.4 Implementation details

TMRLeiden was re-implemented starting from its the original code[1] and trained according to the details in the paper. As for the Transformers, all experiments were performed using the TRANSFORMERS library (Wolf et al., 2019) (see Appendix C). Parameter-tuning was done via grid-search, using different learning rates ($[5e{-}4, 5e{-}5, 5e{-}6]$) and dropout rates (from 0.15 to 0.30, increments of 0.05). All the architectures were trained for 50 epochs on the training set. Learning rate, dropout rate and maximum epoch were chosen evaluating the models on the validation set.

During evaluation all the models were then trained using the best hyperparameters on the concatenation of the training set and the validation set, and tested on the test set. This procedure was repeated five times with different random seeds, and finally we averaged the results over the five runs.

4 Evaluation

The results for the two datasets are shown in Table 1 (we focus on the F1-score, but Precision and Recall are reported in Appendix D). For reference, we reported the scores of the best architecture by Dai et al. (2020), which is the state-of-the-art system on CADEC. At a glance, all systems perform better on CADEC, whose texts belong to a more standardized variety of language. SpanBERT and PubMedBERT emerge as the top performing models, with close F1-scores, and in particular, the SpanBERT models achieve the top score on both datasets, proving that modeling spans gives an important advantage for the identification of ADEs.

	SMM4H		**CADEC**	
Architecture	**F1**	**std**	**F1**	**std**
Dai et al. (2020)	–	–	**68.90**	–
TMRLeiden	60.70	2.08	65.03	1.14
BERT	54.74	1.40	65.20	0.47
BERT+CRF	59.35	1.23	64.36	0.83
SpanBERT	**62.15**	2.17	67.18	0.78
SpanBERT+CRF	59.89	2.16	**67.59**	0.60
PubMedBERT	61.88	0.79	67.16	0.52
PubMedBERT+CRF	59.53	2.07	67.28	0.82
BioBERT	57.83	2.59	65.59	1.10
BioBERT+CRF	58.05	1.45	66.00	0.67
SciBERT	57.75	1.55	65.61	0.54
SciBERT+CRF	58.86	1.55	67.09	0.74
BioClinicalBert	58.03	0.89	64.64	0.53
BioClinicalBert+CRF	59.11	1.99	65.97	0.60

Table 1: F1 scores with standard deviations for all models (our best performing model is in bold).

For both models, the addition of CRF generally determines a slight improvement on CADEC, while it is detrimental on SMM4H. On SMM4H, the F1-scores of BioBERT, SciBERT and BioClinicalBERT consistently improve over the standard BERT, but they are outperformed by its CRF-augmented version, while on CADEC they perform closely to the standard model. The results suggest that in-domain knowledge is consistently useful only when *training is done on in-domain text from scratch*, instead of using general domain text first. SciBERT is also trained from scratch, but on a corpus that is less specific for the biomedical domain than the PubMedBERT one (Gu et al., 2020).

The models are also being compared with TMRLeiden: we can notice that both versions of SpanBERT and PubMedBERT outperform it on CADEC (the differences are also statistically significant for the McNemar test at $p < 0.001$), while only the basic versions of the same models retain an advantage on it on SMM4H (also in this case, the difference is significant at $p < 0.001$).

4.1 Error Analysis

We analyzed the differences between the ADE entities correctly identified by the models and those that were missed, using the text statistics that we previously extracted with TEXTSTAT. As it was

[1] https://github.com/AnneDirkson/SharedTaskSMM4H2019

1	@hospitalpatient have been on humira 2years now n get on off **chest infections** that sometimes need 2diff pills 2sort out should i b worried ?	**4**	i have had no side effects been taking arthrotec a little over a year, have not noticed any side effects. it does help alot i noticed that when there are times when i forget to take it i can't stand or walk for any lengths of time.
2	had a great few hours on my bike but exercise drives my olanzapine **#munchies** . getting fed up with **not being able to fit into summer wardrobe**	**5**	works just fine. if there are any side effects, they are definitely not noticeable. what's with all these older people (70's) complaining about the lack of sex drive ? how much of what you are complaining about is simply related to getting older?
3	this new baccy is just making my **cough** so much worse but ahh well need my nicotine	**6**	what a great store @walmart is: i loss iq points , gained weight & got addicted to nicotine - all in under 15 min from going in !!

Table 2: Examples of ADEs extracted by PubMedBERT (overlined in blue) and SpanBERT (underlined in red). Actual ADEs in bold with gray background. The Samples belong to SMM4H (1–3, 6) and CADEC (4–5).

predictable, it turns out that longer ADE spans are more difficult to identify: for all models, we extracted the average word length of correct and missed spans and we compared them with a two-tailed Mann-Whitney U test, finding that the latter are significantly longer (Z = -6.176, $p < 0.001$). We also extracted the average number of difficult words in the correct and in the missed spans, defined as words with more than two syllables that are not included in the TEXTSTAT list of words of common usage in standard English. We took this as an approximation of the number of "technical" terms in the dataset instances. However, the average values for correct and missed instances do not differ (Z = 0.109, $p > 0.1$), suggesting that the presence of difficult or technical words in a given instance does not represent an inherent factor of difficulty or facilitation. Still, for some of the models – including SpanBERT, PubMedBERT and TMRLeiden – this difference reaches a marginal significance ($p < 0.05$) exclusively on the SMM4H dataset, where correctly identified spans have more difficult words. A possible interpretation is that, as the tweets' language is more informal, such words represent a stronger ADE cue, compared to the more technical language of the CADEC dataset.

Finally, we performed a qualitative analysis, comparing the predictions of SpanBERT and PubMedBERT. We selected the samples on which one of the architectures performed significantly better than the other one in terms of F1-Score, and analyzed them manually. Some significant samples can be found in Table 2. We observed that most of the samples in which PubMedBERT performed better than SpanBERT contained medical terms, which SpanBERT had completely ignored (e.g. Sample 1). The samples in which SpanBERT outperformed the in-domain model contained instead long ADE mentions, often associated with informal descriptions (e.g. Samples 2, 3). As regards false positives, both models make similar errors, which fit into two broad categories: (1) extracting diseases or symptoms of a disease (e.g. Samples 4, 6); (2) not being able to handle general mentions, hypothetical language, negations and similar linguistic constructs (e.g. Sample 5). While the second kind of error requires a deeper reflection, the first one might be addressed by training the model to extract multiple kinds of entities (e.g. both ADEs and Diseases).

5 Conclusions

We presented a comparison between 12 transformers-based models, with the goal of "prescribing" the best option to the researchers working in the field. We also wanted to test whether the span-based objective of SpanBERT and in-domain language pretraining were useful for the task. We can positively answer to the first question, since SpanBERT turned out to be the best performing model on both datasets. As for the in-domain models, PubMedBERT came as a close second after SpanBERT, suggesting that pretraining from scratch with no general domain data is the best strategy, at least for this task.

We have been the first, to our knowledge, to test these two models in a systematic comparison on ADE detection, and they delivered promising results for future research. For the next step, a possible direction would be to combine the strengths of their respective representations: the accurate modeling of text spans on the one side, and deep biomedical knowledge on the other one.

References

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. Flair: An Easy-to-use Framework for State-of-the-art nlp. In *Proceedings of NAACL*.

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly Available Clinical BERT Embeddings. In *Proceedings of the NAACL Workshop on Clinical Natural Language Processing*.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. In *Proceedings of EMNLP*.

Shuai Chen, Yuanhang Huang, Xiaowei Huang, Haoming Qin, Jun Yan, and Buzhou Tang. 2019. HITSZ-ICRC: A Report for SMM4H Shared Task 2019-Automatic Classification and Extraction of Adverse Effect Mentions in Tweets. In *Proceedings of the ACL Workshop on Social Media Mining for Health Applications*.

Xiang Dai. 2018. Recognizing Complex Entity Mentions: A Review and Future Directions. In *Proceedings of ACL 2018, Student Research Workshop*.

Xiang Dai, Sarvnaz Karimi, Ben Hachey, and Cécile Paris. 2020. An Effective Transition-based Model for Discontinuous NER. In *Proceedings of ACL*.

Raminta Daniulaityte, Lu Chen, Francois R Lamy, Robert G Carlson, Krishnaprasad Thirunarayan, and Amit Sheth. 2016. "When 'Bad' Is 'Good'": Identifying Personal Communication and Sentiment in Drug-related Tweets. *JMIR Public Health and Surveillance*, 2(2):e162.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL*.

Anne Dirkson and Suzan Verberne. 2019. Transfer Learning for Health-related Twitter Data. In *Proceedings of the ACL Workshop on Social Media Mining for Health Applications*.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *arXiv preprint arXiv:2007.15779*.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2019. SpanBERT: Improving Pre-training by Representing and Predicting Spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Sarvnaz Karimi, Alejandro Metke-Jimenez, Madonna Kemp, and Chenchen Wang. 2015a. Cadec: A Corpus of Adverse Drug Event Annotations. *Journal of Biomedical Informatics*, 55:73–81.

Sarvnaz Karimi, Chen Wang, Alejandro Metke-Jimenez, Raj Gaire, and Cecile Paris. 2015b. Text and Data Mining Techniques in Adverse Drug Reaction Detection. *ACM Computing Surveys (CSUR)*, 47(4):1–39.

John Lafferty, Andrew Mccallum, and Fernando Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of ICML*.

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.

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 ACL Workshop on Social Media Mining for Health Applications*.

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 ACL Workshop on Social Media Mining for Health Applications*.

Azadeh Nikfarjam, Abeed Sarker, Karen O'Connor, Rachel E. Ginn, and Graciela Gonzalez-Hernandez. 2015. Pharmacovigilance from Social Media: Mining Adverse Drug Reaction Mentions Using Sequence Labeling with Word Embedding Cluster Features. *Journal of the American Medical Informatics Association : JAMIA*, 22:671 – 681.

Nishant Nikhil and Shivansh Mundra. 2018. Neural DrugNet. In *Proceedings of the EMNLP Workshop on Social Media Mining for Health Applications*.

Sean Papay, Roman Klinger, and Sebastian Padó. 2020. Dissecting Span Identification Tasks with Performance Prediction. In *Proceedings of EMNLP*.

Michael Paul, Abeed Sarker, John Brownstein, Azadeh Nikfarjam, Matthew Scotch, Karen Smith, and Graciela Gonzalez. 2016. Social Media Mining for Public Health Monitoring and Surveillance. In *Biocomputing 2016*, pages 468–479.

Abeed Sarker, Rachel Ginn, Azadeh Nikfarjam, Karen O'Connor, Karen Smith, Swetha Jayaraman, Tejaswi Upadhaya, and Graciela Gonzalez. 2015. Utilizing Social Media Data for Pharmacovigilance: A Review. *Journal of Biomedical Informatics*, 54:202–212.

Abeed Sarker and Graciela Gonzalez. 2015. Portable Automatic Text Classification for Adverse Drug Reaction Detection via Multi-corpus Training. *Journal of Biomedical Informatics*, 53:196–207.

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

Isabel Segura-Bedmar, Paloma Martínez, and María Herrero-Zazo. 2013. SemEval-2013 Task 9 : Extraction of Drug-Drug Interactions from Biomedical Texts (DDIExtraction 2013). In *Proceedings of SemEval*.

Gabriel Stanovsky, Daniel Gruhl, and Pablo Mendes. 2017. Recognizing Mentions of Adverse Drug Reaction in Social Media Using Knowledge-Infused Recurrent Models. In *Proceedings of EACL*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Proceedings of NIPS*.

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 ACL Social Media Mining for Health Applications (# SMM4H) Workshop & Shared Task*.

Davy Weissenbacher, Abeed Sarker, Michael Paul, and Graciela Gonzalez. 2018. Overview of the Social Media Mining for Health (SMM4H) Shared Tasks at EMNLP 2018. In *Proceedings of the EMNLP Workshop on Social Media Mining for Health Applications*.

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. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *ArXiv*, abs/1910.03771.

Chuhan Wu, Fangzhao Wu, Junxin Liu, Sixing Wu, Yongfeng Huang, and Xing Xie. 2018. Detecting Tweets Mentioning Drug Name and Adverse Drug Reaction with Hierarchical Tweet Representation and Multi-head Self-attention. In *Proceedings of the EMNLP Workshop on Social Media Mining for Health Applications*.

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 EMNLP*.

A Text statistics for datasets

Some statistics for the texts of the two datasets have been extracted with the TEXTSTAT Python package and reported reported in Table A: we extracted the counts of syllables, lexicon (how many different word types are being used), sentences and characters. Difficult words refers to the number of polysyllabic words with Syllable Count > 2 that are not included in the list of words of common usage in English.

Metric	CADEC	SMM4H
Syllable Count	116 ± 2.7	26 ± 0.2
Lexicon Count	83 ± 1.9	18 ± 0.1
Sentence Count	5 ± 0.1	1 ± 0.0
Character Count	461 ± 10.5	104 ± 0.7
Difficult Words	14 ± 0.3	4 ± 0.1

Table 3: Average metrics per dataset, computed by the TEXTSTAT Python library.

B Best hyperparameters on the two datasets

Table 4 reports the best hyperparameters for all architectures on SMM4H and CADEC, respectively.

	SMM4H			CADEC		
Architecture	**lr**	**dropout**	**epoch**	**lr**	**dropout**	**epoch**
BERT	$5e{-}5$	0.20	4	$5e{-}5$	0.25	11
BERT+CRF	$5e{-}5$	0.15	6	$5e{-}5$	0.15	7
SpanBERT	$5e{-}5$	0.25	43	$5e{-}5$	0.25	19
SpanBERT+CRF	$5e{-}5$	0.15	14	$5e{-}5$	0.15	11
PubMedBERT	$5e{-}5$	0.25	21	$5e{-}5$	0.15	7
PubMedBERT+CRF	$5e{-}5$	0.25	13	$5e{-}5$	0.25	16
BioBERT	$5e{-}5$	0.20	8	$5e{-}5$	0.25	12
BioBERT+CRF	$5e{-}5$	0.15	6	$5e{-}5$	0.20	9
SciBERT	$5e{-}5$	0.15	7	$5e{-}5$	0.15	6
SciBERT+CRF	$5e{-}5$	0.25	13	$5e{-}5$	0.25	12
BioClinicalBERT	$5e{-}5$	0.25	10	$5e{-}5$	0.25	6
BioClinicalBERT+CRF	$5e{-}5$	0.25	12	$5e{-}5$	0.25	10

Table 4: Best hyperparameters for all Transformer-based architectures on SMM4H and CADEC.

C General information on the models

Table 5 is a summary of the information about the version of all Transformer-based models used and their pretraining methods.

D Detailed metrics of all the models

Table 6 and 7 report as Strict and Partial metrics the F1-score, Precision and Recall calculated for all architectures on SMM4H and CADEC respectively. Results are the average over five runs.

The Partial scores are standard metrics for this task (Weissenbacher et al., 2019) and take into account "partial"matches, in which it is sufficient for a system prediction to partially overlap with the gold annotation to be considered as a true match.

Name	**Version**	**Vocabulary**	**Pretraining**	**Pretraining Corpus**
BERT	base uncased	Wikipedia+BookCorpus	from scratch	Wikipedia+BookCorpus
SpanBERT	base cased	Wikipedia+BookCorpus	from scratch	Wikipedia+BookCorpus
PubMedBERT	base uncased abstract+fulltext	PubMed	from scratch	PubMed+PMC
BioBERT	base v1.1 (+PubMed)	Wikipedia+BookCorpus	from BERT	PubMed
BioBERT(v1.0)	base v1.0 (+PubMed+PMC)	Wikipedia+BookCorpus	from BERT	PubMed+PMC
SciBERT	scivocab cased	Semantic Scholar	from scratch	Semantic Scholar
BioClinicalBERT	bio+clinical	Wikipedia+BookCorpus	from BioBERT	MIMIC-III

Table 5: Information about the version of all the Transformer-based models used and their pretraining.

Strict				**Partial**		
F1	**P**	**R**	**Architecture**	**F1**	**P**	**R**
60.70 ± 2.08	**68.36 ± 2.41**	54.59 ± 1.97	TMRLeiden	66.08 ± 1.79	**74.42 ± 2.11**	59.43 ± 1.76
54.74 ± 1.40	48.50 ± 1.67	62.84 ± 1.12	BERT	64.53 ± 1.09	57.17 ± 1.52	74.08 ± 0.78
59.35 ± 1.23	54.12 ± 1.19	65.69 ± 1.34	BERT+CRF	68.35 ± 0.64	62.33 ± 0.74	75.66 ± 0.68
62.15 ± 2.17	54.54 ± 3.06	**72.31 ± 1.30**	SpanBERT	69.38 ± 1.60	60.88 ± 2.74	**80.74 ± 1.08**
59.89 ± 2.16	54.86 ± 3.10	66.05 ± 1.93	SpanBERT+CRF	68.09 ± 1.51	62.35 ± 2.79	75.10 ± 1.72
61.88 ± 0.79	58.70 ± 0.83	65.45 ± 1.39	PubMedBERT	**69.82 ± 0.60**	66.23 ± 0.86	73.84 ± 1.26
59.53 ± 2.07	55.29 ± 2.19	64.49 ± 2.27	PubMedBERT+CRF	67.94 ± 1.48	63.10 ± 1.69	73.61 ± 1.84
55.22 ± 1.71	49.85 ± 1.76	61.89 ± 1.78	BioBERT v1.0	64.25 ± 1.09	58.00 ± 1.22	72.02 ± 1.30
57.83 ± 2.59	53.68 ± 3.20	62.72 ± 2.30	BioBERT	66.58 ± 1.34	61.79 ± 2.25	72.23 ± 1.42
58.05 ± 1.45	54.44 ± 2.18	62.22 ± 1.22	BioBERT+CRF	66.30 ± 0.85	62.17 ± 1.83	71.07 ± 1.15
57.75 ± 1.55	53.49 ± 0.97	62.75 ± 2.54	SciBERT	66.49 ± 0.83	61.61 ± 0.61	72.25 ± 1.89
58.86 ± 1.55	52.94 ± 2.27	66.35 ± 1.86	SciBERT+CRF	67.12 ± 0.97	60.36 ± 1.93	75.67 ± 1.99
58.03 ± 0.89	51.63 ± 1.51	66.26 ± 0.46	BioClinicalBERT	66.90 ± 0.57	59.52 ± 1.29	76.39 ± 0.99
59.11 ± 1.99	52.35 ± 2.55	67.92 ± 1.55	BioClinicalBERT+CRF	67.41 ± 1.19	59.69 ± 1.92	77.48 ± 1.40

Table 6: Results on SMM4H, F1-scores, Precision and Recall calculated as Strict and Partial metrics, with standard deviations for all models.

Strict				**Partial**		
F1	**P**	**R**	**Architecture**	**F1**	**P**	**R**
65.03 ± 1.14	**67.50 ± 1.01**	62.75 ± 1.26	TMRLeiden	77.08 ± 0.78	**79.99 ± 0.60**	74.36 ± 0.97
65.20 ± 0.47	62.86 ± 0.52	67.72 ± 0.70	BERT	77.73 ± 0.28	74.95 ± 0.57	80.74 ± 0.47
64.36 ± 0.83	62.47 ± 0.97	66.36 ± 0.79	BERT+CRF	77.23 ± 0.45	74.97 ± 0.72	79.63 ± 0.41
67.18 ± 0.78	65.84 ± 0.94	**68.57 ± 0.78**	SpanBERT	79.18 ± 0.61	77.60 ± 0.79	**80.82 ± 0.72**
67.59 ± 0.60	67.09 ± 0.54	68.10 ± 0.78	SpanBERT+CRF	**79.43 ± 0.27**	78.84 ± 0.24	80.02 ± 0.60
67.16 ± 0.52	66.60 ± 0.67	67.73 ± 0.57	PubMedBERT	79.13 ± 0.23	78.47 ± 0.51	79.81 ± 0.42
67.28 ± 0.82	66.69 ± 0.99	67.88 ± 0.91	PubMedBERT+CRF	79.12 ± 0.43	78.43 ± 0.72	79.83 ± 0.71
65.54 ± 0.47	64.24 ± 0.48	66.90 ± 0.46	BioBERT v1.0	77.86 ± 0.34	76.32 ± 0.36	79.47 ± 0.33
65.59 ± 1.10	64.86 ± 1.39	66.34 ± 0.85	BioBERT	78.17 ± 0.75	77.30 ± 1.13	79.06 ± 0.48
66.00 ± 0.67	65.52 ± 0.97	66.48 ± 0.63	BioBERT+CRF	78.24 ± 0.43	77.68 ± 0.81	78.82 ± 0.58
65.61 ± 0.54	64.46 ± 0.70	66.80 ± 0.50	SciBERT	78.05 ± 0.19	76.69 ± 0.36	79.47 ± 0.46
67.09 ± 0.74	65.99 ± 0.74	68.23 ± 0.80	SciBERT+CRF	79.01 ± 0.35	77.72 ± 0.36	80.35 ± 0.50
64.64 ± 0.53	61.99 ± 0.51	67.53 ± 0.56	BioClinicalBERT	76.95 ± 0.35	73.80 ± 0.36	80.39 ± 0.37
65.97 ± 0.60	64.23 ± 1.16	67.82 ± 0.60	BioClinicalBERT+CRF	77.98 ± 0.49	75.92 ± 1.26	80.17 ± 0.53

Table 7: Results on CADEC, F1-scores, Precision and Recall calculated as Strict and Partial metrics, with standard deviations for all models.