KFU NLP Team at SMM4H 2021 Tasks: Cross-lingual and Cross-modal BERT-based Models for Adverse Drug Effects

Andrey Sakhovskiy
Kazan Federal University
Kazan, Russia

Zulfat Miftahutdinov
Kazan Federal University
Kazan, Russia

Elena Tutubalina
Kazan Federal University
Kazan, Russia
HSE University
Moscow, Russia

{andrey.sakhovskiy, zulfatmi, tutubalinaev}@gmail.com

Abstract

This paper describes neural models developed for the Social Media Mining for Health (SMM4H) 2021 Shared Task. We participated in two tasks on classification of tweets that mention an adverse drug effect (ADE) (Tasks 1a & 2) and two tasks on extraction of ADE concepts (Tasks 1b & 1c). For classification, we investigate the impact of joint use of BERT-based language models and drug embeddings obtained by chemical structure BERT-based encoder. The BERT-based multimodal models ranked first and second on classification of Russian (Task 2) and English tweets (Task 1a) with the F1 scores of 57% and 61%, respectively. For Task 1b and 1c, we utilized the previous year's best solution based on the EnDR-BERT model with additional corpora. Our model achieved the best results in Task 1c, obtaining an F1 of 29%.

1 Introduction

Text classification, named entity recognition, and medical concept normalization in free-form texts are crucial steps in every text-mining pipeline. Here we focus on discovering adverse drug effects (ADE) concepts in Twitter messages as part of the Social Media Mining for Health (SMM4H) 2021 Shared Task (Magge et al., 2021).

This work is based on the participation of our team in four subtasks of two tasks. Task 1 consists of three subtasks, namely 1a, 1b, and 1c each of which corresponds to classification, extraction, and normalization of ADEs. For Task 2, train, dev, and test sets include Russian tweets annotated with a binary label indicating the presence or absence of ADEs. For the 1b task, named entity recognition aims to detect the mentions of ADEs. Task 1c is designed as an end-to-end problem, intended to perform full evaluation of a system operating in real conditions: given a set of raw tweets, the system has to find the tweets that are mentioning ADEs, find the spans of the ADEs, and normalize them with respect to a given knowledge base (KB). These tasks are especially challenging due to specific characteristics of user-generated texts from social networks which are noisy, containing misspelled words, abbreviations, emojis, etc. The source code for our models is freely available[1].

The paper is organized as follows. We describe our experiments on the multilingual and multimodal classification of Russian and English tweets for the presence or absence of adverse effects in Section 2. In Section 3, we describe our pipeline for named entity recognition (NER) and medical concept normalization (MCN). Finally, we conclude this paper in Section 4.

2 Tasks 1a & 2: multilingual classification of tweets

The objective of Tasks 1a & 2 is to identify whether a tweet in English (Task 1a) or Russian (Task 2) mentions an adverse drug effect.

2.1 Data

For the English task, we used the original dev set provided by the organizers of the SMM4H 2021. For the Russian task, we sampled 1,000 non-repeating tweets from the original dev set as the new dev set and added the remaining tweets to the training set. Table 1 presents the statistics on Task 1a and Task 2 data. As can be seen from the table, the classes are highly imbalanced for both the English and the Russian corpora.

We preprocessed datasets for tasks 1a and 2 in a similar manner. During preprocessing, we: (i) replaced all URLs with the word "link"; (ii) replaced all user mentions with @username placeholder; (iii) replaced some emojis with a textual representation (e.g., laughing emojis with the word laughing; pill and syringe emojis with the corresponding

[1] https://github.com/Andoree/smm4h_2021_classification

Proceedings of the Sixth Social Media Mining for Health Workshop 2021, pages 39–43

Dataset	# Tweets	# Positive samples (ADE presence)
Task 1a (English tweets)		
Train	17,385	1,235 (7.1%)
Dev	914	65 (7.1%)
Test	10,984	–
Task 2 (Russian tweets)		
Train	10,609	980 (9.2%)
Dev	1,000	92 (9.2%)
Test	9,095	–

Table 1: Task 1a and Task 2 data statistics

words); (iv) replaced ampersand's HTML representation "&" with "&". As training sets are highly imbalanced, we applied the positive class over-sampling so that each training batch contained roughly the same number of positive and negative samples. However, we did not observe a significant performance improvement for the Russian subtask, so we applied the technique for the English subtask only. Following (Miftahutdinov et al., 2020), for Task 2, we combined the English and the Russian training sets.

2.2 Models

For our experiments, we used neural models based on Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) as they have achieved state-of-the-art results in the biomedical domain. In particular, BERT-based models proved efficient at the SMM4H 2020 Shared Task (Gonzalez-Hernandez et al., 2020). We used the following BERT-based models:

(1) RoBERTa$_{large}$[2] (Liu et al., 2019), a modification of BERT that is pretrained on 160GB of English texts with dynamic masking;

(2) EnRuDR-BERT[3] (Tutubalina et al., 2020), pretrained on English (Tutubalina et al., 2017) and Russian (Tutubalina et al., 2020) corpora of 5M health-related texts in English and Russian;

(3) ChemBERTa[4] (Chithrananda et al., 2020), a RoBERTa$_{base}$-based model that is pretrained on compounds from ZINC (Irwin and Shoichet, 2005) and is designed for drug design, chemical modelling and molecular properties prediction.

2.3 Experiments

For both tasks, we investigated the efficacy of the multimodal classification approach. For each tweet, we found its drug mentions, represented the chemical structure of each drug as a Simplified molecular-input line-entry system (SMILES) string, encoded the string using ChemBERTa, and took the final [CLS] embedding as drug embedding. Thus, we matched each tweet with a drug embedding. For tweets that contain no drug mentions, we encoded an empty string. We compared the following text-molecule combination strategies: (i) concatenation of the drug and the text embeddings, (ii) one cross-attention layer (Vaswani et al., 2017) from molecule encoder to text encoder. For concatenation architecture, we did not fine-tune ChemBERTa on the training set, whereas for cross-attention models, we trained both text and drug encoder.

For both Task 1a and Task 2, we adopted pretrained models from HuggingFace (Wolf et al., 2019) and fine-tuned them using PyTorch (Paszke et al., 2019). We trained each RoBERTa$_{large}$ model for 10 epochs with the learning rate of $1 * 10^{-5}$ using Adam optimizer (Kingma and Ba, 2014). We set batch size to 32 and maximum sequence size to 128. For EnRuDR-BERT we used the learning rate of $3 * 10^{-5}$, batch size to 64, and sequence to 128. For ChemBERTa, we used a sequence length of 256. For classification, we used a fully-connected network with one hidden layer, GeLU (Hendrycks and Gimpel, 2016) activation, a dropout probability of 0.3, and sigmoid as the final activation. To handle a high variance of BERT-based models' performance that varies across different initializations of classification layers, for each training setup, we trained 10 models and weighed their predictions. We tried two weighing strategies: (i) majority voting and (ii) sigmoid-based weighing. For (ii), we used predicted positive class probabilities to train a Scikit-learn's (Pedregosa et al., 2011) logistic regression on the validation set. For all experiments, we used a classification threshold of 0.5.

Table 2 shows the performance of our systems for Task 1a and Task 2 in terms of precision (P), recall (R), and F1-score (F1). Based on the results, we can draw the following conclusions. First, for the English task, the concatenation of text and chemical features increases the F1-score by 3%

[2]https://huggingface.co/roberta-large
[3]https://huggingface.co/cimm-kzn/enrudr-bert
[4]https://huggingface.co/seyonec/ChemBERTa_zinc250k_v2_40k

Model set-up	P	R	F1
Task 1a (English tweets)			
RoBERTa	–	–	0.58
RoBERTa + ChemBERTa concatenation	–	–	**0.61**
∗ RoBERTa + ChemBERTa concatenation + over-sampling	0.55	0.68	**0.61**
∗ RoBERTa + ChemBERTa concatenation + over-sampling + sigmoid	0.59	0.56	0.58
Average scores provided by organizers	0.51	0.41	0.44
Task 2 (Russian tweets)			
EnRuDR-BERT + Ru train	–	–	0.55
EnRuDR-BERT + Ru train + ChemRoBERTa + cross-attention	–	–	0.54
EnRuDR-BERT + RuEn train	–	–	0.54
∗ EnRuDR-BERT + RuEn train + ChemRoBERTa cross-attention	0.58	0.57	**0.57**
∗ EnRuDR-BERT + RuEn train + ChemRoBERTa cross-attention + sigmoid	0.77	0.35	0.48
Average scores provided by organizers	0.55	0.56	0.51

Table 2: Text classification results on the SMM4H 2021 Task 1a and Task 2 test sets. For all set-ups except the ones with "sigmoid", we used majority voting. Our official submissions for the SMM4H 2021 are denoted by ∗.

compared to text-only classification. Second, for the Russian task, neither the bilingual approach nor the use of chemical features shows a performance improvement when used separately, but the joint use of bilingual data and cross-modality with cross-attention results in an F1-score growth of 2% compared to text-only monolingual classification. Third, the results of this year showed a smaller gap between F1-scores on Russian and English test sets than last year.

3 Tasks 1b & 1c: extraction and normalization of ADEs

The 1b task's objective is to detect ADE mentions. Task 1c is designed as an end-to-end task. Systems have a free-form tweet as input and should be able to produce a set of extracted medical concepts. For this task, we develop a pipeline that (i) first detect ADE mentions and then (ii) link extracted ADEs to the concepts from the medical dictionary for regulatory activities (MedDRA) (Brown et al., 1999).

Following the best results in SMM4H 2020 Task 3 (Miftahutdinov et al., 2020), we utilize a EnDR-BERT model[5] with dictionary based features for the named entity recognition (NER) task. We adopted the dictionaries from (Miftahutdinov et al., 2017). As in the best solution of the SMM4H 2020 Task 3, we adopted extra training data for the NER task, we used the CSIRO Adverse Drug Event Corpus (CADEC) (Karimi et al., 2015) and COMETA corpus (Basaldella et al., 2020).

For the normalization task, we applied two models: (i) a classifier (Miftahutdinov et al., 2020; Miftahutdinov and Tutubalina, 2019), (ii) a novel neural model based on similarity distance of BERT vectors of concepts (Miftahutdinov et al., 2021). Following (Miftahutdinov et al., 2020), we utilize additional data for training. Other corpora are filtered to match a vocabulary of the SMM4H 2021 train set. We combined two models based on a threshold. For instance, given (i) prediction c_{bs} from from BERT-based similarity method with the distance equals to d and (ii) prediction c_{clf} from the classification approach, the final prediction is set to c_{bs}, if d is less than a threshold, and to c_{clf}, otherwise. For more detailed description of NER and end-to-end entity linking model please refer to (Miftahutdinov et al., 2020).

Table 3 shows a comparison of the model to the official average scores computed using the participants' submissions. Our NER model achieved below average results (40% vs 42%). We believe that the results are related to additional training of the model on non-target texts (reviews). Yet, with lower results in Task 1b and the top ranked results in Task 1c, it becomes clear that that the advantage of our pipeline is the two-component model for medical concept normalization. To sum up, the pipeline ranked first at SMM4H 2021 Task 1c and obtained the F1 score of 29% on extraction of MedDRA concepts.

[5]https://huggingface.co/cimm-kzn/endr-bert

Run name	P	R	F1
ADE Detection Evaluation (Task 1b)			
KFU NLP Team	0.42	0.38	0.40
Average scores	0.49	0.46	**0.42**
End-to-End Evaluation (Task 1c)			
KFU NLP Team	0.30	0.28	**0.29**
Average scores	0.23	0.22	0.22

Table 3: Performance of our models in SMM4H 2021 Task 1b and 1c (official results).

4 Conclusion

In this work, we have explored an application of domain-specific BERT models pretrained on health-related user reviews in English and Russian to the task of multilingual and multimodal text classification, extraction, and normalization of adverse drug effects. Our experiments show that multimodal architecture for classification of tweets outperforms other strong baselines and text classifiers. Besides, our BERT-based pipeline for extraction on MedDRA concepts ranked 1st in Task 1c.

We foresee two directions for future work. First, future research will explore how different drug representation models and pretraining approaches affect classification performance. Second, a potential direction is to verify the efficacy of multimodal classification for languages other than Russian and English.

Acknowledgements

The work on neural models has been supported by the Russian Science Foundation grant # 18-11-00284. Part of NER experiments was carried out by Elena Tutubalina on a lab computer (workstation) at HSE University within the framework of the HSE University Basic Research Program.

References

Marco Basaldella, Fangyu Liu, Ehsan Shareghi, and Nigel Collier. 2020. COMETA: A corpus for medical entity linking in the social media. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3122–3137, Online. Association for Computational Linguistics.

Elliot G Brown, Louise Wood, and Sue Wood. 1999. The medical dictionary for regulatory activities (meddra). *Drug safety*, 20(2):109–117.

Seyone Chithrananda, Gabe Grand, and Bharath Ramsundar. 2020. Chemberta: Large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885*.

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.

Graciela Gonzalez-Hernandez, Ari Z. Klein, Ivan Flores, Davy Weissenbacher, Arjun Magge, Karen O'Connor, Abeed Sarker, Anne-Lyse Minard, Elena Tutubalina, Zulfat Miftahutdinov, and Ilseyar Alimova, editors. 2020. *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task*. Association for Computational Linguistics, Barcelona, Spain (Online).

Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.

John J Irwin and Brian K Shoichet. 2005. Zinc- a free database of commercially available compounds for virtual screening. *Journal of chemical information and modeling*, 45(1):177–182.

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.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

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, Ari Klein, Ivan Flores, Ilseyar Alimova, Mohammed Ali Al-garadi, Antonio Miranda-Escalada, Zulfat Miftahutdinov, Eulàlia Farré-Maduell, Salvador Lima López, Juan M Banda, Karen O'Connor, Abeed Sarker, Elena Tutubalina, Martin Krallinger, Davy Weissenbacher, and Graciela Gonzalez-Hernandez. 2021. Overview of the sixth social media mining for health applications (# smm4h) shared tasks at naacl 2021. In *Proceedings of the Sixth Social Media Mining for Health Applications Workshop & Shared Task*.

Z.Sh. Miftahutdinov, E.V. Tutubalina, and A.E. Tropsha. 2017. Identifying disease-related expressions in reviews using conditional random fields. *Komp'juternaja Lingvistika i Intellektual'nye Tehnologii*, 1(16):155–166.

Zulfat Miftahutdinov, Artur Kadurin, Roman Kudrin, and Elena Tutubalina. 2021. Drug and disease interpretation learning with biomedical entity representation transformer. *Proceedings of the 43rd European Conference on Information Retrieval*.

Zulfat Miftahutdinov, Andrey Sakhovskiy, and Elena Tutubalina. 2020. KFU NLP team at SMM4H 2020 tasks: Cross-lingual transfer learning with pretrained language models for drug reactions. In *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task*, pages 51–56, Barcelona, Spain (Online). Association for Computational Linguistics.

Zulfat Miftahutdinov and Elena Tutubalina. 2019. Deep neural models for medical concept normalization in user-generated texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 393–399.

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 H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Elena Tutubalina, Ilseyar Alimova, Zulfat Miftahutdinov, Andrey Sakhovskiy, Valentin Malykh, and Sergey Nikolenko. 2020. The russian drug reaction corpus and neural models for drug reactions and effectiveness detection in user reviews. *Bioinformatics*. Btaa675.

EV Tutubalina, Z Sh Miftahutdinov, RI Nugmanov, TI Madzhidov, SI Nikolenko, IS Alimova, and AE Tropsha. 2017. Using semantic analysis of texts for the identification of drugs with similar therapeutic effects. *Russian Chemical Bulletin*, 66(11):2180–2189.

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 the 31st International Conference on Neural Information Processing Systems*, pages 6000–6010.

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*.