BOUN-TABI@SMM4H'22: Text-to-Text Adverse Drug Event Extraction with Data Balancing and Prompting

Gökçe Uludoğan* and **Zeynep Yirmibeşoğlu***

Department of Computer Engineering
Bogazici University
Istanbul 34342, Turkey
{gokce.uludogan, zeynep.yirmibesoglu}@boun.edu.tr

Abstract

This paper describes models developed for the Social Media Mining for Health 2022 Shared Task. We participated in two subtasks: classification of English tweets reporting adverse drug events (ADE) (Task 1a) and extraction of ADE spans in such tweets (Task 1b). We developed two separate systems based on the T5 model, viewing these tasks as sequence-to-sequence problems. To address the class imbalance, we made use of data balancing via over- and undersampling on both tasks. For the ADE extraction task, we explored prompting to further benefit from the T5 model and its formulation. Additionally, we built an ensemble model, utilizing both balanced and prompted models. The proposed models outperformed the current state-of-the-art, with an F1 score of 0.655 on ADE classification and a Partial F1 score of 0.527 on ADE extraction.

1 Introduction

This paper describes the models developed for the classification and extraction of Adverse Drug Event (ADE) mentions in English tweets (Tasks 1a and 1b) in the Social Media Mining for Health (SMM4H) 2022 Shared Task (Weissenbacher et al., 2022). We viewed the ADE classification and extraction tasks as sequence-to-sequence problems and fine-tuned the Text-to-Text Transfer Transformer (T5) model (Raffel et al., 2019), which is a pretrained model where the input and the output are both represented as text rather than class labels or spans.

Observing a significant class imbalance in the training data, we made use of over- and undersampling to balance out positive and negative class instances. After this operation, with an equal number of positive and negative instances (1:1 ratio), we obtained an 8% improvement in F1 score for ADE classification, and a 2.2% improvement in Strict F1 score for ADE extraction, showing the power of a balanced class distribution.

Prompt-based learning, or prompting is a method where the input and output text is modified with a string template such that training resembles language model training, where the model fills out missing information (finds the label) (Liu et al., 2021). We have made use of prompting for the ADE extraction task with three different templates, obtaining higher partial, but lower strict F1 scores compared to raw and balanced models. In the end, we have also put together the strengths of our balanced and prompted models in ensemble for the ADE extraction task.

2 Methodology

2.1 Data and Preprocessing

The task organizers provided the second version of the dataset presented in (Magge et al., 2021; Weissenbacher et al., 2022). For both tasks, we split the training data into a training set and a development set with a 80:20 ratio and used the validation set provided by the organizers to evaluate the models. We cleaned tweets with `preprocessor` library[1] and converted emojis to their text aliases using `emoji` library[2].

2.2 Model

In this work, the T5 model suggested by Raffel et al. (Raffel et al., 2019) is employed for ADE classification and extraction. This is an encoder-decoder model pre-trained on multiple tasks, whose architecture resembles that of the original Transformer (Vaswani et al., 2017), and the pre-training objective is to reconstruct randomly corrupted spans. The model can be fine-tuned on multiple tasks by adding a task-denoting prefix in front of the input text, such as "assert ade", or "ner ade". We fine-tuned the T5 model separately for the ADE clas-

*Equal contribution: order determined by a coin flip.

[1]https://github.com/s/preprocessor
[2]https://github.com/carpedm20/emoji

Proceedings of the 29th International Conference on Computational Linguistic, pages 31–34
October 12–17, 2022.

sification and extraction tasks with raw, balanced (over- and undersampled) and prompted datasets.

2.3 Data Balancing

Only 7.06% of the tweets in our training set contained ADE mentions (positive instances). To eliminate this imbalance between positive (ADE) and negative (noADE) instances, we oversampled the 982 ADE tweets up to 6463, and undersampled the 12926 noADE tweets by half using the `imbalanced-learn` library (Lemaître et al., 2017), such that there's a 1:1 ratio. As a second approach, we once again oversampled the 982 ADE tweets up to 6463, and used all of the noADE tweets, obtaining a 2:1 (noADE:ADE) ratio.

2.4 Prompting

Prompting is a method where templates are applied to input and output text for prediction tasks, so that the probability of text can directly be modeled (Liu et al., 2021). This method shows useful in low-resource scenarios with few labeled data (Schick and Schütze, 2021; Gao et al., 2021). Since there are only 982 instances of ADE mentions in the training set, we explored prompting for the ADE extraction task where the goal is to identify ADE mentions in tweets. To this end, we used three templates illustrated in Table 1 to transform data.

2.5 Ensemble Modeling

In neural network training, a change in seed, initialization of parameters or a slight change in the training set alters the outcome of the model. Therefore, we ensembled the predictions of our ADE extraction models, such that the strengths and weaknesses of the models can be compensated by each other. We first applied majority voting and chosen the span predicted at least half of the models for each tweet if there exists such a span. We combined the predictions of different models for the rest of tweets by taking the intersection of the predicted spans.

2.6 Implementation Details

We adapted the implementation[3] of Raval et al. (2021), and fine-tuned the T5 model separately for the classification and extraction tasks. Our models were trained on a single 12 GB GPU for 12 epochs. We took the maximum sequence length as 130, and batch size as 16. We made use of the AdamW optimizer (Loshchilov and Hutter, 2019) with an initial learning rate of *1e-4*. The source code is available at `https://github.com/gokceuludogan/boun-tabi-smm4h22`.

[3]https://github.com/shivamraval98/MultiTask-T5_AE

3 Results

3.1 Task 1a: ADE Classification

Model	Precision	Recall	F1
Raw data	0.75	0.69	0.72
Balanced data (1:1 ratio)	0.75	0.86	0.80
Balanced data (2:1 ratio)	0.73	0.86	0.79

Table 2: Comparison of models on ADE classification task (i.e. Task 1a) on the validation set

Table 2 presents the performance of different models on the validation set with respect to precision, recall and F1 metrics for the ADE class. Unsurprisingly, the models with balanced data outperform the model trained on raw data where tweets mentioning ADEs are rare. The model trained on the balanced data with 1:1 positive/negative ratio performed the best with an F1 score of 0.80 on the validation set. This model was then used to produce predictions for our official submission. Our model obtained higher scores compared to the mean of all submissions across metrics and achieved the state-of-the art result exceeding the performance reported in Magge et al. (2021) (see Table 3).

Model	Precision	Recall	F1
Our model	0.688	0.625	0.655
Mean	0.646	0.497	0.562
Magge et al. (2021)	0.61	0.64	0.63

Table 3: Comparison of our model with the mean of all submissions and the state-of-the-art (Magge et al., 2021) on ADE classification task (i.e. Task 1a) on the test set.

3.2 Task 1b: ADE Extraction

Results for the ADE extraction task on the validation set are reported in Table 4. The models are evaluated with two metrics, Strict F1 score where only perfect matches are considered as true matches and Partial F1 score in which partial matches (i.e. overlapping spans) are also taken into account. The models trained with balanced data achieved the best Strict F1 scores, yet they lag behind all prompting models in terms of Partial F1 score. The highest score in Partial F1 metric is obtained with the

Templates		Input	Output
T1	ADE noADE	Is there a negative drug effect in : [X]	[Y] is a negative drug effect. There isn't a negative drug effect.
T2	ADE noADE	Did the patient suffer from a side effect? [X]	Yes, the patient suffered from [Y]. No, the patient didn't suffer from a side effect.
T3	ADE noADE	[X] Did the patient suffer from a side effect?	Yes, the patient suffered from [Y]. No, the patient didn't suffer from a side effect.

Table 1: Prompting templates for the ADE extraction task given input [X] and output ADE span [Y]. T1, T2 and T3 denote Templates 1, 2 and 3.

Model	Partial F1	Strict F1
Raw data	0.605	0.481
Balanced data (1:1)	0.612	0.503
Balanced data (2:1)	0.639	0.482
Prompt/T1	0.636	0.424
Prompt/T2	0.662	0.408
Prompt/T3	0.638	0.393
Ensemble	0.657	0.500

Table 4: Comparison of models on ADE extraction task (i.e. Task 1b) on the validation set

Model	Partial			Strict		
	P	R	F1	P	R	F1
Our model	0.507	**0.549**	**0.527**	**0.384**	**0.412**	**0.398**
Mean	**0.539**	0.517	**0.527**	0.344	0.339	0.341
Magge et al. (2021)	0.53	0.38	0.44	-	-	-

Table 5: Comparison of our model with the mean of all submissions and the state-of-the-art (Magge et al., 2021) on ADE extraction task (i.e. Task 1b) on the test set. P and R denote precision and recall, respectively.

model using the second prompt template. However, this model's Strict F1 score is significantly lower than the best performing model (0.408 vs 0.503). The ensemble model addressing the performance gap between the best models with respect to Strict F1 and Partial F1 metrics combined the predictions of Balanced data (1:1) and Prompt/T2, and obtained the second best score across metrics. Observing the success of the ensemble model on both overlapping and strict metrics, we used its predictions as our official submission to Task 1b. As seen in Table 5, the model beats the state-of-the-art (Magge et al., 2021) with respect to Partial Recall and F1 metrics as well as obtaining significantly higher strict scores than the average of all official submissions on the test set. Yet, its Partial F1 score was at par with the other submissions, suggesting that our model achieves acceptable results for overlapping spans, but the strength of our model lies rather in its detection of perfect spans.

3.3 Analysis

We analyzed the performance of our submissions on the validation set provided by the organizers. We observed that our ADE classification model, trained on the balanced data with 1:1 ratio, performed fairly well in detecting ADE mentions with 56 correct predictions out of 65 ADE mentions. However, the model also made 19 incorrect ADE predictions (i.e. false positives). Similar results were also observed in our ensemble ADE extraction model. The model predictions included a lot of false positives (42) in contrast to relatively few false negatives (14).

4 Conclusion

In this study, we applied the T5 model and its problem formulation (i.e. viewing tasks as sequence-to-sequence problems) to ADE classification and extraction tasks. We used over- and undersampling to address class imbalance, a major challenge in ADE extraction from social media. In addition, we explored prompting methods to further benefit from the T5 language model and its sequence-to-sequence formulation. Our official submission for the ADE classification task made use of data balancing via over- and undersampling, and achieved an F1 score of 0.655, outperforming the state-of-the-art. For the ADE extraction task, we ensembled our balanced and prompted models to increase generalization for both partial and exact matches. Our submission of the ensemble model achieved a Partial F1 score of 0.527, beating the current state-of-the art and performing at par with the mean of all submissions. In the future, we intend to experiment with other methods of dealing with imbal-

anced data, such as data augmentation. The ADE classification and extraction tasks is also planned to be used in a multi-task learning scenario with the T5 model.

Acknowledgements

The models trained in this paper were fully performed at Boğaziçi University Telecommunications and Informatics Technologies Research Center servers (TETAM resources).

References

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830, Online. Association for Computational Linguistics.

Guillaume Lemaître, Fernando Nogueira, and Christos K. Aridas. 2017. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18(17):1–5.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Shivam Raval, Hooman Sedghamiz, Enrico Santus, Tuka Alhanai, Mohammad Ghassemi, and Emmanuele Chersoni. 2021. Exploring a unified Sequence-To-Sequence Transformer for medical product safety monitoring in social media. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3534–3546, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021. Few-shot text generation with natural language instructions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 390–402, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

Davy Weissenbacher, Ari Z. Klein, Luis Gascó, Darryl Estrada-Zavala, Martin Krallinger, Yuting Guo, Yao Ge, Abeed Sarker, Ana Lucia Schmidt, Raul Rodriguez-Esteban, Mathias Leddin, Arjun Magge, Juan M. Banda, Vera Davydova, Elena Tutubalina, and Graciela Gonzalez-Hernandez. 2022. Overview of the seventh social media mining for health applications #SMM4H shared tasks at COLING 2022. In *Proceedings of the Seventh Social Media Mining for Health (#SMM4H) Workshop and Shared Task*.