HaleLab_NITK@SMM4H'22: Adaptive Learning Model for Effective Detection, Extraction and Normalization of Adverse Drug Events from Social Media Data

Reshma Unnikrishnan, Sowmya Kamath S, Ananthanarayana V S
Department of Information Technology
National Institute of Technology Karnataka, Surathkal, Mangalore 575025, India
{reshmau.197it008, sowmyakamath, anvs} @nitk.edu.in

Abstract

This paper describes the techniques designed for detecting, extracting and normalizing adverse events from social data as part of the submission for the Shared task, Task 1-SMM4H'22. We present an adaptive learner mechanism for the foundation model to identify Adverse Drug Event (ADE) tweets. For the detected ADE tweets, a pipeline consisting of a pre-trained question-answering model followed by a fuzzy matching algorithm was leveraged for the span extraction and normalization tasks. The proposed method performed well at detecting ADE tweets, scoring an above-average F1 of 0.567 and 0.172 overlapping F1 for ADE normalization. The model's performance for the ADE extraction task was lower, with an overlapping F1 of 0.435.

1 Introduction

Social media data is extensively used for a wide variety of informed decision-making applications in varied domains like e-business, healthcare, policy making etc (Aichner et al., 2021; Unnikrishnan et al., 2021; Saini et al., 2022). Analyzing drug post-marketing management policies is a major aspect of pharmacovigilance (Nikfarjam et al., 2015) and has received significant research attention in recent years. As part of the Shared task, participants were provided with Twitter data (Magge et al., 2021b; Weissenbacher et al., 2022) that includes ADE mentions for detection, extraction and normalization purposes to aid pharmacovigilance studies. While detecting adverse reactions from social data having informal language is a complex task, the availability of valid ADE samples is scarce, making the data highly imbalanced (Klein et al., 2020; Magge et al., 2021a; Weissenbacher et al., 2022). Each of the tasks are correlated, thus, it is necessary to ensure that authentic ADE tweet samples are detected before utilizing them for ADE extraction and normalization processes. With these objectives, our work encompasses the design of an adaptive learning mechanism for the foundation model for effective ADE identification, extraction and normalization.

2 Methodology

Extracting ADEs from informal social data is challenging due to the short text length and style-variant nature (Weissenbacher et al., 2022). As stated earlier, we focused on developing an efficient adaptive learning method for foundation models due to the association between the three subtasks. We adopted RoBERTa (Liu et al., 2019) as the foundation model for applying adaptive learning techniques. The empirical and theoretical aspects of the data-specific features learnt by the higher-order layers were explored, for understanding the contribution of different layers in a foundation model and relearning task-specific requirements.

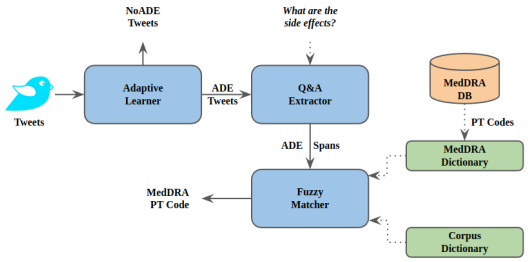

Figure 1: Proposed Workflow

During knowledge distillation, it is essential to ensure that the pre-trained weights are not affected by learning divergence, due to absence of previously learned weights. With this in mind, we employed mixout (Lee et al., 2019) with the weight reinitialization technique (Li et al., 2020), which helps adaptively retrain the last few layers. While mixout helps retain the weights from the base foundation model without losing the pre-trained features, weight reinitialization helps relearn these specific weights for the higher order layers to pre-

Proceedings of the 29th International Conference on Computational Linguistic, pages 95–97
October 12–17, 2022.

Table 1: Results on Test set

Task		Precision	Recall	F1
Mean	1a	0.646	0.497	0.562
Ours	1a	**0.674**	0.489	**0.567**

Approach	Task	Overlapping			Strict		
		Precision	*Recall*	*F1*	*Precision*	*Recall*	*F1*
Mean	1b	0.539	0.517	0.527	0.344	0.339	0.341
	1c	0.120	0.112	0.116	0.085	0.082	0.083
Ours	1b	**0.562**	0.354	0.435	0.194	0.114	0.144
	1c	**0.232**	**0.137**	**0.172**	**0.088**	0.052	0.065

dict ADE tweets.

The span extraction task was approached as a Question Answering (QA) task, rather than as a conventional named entity recognition (NER) task (Dima et al., 2021; Balumuri et al., 2021). To extract the side effects or adverse medical events from the Adaptive Learner output, Roberta-base (Liu et al., 2019) fine-tuned on the SQuAD2.0 dataset (Rajpurkar et al., 2018) was trained for a batch size of 96 for two epochs, a learning rate of 3e-5, handling a maximum sequence length of 386 and query length of 96. Since all the spans here represent ADEs on individuals, we considered a fixed query ("*What are the side effects?*") to extract spans from ADE tweets.

Due to fewer number of ADE tweets with varied MedDRA codes[1](Mozzicato, 2009) (no duplicates), formulating it as a supervised classification problem is practically unfeasible. In view of this, the conventional fuzzy matching algorithm built on Levenshtein distance (Yujian and Bo, 2007) based similarity measure was used for mapping the extracted spans to MedDRA codes. We created two dictionaries: MedDRA terms and their codes (Preferred Terms - PT) from the MedDRA database and the dictionary from the gold standard train data with spans and MedDRA codes. The spans extracted from the QA model were matched against the two dictionary terms, and the MedDRA code for one that weighted more was used as the label for the ADE spans.

3 Experiments and Observation

The SMM4H'22 task organizers provided data (Magge et al., 2021b; Weissenbacher et al., 2022) for this experimentation. We observed that the data was highly imbalanced along with a unique set of ADE events, thus increasing the learning complexity of a system. For ADE detection, experimentation on weight retraining varied between 1 to 4 layers and on mixout from 0.5 to 0.7 were performed. Empirically, we chose to reinitialize the top 4 layers by fixing the mixout to 0.7. While approaching the entity recognition task as QA, we found that it returned context chunks comprising adverse events rather than short entities. Since the MedDRA database holds proper clinical expressions, we believe that including the gold standard data covering the colloquial mentions of adverse effects as another dictionary helps improve the span normalization.

Table 2 presents the observed results for all three subtasks on the validation set, which achieved promising performance in terms of precision, recall and F1 score for tasks 1a and 1b. Task 1c falls into the 0.2 range for all, which is still better than all system submissions' baseline mean average overlapping score on test data. While observing the test data scores in Table 1, task 1a attains above average F1 score of 0.567 and an improved overlapping precision, recall and F1 score for task 1c. In contrast, task 1b exhibits degraded performance across all measures in the test set compared to the validation data performance. As stated earlier, the measurement fails to give a better score in test data due to the return of context chunks rather than word entities from the QA model. For ensuring reproducibility of the results, the source code of the proposed approach is made available publicly[2].

4 Conclusion

For addressing the SMM4H'22 Shared task 1, we concentrated on developing an adaptive learning technique for the foundation model using Roberta

[1]https://www.meddra.org/

[2]https://github.com/Reshma-U/SMM4H-22

Table 2: Results on Validation set

Shared Task	Precision	Recall	F1
Task 1a	0.71	0.72	0.72
Task 1b	1.00	0.63	0.77
Task 1c	0.24	0.22	0.21

base for ADE detection. While achieving a promising F1 measure of 0.72 on the validation set, the same model on the test set produced an above-average F1 of 0.567. We approached ADE span extraction as a QA task that showed degraded results due to the return of contextual chunks rather than word entities. Fuzzy matching for span normalization gave above-average overlapping P, R and F1 scores. In future, we plan to fine-tune the QA model to return adverse word level entities for improved performance in the span prediction task.

References

Thomas Aichner, Matthias Grünfelder, Oswin Maurer, and Deni Jegeni. 2021. Twenty-five years of social media: a review of social media applications and definitions from 1994 to 2019. *Cyberpsychology, behavior, and social networking*, 24(4):215–222.

Spandana Balumuri, Sony Bachina, and Sowmya Kamath. 2021. Sb_nitk at mediqa 2021: Leveraging transfer learning for question summarization in medical domain. In *Proceedings of the 20th workshop on biomedical language processing*, pages 273–279.

George-Andrei Dima, Dumitru-Clementin Cercel, and Mihai Dascalu. 2021. Transformer-based multi-task learning for adverse effect mention analysis in tweets. In *Sixth Social Media Mining for Health (# SMM4H) Workshop and Shared Task*, pages 44–51.

Ari Klein, Ilseyar Alimova, Ivan Flores, Arjun Magge, Zulfat Miftahutdinov, Anne-Lyse Minard, Karen O'connor, Abeed Sarker, Elena Tutubalina, Davy Weissenbacher, et al. 2020. Overview of the fifth social media mining for health applications (# smm4h) shared tasks at coling 2020. In *Fifth Social Media Mining for Health Applications Workshop & Shared Task*, pages 27–36.

Cheolhyoung Lee, Kyunghyun Cho, and Wanmo Kang. 2019. Mixout: Effective regularization to finetune large-scale pretrained language models. *arXiv preprint arXiv:1909.11299*.

Xingjian Li, Haoyi Xiong, Haozhe An, Cheng-Zhong Xu, and Dejing Dou. 2020. Rifle: Backpropagation in depth for deep transfer learning through re-initializing the fully-connected layer. In *International Conference on Machine Learning*, pages 6010–6019. PMLR.

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, Antonio Miranda-Escalada, Eulalia Al-Garadi, Farre, Salvador Lima-López, et al. 2021a. Overview of the sixth social media mining for health applications (# smm4h) shared tasks at naacl 2021. In *Sixth Social Media Mining for Health (# SMM4H) Workshop and Shared Task*, pages 21–32.

Arjun Magge, Elena Tutubalina, Zulfat Miftahutdinov, Ilseyar Alimova, Anne Dirkson, Suzan Verberne, Davy Weissenbacher, and Graciela Gonzalez-Hernandez. 2021b. 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.

Patricia Mozzicato. 2009. Meddra. *Pharmaceutical Medicine*, 23(2):65–75.

Azadeh Nikfarjam, Abeed Sarker, Karen O'connor, Rachel Ginn, and Graciela Gonzalez. 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*, 22(3):671–681.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*.

Gurdeep Saini, Naveen Yadav, and Sowmya Kamath S. 2022. Ensemble neural models for depressive tendency prediction based on social media activity of twitter users. In *Security, Privacy and Data Analytics*, pages 211–226. Springer.

Reshma Unnikrishnan, S Sowmya Kamath, and VS Ananthanarayana. 2021. Benchmarking shallow and deep neural networks for contextual representation of social data. In *2021 IEEE 18th India Council International Conference (INDICON)*, pages 1–8. IEEE.

Davy Weissenbacher, Ari Z. Klein, Luis Gascó, Darryl Estrada-Zavala, 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 *Seventh Social Media Mining for Health (#SMM4H) Workshop and Shared Task*.

Li Yujian and Liu Bo. 2007. A normalized levenshtein distance metric. *IEEE transactions on pattern analysis and machine intelligence*, 29(6):1091–1095.