BIT@UA at #SMM4H 2024 Tasks 1 and 5: finding adverse drug events and children's medical disorders in English tweets

Luís Carlos Afonso 0009-0005-6728-3089, **João Rafael Almeida** 0000-0003-0729-2264, **Rui Antunes** 0000-0003-3533-8872, and **José Luís Oliveira** 0000-0002-6672-6176

IEETA/DETI, LASI, University of Aveiro, Aveiro, Portugal

Abstract

In this paper, we present our proposed systems, for Tasks 1 and 5 of the #SMM4H-2024 shared task (Social Media Mining for Health), responsible for identifying health-related aspects in English social media text. Task 1 consisted of identifying text spans mentioning adverse drug events and linking them to unique identifiers from the medical terminology MedDRA, whereas in Task 5 the aim was to distinguish tweets that report a user having a child with a medical disorder from tweets that merely mention a disorder.

For Task 1, our system, composed of a pre-trained RoBERTa model and a random forest classifier, achieved 0.397 and 0.295 entity recognition and normalization F1-scores respectively. In Task 5, we obtained a 0.840 F1-score using a pre-trained BERT model.

1 Introduction

Social media text, such as tweets from Twitter, holds a vast amount of textual information and can be a solid source for clinical findings (Dreisbach et al., 2019). Several research initiatives have been pursued to promote the development of data mining solutions from social media to foster healthier lives (Weissenbacher et al., 2018, 2019). Text from social media has been used by biomedical NLP (Natural Language Processing) researchers for different purposes including sentiment analysis (Yang et al., 2016), disease normalization (Tutubalina et al., 2018), suicide attempt prediction (Coppersmith et al., 2016), and classification of depression users (Trifan et al., 2020).

The 9th Social Media Mining for Health Research and Applications (#SMM4H-2024) Workshop continues this research endeavor promoting seven different tasks (Xu et al., 2024). In this work, we describe our participation in the #SMM4H 2024 shared task where we present our developed systems for Tasks 1 and 5, both consisting in mining English tweets from Twitter.

In Sections 2 and 3 we present the datasets in use and the methodology followed, and discuss the results obtained during the official challenge for Tasks 1 and 5 respectively. Finally, we draw some conclusions in Section 4 and present future lines of research.

2 Task 1: adverse drug events

In Task 1, participants had to develop systems to automatically identify mentions of Adverse Drug Events (ADEs) and link them to unique identifiers from the standard terminology MedDRA (Fescharek et al., 2014). Past research work also explored the identification of ADEs from medical case reports and social media (Gurulingappa et al., 2012; Liu and Chen, 2015).

2.1 Dataset

The dataset is composed of a few files annotated with entity mentions, representing adverse drug events, linked with unique MedDRA identifiers. All of the ADE annotations have an associated text span (character start and end offsets) as well as the unique MedDRA identifier. A unique tweet identifier is also used to specify a tweet within the dataset.

The organizers split the dataset into three subsets—training, development, and testing. During the challenge, participants had access to the training and development subsets, containing tweets annotated with ADE mentions, to develop their systems. Table 1 presents dataset statistics.

The training subset contains 18 185 tweets of which only 1 239 tweets contain ADE annotations, and 16 946 tweets do not contain any entity mention. It is annotated with a total of 1 711 entities meaning that some of the tweets have more than one ADE mention. The development subset fol-

Proceedings of The 9th Social Media Mining for Health Applications (#SMM4H) Workshop and Shared Tasks, pages 158–162
August 15, 2024 ©2024 Association for Computational Linguistics

Table 1: Task 1 dataset statistics.

	Training	Development	Testing*
# Tweets	18 185	965	11 799
with entities	1239	65	—
# Entities	1711	87	—

* Participants had no access to the gold standard entities in the testing subset during the challenge.

lows a similar distribution containing a total of 965 tweets and 87 entity annotations where only 65 tweets have at least one ADE annotation. For the challenge official evaluation, participants had to submit their predictions on the blind testing subset, composed of 11 799 tweets.

2.2 Method

In this subsection we detail our approach for detecting adverse drug events in tweets. Our strategy was based on a two-phase workflow where we first identify the spans of ADE mentions and then link them to MedDRA identifiers.

2.2.1 Named entity recognition

We treated the problem of detecting ADE mentions in tweets as a sequence labeling task. For simplicity, each token is attributed an I (Inside) or O (Outside) tag to specify if a token belongs to an ADE mention or not. Contrarily to the standard IOB (Inside, Outside, Beginning) tagging scheme, we employed this IO simpler scheme since the mentions were scarce and no adjacent entities were found in the training and development subsets.

Then, we experimented different machine learning classifiers for performing this token-level classification and obtained the best preliminary results using a RoBERTa model[1] that was (i) pre-trained in PubMed abstracts, (ii) trained on datasets annotated with mentions of diseases (Liu et al., 2019), and (iii) then fine-tuned, by us, using the training data for detecting ADEs.

We also experimented applying a filtering stage as a binary document classification task, before the NER module, to remove documents that did not contain ADEs but this did not prove beneficial and was therefore discarded.

2.2.2 Entity normalization

The final step in our pipeline involved entity normalization, responsible for linking the detected ADEs to unique MedDRA identifiers. This task was tackled as a classification problem where each detected entity needed to be assigned the correct MedDRA identifier.

Prior to classification, we applied standard preprocessing techniques to the textual data such as lowercasing, stop words removal, stemming, and lemmatization. We used a combination of features for the classification task, including token counts and TF-IDF (term frequency–inverse document frequency) features calculated using the the scikit-learn library (Pedregosa et al., 2011).

After experimenting with various machine learning models, also from scikit-learn, our final system employed a random forest classifier with 10 estimators. This choice was based on its superior performance on the development subset compared to other tested classifiers, and offered a good balance between computational efficiency and classification accuracy for this particular task. The classifier was trained on the provided training data, learning to map textual representations of ADEs to their corresponding MedDRA identifiers.

2.3 Results and discussion

During the implementation phase, we obtained an F1-score of 0.453 for entity normalization on the development subset. From Table 2 we observe that our normalization result, on the testing subset, deteriorated considerably (0.453 vs 0.295). We suspect that one of the reasons for this performance drop was because our model was only able to assign identifiers that were seen during the training phase.

We noticed that one of the main challenges in mapping detected entities to MedDRA identifiers was dealing with the variability in how ADEs are expressed in social media text and informal writing. This includes handling synonyms, abbreviations, and misspelled terms that may all refer to the same underlying medical concept.

Table 2: Task 1 official results on the testing subset. F1-score metric is employed. Norm.: entity normalization.

	NER	Norm
Our submission	0.397	0.295
Mean*	0.327	0.283
Median*	0.376	0.293
Baseline* (Magge et al., 2021)	0.481	0.439

* The organizers shared the results of a baseline model, and the mean and median of all submissions by the participating teams.

[1] https://huggingface.co/raynardj/ner-disease-ncbi-bionlp-bc5cdr-pubmed

3 Task 5: children's medical disorders

In Task 5, participants were asked to develop a system to identify tweets that mention a user having a child with a medical disorder, from tweets solely referring a disorder. This was considered a binary text classification task.

3.1 Dataset

The dataset is composed of a few files with tweets classified with a gold standard label that is either 0 or 1. A label of 1 represents a *true case* for this task, meaning that the linked tweet mentions a user having a child with a medical disorder, and a label of 0 represents the opposite scenario (*negative case*). Each tweet also has associated an unique identifier that represents the tweet uniquely within the dataset.

The organizers split the dataset in three subsets, publicly distributing the gold standard labels for the training and development subsets, while keeping the labels for the testing subset unknown for the participants.

Some statistics can be seen in Table 3 about each split of the original dataset where 'Positive' and 'Negative' refer to tweets with a label of 1 and 0 respectively. Participants had to submit their predictions for the unlabeled testing subset containing 10 000 tweets.

Table 3: Task 5 dataset statistics.

	Training	Development	Testing*
# Tweets	7398	389	10 000
Positive	2280	135	—
Negative	5118	254	—

* Participants had no access to the gold standard labels in the testing subset during the challenge.

3.2 Method

Here we detail our approach for detecting tweets posted by users that report having a child with a disorder. Our strategy was based on the assumption that the task can be tackled as a simple text classification problem.

We experimented with different traditional models from the scikit-learn library (Pedregosa et al., 2011)—naive Bayes, SVM with a linear kernel, Logistic Regression—and XGBoost (Chen and Guestrin, 2016). For text representation we tried two different well-known approaches, available in scikit-learn, for converting a collection of text documents:

1. Count vectorizer—to obtain a matrix of token counts; and
2. TF-IDF vectorizer—to obtain a matrix of term frequency–inverse document frequency features.

In a later stage we employed a pre-trained BERT variant[2] to inspect how a more complex model would compare (Devlin et al., 2019).

3.3 Results and discussion

The results for all the aforementioned models, varying the vectorizer (text representation features) for the traditional classifiers and the number of epochs for the BERT model, are presented in Table 4 and Table 5 respectively. As one can observe, the BERT model achieved the best results after being fine-tuned for at least 4 epochs after which the performance changes were not significant.

For the official submission, with the final predictions on the blind testing subset, we employed the BERT model fine-tuned for 8 epochs since it achieved the highest preliminary result on the development subset (0.8968 F1-score).

Table 6 presents the official challenge results on the blind testing subset where we observe that

[2] https://huggingface.co/google-bert/bert-base-uncased

Table 4: Task 5 results with traditional classifiers by applying 5-fold cross-validation on the training subset.

Classifier	Vectorizer	F1-score	Accuracy
Naive Bayes	Count	0.6833	0.7693
	TF-IDF	0.4145	0.6934
SVM	Count	0.7117	0.7814
	TF-IDF	0.7057	0.7830
Logistic Regression	Count	0.7255	0.7776
	TF-IDF	0.7085	0.7813
XGBoost	Count	0.7431	0.7922
	TF-IDF	0.7358	0.7862

Table 5: Task 5 results with a BERT model, fine-tuned for different numbers of epochs, by applying 5-fold cross-validation on the training subset.

Epochs	F1-score	Accuracy
2	0.8216	0.8592
4	0.8499	0.8778
8	0.8428	0.8628
16	0.8467	0.8708
32	0.8520	0.8755

Table 6: Task 5 official results on the testing subset.

	F1-score
Our submission	0.840
Mean*	0.822
Median*	0.901

* The organizers shared the mean and median results of all submissions by the participating teams.

despite our system achieved an F1-score 1.8 percentage points above the mean result it lags behind the median result, showing that there is significant room for improvement.

4 Conclusions

In this work, we presented machine learning models to detect ADEs (Task 1) and users reporting having children with medical disorders in English tweets (Task 5). We obtained more competitive results in Task 1 being slightly above the median. In Task 5, our classification F1-score was 6.1 percentage points below the median result demonstrating that our approach still holds great potential for improvement.

In Task 1, the most relevant aspect to enhance would be for our system to be able to link ADEs to identifiers that were not seen during the training phase. Such system should be able to consult the full MedDRA terminology and normalize any ADE mention to the respective identifier. We also hypothesize that the adoption of the BIO tagging scheme for NER (Task 1) could be beneficial and a more careful hyperparameter optimization through grid search could improve results on both tasks.

From our experiments, we conclude that BERT-based models achieved the best performance in entity recognition and document classification proving to be on par with the state-of-the-art.

5 Funding

This work has received support from the EU/EFPIA Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No 806968. Rui Antunes is funded under the project UIDB/00127/2020[3].

References

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: a scalable tree boosting system. In *22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, San Francisco, California, USA. ACM.

Glen Coppersmith, Kim Ngo, Ryan Leary, and Anthony Wood. 2016. Exploratory analysis of social media prior to a suicide attempt. In *Third Workshop on Computational Linguistics and Clinical Psychology*, pages 106–117, San Diego, California, USA. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *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, USA. Association for Computational Linguistics.

Caitlin Dreisbach, Theresa A. Koleck, Philip E. Bourne, and Suzanne Bakken. 2019. A systematic review of natural language processing and text mining of symptoms from electronic patient-authored text data. *International Journal of Medical Informatics*, 125:37–46.

Reinhard Fescharek, Jürgen Kübler, Ulrich Elsasser, Monika Frank, and Petra Güthlein. 2014. Medical dictionary for regulatory activities (MedDRA). *International Journal of Pharmaceutical Medicine*, 18(5):259–269.

Harsha Gurulingappa, Abdul Mateen Rajput, Angus Roberts, Juliane Fluck, Martin Hofmann-Apitius, and Luca Toldo. 2012. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *Journal of Biomedical Informatics*, 45(5):885–892.

Xiao Liu and Hsinchun Chen. 2015. A research framework for pharmacovigilance in health social media: Identification and evaluation of patient adverse drug event reports. *Journal of Biomedical Informatics*, 58:268–279.

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: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.

[3] https://doi.org/10.54499/UIDB/00127/2020

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Alina Trifan, Rui Antunes, Sérgio Matos, and Jose Luís Oliveira. 2020. Understanding depression from psycholinguistic patterns in social media texts. In *42nd European Conference on Information Retrieval*, pages 402–409, Online. Springer Nature.

Elena Tutubalina, Zulfat Miftahutdinov, Sergey Nikolenko, and Valentin Malykh. 2018. Medical concept normalization in social media posts with recurrent neural networks. *Journal of Biomedical Informatics*, 84:93–102.

Davy Weissenbacher, Abeed Sarker, Arjun Magge, Ashlynn Daughton, Karen O'Connor, Michael J. Paul, and Graciela Gonzalez-Hernandez. 2019. Overview of the Fourth Social Media Mining for Health (SMM4H) Shared Tasks at ACL 2019. In *Fourth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 21–30, Florence, Italy. Association for Computational Linguistics.

Davy Weissenbacher, Abeed Sarker, Michael J. Paul, and Graciela Gonzalez-Hernandez. 2018. Overview of the Third Social Media Mining for Health (SMM4H) Shared Tasks at EMNLP 2018. In *2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task*, pages 13–16, Brussels, Belgium. Association for Computational Linguistics.

Dongfang Xu, Guillermo Lopez Garcia, Lisa Raithel, Roland Roller, Philippe Thomas, Eiji Aramaki, Shuntaro Yada, Pierre Zweigenbaum, Sai Tharuni Samineni, Karen O'Connor, Yao Ge, Sudeshna Das, Abeed Sarker, Ari Klein, Lucia Schmidt, Vishakha Sharma, Raul Rodriguez-Esteban, Juan Banda, Ivan Flores Amaro, Davy Weissenbacher, and Graciela Gonzalez-Hernandez. 2024. Overview of the 9th Social Media Mining for Health (#SMM4H) Research and Applications Workshop and Shared Tasks at ACL 2024. In *Proceedings of The 9th Social Media Mining for Health Research and Applications Workshop and Shared Tasks*, Bangkok, Thailand. Association for Computational Linguistics.

Fu-Chen Yang, Anthony J.T. Lee, and Sz-Chen Kuo. 2016. Mining health social media with sentiment analysis. *Journal of Medical Systems*, 40(11):236.