LHS712NV at #SMM4H 2024 Task 4: Using BERT to classify Reddit posts on non-medical substance use

Valeria Fraga,[1] Neha Nair,[1] Dalton Simancek,[2] V.G.Vinod Vydiswaran[2,1]
[1]School of Information, [2]Department of Learning Health Sciences
University of Michigan, Ann Arbor
{vfraga, nehakkn, daltonsi, vgvinodv}@umich.edu

Abstract

This paper summarizes our participation in the Shared Task 4 of #SMM4H 2024. Task 4 was a named entity recognition (NER) task identifying clinical and social impacts of non-medical substance use in English Reddit posts. We employed the Bidirectional Encoder Representations from Transformers (BERT) model to complete this task. Our team achieved an F1-score of 0.892 on a validation set and a relaxed F1-score of 0.191 on the test set.

1 Introduction

Substance use, whether prescription or illicit, poses a significant public health challenge, contributing to addiction, overdose, and various associated health concerns (Lo et al., 2020 Apr). Understanding the clinical and social ramifications of non-medical substance use is crucial for enhancing the treatment of substance use disorder, informing intervention strategies, and developing preventive measures (Xu et al., 2024). This paper addresses a named entity recognition (NER) task focused on identifying two key entity types: clinical and social impacts. Clinical impacts refer to the effects of substance use on individuals' health and well-being, while social impacts encompass broader societal consequences (Xu et al., 2024).

The Social Media Mining for Health (#SMM4H) shared tasks aim to leverage natural language processing (NLP) techniques to extract valuable health insights from social media data. In #SMM4H 2024, we participated in Task 4 on extraction of the clinical and social impacts of non-medical substance use from Reddit posts. We were particularly motivated to participate in this task because of the pressing public health concern surrounding non-medical substance use. Understanding the clinical and social impacts of such usage is crucial for developing effective interventions and prevention strategies. By leveraging advanced, deep-learning based NLP models, we aimed to contribute to this understanding and advance the field of health informatics through innovative data analysis techniques.

2 System Description

2.1 Data Preprocessing

The #SMM4H 2024 Task 4 is a named entity recognition task with a goal to identify two entities – *Clinical Impacts* and *Social Impacts* – in Reddit posts. Individual sequences of tokens are tagged with one of these two labels or *No Label (-)* to denote neither of the two desired classes. (Ge et al., 2024) To perform this task, we use the Bidirectional Encoder Representations from Transformers (BERT) model (Devlin et al., 2019), which is highly effective at capturing contextual information from both the previous and subsequent sequences of text tokens. The process begins with data preprocessing, where the dataset is first split into training and test sets. Then, the text is tokenized, with each token being assigned one of the following labels: *Clinical Impacts*, *Social Impacts*, or *No Label (-)*.

2.2 BERT Fine-tuning and Evaluation

The model is then fine-tuned using the HuggingFace Transformers library, with training parameters such as batch size, learning rate, and the number of epochs specified. The selection of BERT for this task is underscored by its ability to comprehend the contextual nuances of words, enabling accurate identification of named entities. Fine-tuning pre-trained BERT models often results in enhanced performance on downstream tasks such as NER, making it a popular choice in natural language processing applications.

During training, four performance metrics, viz. precision, recall, F1 score, and accuracy, are computed to assess the model's performance. Subsequently, the model is evaluated on a separate validation dataset to gauge its effectiveness. Predictions are generated, and the four metrics are calculated and displayed, along with a visualization

Proceedings of The 9th Social Media Mining for Health Applications (#SMM4H) Workshop and Shared Tasks, pages 146–148

Epoch	Precision	Recall	F1	Accuracy
1	0.929	0.988	0.958	0.929
2	0.938	0.983	0.960	0.937
3	0.940	0.980	0.960	0.938

Table 1: Model evaluation of training data

Run	Precision	Recall	F1	Accuracy
1	0.892	0.892	0.892	0.892

Table 2: Model evaluation of validation data

of the confusion matrix. Finally, the trained model is tested on a separate dataset, and the predictions are saved for further analysis or task submission.

3 Results

3.1 Model Performance on training and validation

The model's performance was evaluated using training and validation datasets. The training phase comprised three epochs, each reporting training loss, validation loss, precision, recall, F1 score, and accuracy metrics (Table 1). Additionally, results on the validation data (performance metrics in Table 2 and confusion matrix in Figure 1) provided further insights into the model's performance. These results showed that the model did not identify any entities in the Social Impacts class and tended to label entities to the majority class (No label).

3.2 Test Data Analysis

Although the model seemed to perform well overall with the training and validation data, it did not perform as well with the test data. With the test data, the model had a token-level F1 score of 17.15% and a relaxed F1 score of 19.12%.

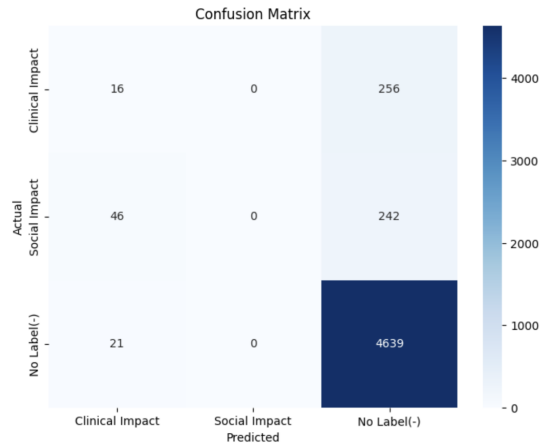

Figure 1: Confusion matrix of validation data

Labels	Accuracy (%)
No label	99.5
Clinical Impacts	5.9
Social Impacts	0.0

Table 3: The percent accuracy of the model's predicted labels

These findings collectively underscore the model's efficacy in classifying instances with *No label (-)* accurately while indicating room for improvement, particularly in distinguishing *Social Impact* instances (Table 3). Further optimization efforts may be warranted to enhance the model's performance across all classes.

4 Conclusion

This paper discusses our submission for #SMM4H 2024 Task 4 on named entity recognition (NER) for identifying clinical and social impacts of substance use in Reddit posts using Large Language Models (LLMs), particularly the BERT model. While the BERT model demonstrated high overall performance, there were notable disparities in its efficacy across different metrics and datasets. Challenges remain in accurately identifying *Social Impact* instances, highlighting the need for further refinement. Despite these challenges, the BERT model shows promise in automating the detection of clinical impacts in Reddit posts. To facilitate continued development, we have shared the implementation code for our system participation on GitHub.[1]

Limitations

In spite of the promising results obtained, it is crucial to acknowledge several limitations inherent in our study. The utilization of Reddit data presents inherent unpredictability due to the dynamic and unregulated nature of user-generated content. This variability may introduce noise and biases into our model, potentially impacting its generalizability to broader contexts. The size of the training data poses a potential constraint. With only 800 posts available for training, the dataset may not capture the full spectrum of patterns and nuances present in the data, thereby limiting the model's ability to discern complex relationships and generalize effectively. Inconsistencies in model performance raise concerns about reliability and robustness. Despite employing advanced deep-learning

[1]The software code, written in Python, is available at: https://github.com/NLP4HealthUMich/SMM4H2024-Task4

based techniques such as the BERT model, we observed fluctuating F1 scores, ranging from 0.4 to 0.9. These inconsistencies suggest potential issues in model stability or sensitivity to varying conditions, warranting further investigation and refinement. Finally, we did not use the "entity or not" column in the training of our model or investigate the results of other models besides the BERT model, which could have added much-needed context to the model and resulted in better performance. In light of these limitations, a cautious interpretation of the results is advised. Future research should address these constraints to enhance the validity and applicability of the findings and explore the performance of other models on the data.

Ethics Statement

Firstly, we acknowledge that the dataset provided to us was already anonymized. We prioritize the privacy and confidentiality of Reddit users, and we are committed to maintaining the anonymity of individuals whose posts are included in our dataset. We are dedicated to ensuring that our research contributes positively to understanding and mitigating substance use disorders. By identifying the clinical and social impacts of substance use, we aim to provide valuable insights that can inform intervention strategies and support efforts to address this pressing public health issue.

The Reddit posts for this study were sourced through the #SMM4H 2024 shared task coordinators and were accessed under the guidance of an academic advisor, strictly for research purposes. The data was exclusively utilized for participation in #SMM4H 2024 Task 4 and no other uses. We acknowledge our study's limitations, including the potential biases and uncertainties inherent in working with anonymized data. We remain transparent about these limitations and encourage future research to address them through robust methodologies and data validation techniques.

Overall, our research is conducted with the utmost integrity and respect for ethical considerations. We are dedicated to advancing our understanding of natural language processing techniques for health insights while upholding ethical standards and promoting the well-being of individuals and communities affected by substance use.

References

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.

Yao Ge, Sudeshna Das, Karen O'Connor, Mohammed Ali Al-Garadi, Graciela Gonzalez-Hernandez, and Abeed Sarker. 2024. Reddit-impacts: A named entity recognition dataset for analyzing clinical and social effects of substance use derived from social media. DOI:10.48550/arXiv.2405.06145.

T. Wing Lo, Jerf W. K. Yeung, and Cherry H. L. Tam. 2020 Apr. Substance abuse and public health: A multilevel perspective and muliple responses. *Int J Environ Res Public Health*, 17(7):2610. DOI: 10.3390/ijerph17072610.

Dongfang Xu, Guillermo Lopez Garcia, Lisa Raithel, Rolland 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 applications (#SMM4H) 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.