UKYNLP@SMM4H2024: Language Model Methods for Health Entity Tagging and Classification on Social Media (Tasks 4 & 5)

Motasem S Obeidat
University of Kentucky
obeidat.s.motasem@uky.edu

Vinu H Ekanayake
University of Kentucky
vinu.ekanayake@uky.edu

Md Sultan Al Nahian
University of Kentucky
mna245@uky.edu

Ramakanth Kavuluru
University of Kentucky
ramakanth.kavuluru@uky.edu

Abstract

We describe the methods and results of our submission to the 9th Social Media Mining for Health Research and Applications (SMM4H) 2024 shared tasks 4 and 5. Task 4 involved extracting the clinical and social impacts of non-medical substance use and task 5 focused on the binary classification of tweets reporting children's medical disorders. We employed encoder language models and their ensembles, achieving the top score on task 4 and a high score for task 5.

1 Introduction

In today's digital landscape, social media platforms have evolved beyond mere communication channels, transforming into vital sources of real-time, user-generated data that reflect a wide array of public experiences and concerns. This transformation is particularly pertinent in the realm of public health, where social media discussions provide invaluable insights into both prevalent health issues and the personal impacts of various conditions. The critical analysis of these online dialogues is essential for understanding and addressing two significant public health challenges: nonmedical substance use and children's health disorders.

Task 4 and Task 5, conducted as part of the 9th Social Media Mining for Health Research and Applications (SMM4H) workshop, illustrate the potential of leveraging social media for health research. Task 4 focuses on extracting the clinical and social impacts of nonmedical substance use from Reddit discussions. Understanding these impacts is vital for developing nuanced interventions and educational programs aimed at mitigating the adverse outcomes of substance misuse. The analysis of user discussions can reveal the multifaceted consequences of such behavior, informing targeted interventions and effective medication strategies.

Task 5 addresses the binary classification of tweets related to children's health disorders, distinguishing between tweets from users reporting a child with disorders such as ADHD, ASD, delayed speech, or asthma from those that merely mention these conditions without indicating a personal effect. This task underlines the necessity for innovative data collection methods that can complement traditional epidemiological approaches, which often face logistical and financial barriers. By analyzing Twitter data, researchers can access a broader dataset, uncovering patterns and insights with speed and scale infeasible in conventional studies. This capability is crucial for developing targeted public health interventions and building support services that meet the needs of families dealing with children's health disorders.

2 Datasets

2.1 Task 4: Extracting clinical and social impacts of nonmedical substance use

The Task 4 dataset consists of 26,126 Reddit posts (60% for training, 20% for validation, and 20% for testing/evaluation). Only 318 posts (approximately 1.22%) were annotated for clinical impacts (health, physical condition, and mental well-being) or social impacts (effects on social relationships, community dynamics, and broader societal issues) (Ge et al., 2024). Thus this dataset has high imbalance in the sense that most posts would not contain any task specific entities.

2.2 Task 5: Binary classification of tweets reporting children's medical disorders

Task 5 uses a dataset of tweets, specifically targeting discussions where users reported their pregnancy and mentioned health conditions affecting their children — ADHD, ASD, delayed speech, or asthma. For binary classification, tweets were labeled as "1" if they reported a user having a child with a disorder and "0" if they merely mentioned

Proceedings of The 9th Social Media Mining for Health Applications (#SMM4H) Workshop and Shared Tasks, pages 124–129

Method (# parameters)	Strict Precision	Strict Recall	Strict F1
ALBERT CW0 (BiLSTM) (223M)	15.4	25.8	19.3
BERT CW100 (No BiLSTM) (110M)	11.5	23.9	15.5
BERT CW0 (No BiLSTM) (110M)	14.3	26.4	18.6
BERT CW0 (BiLSTM) (110M)	17.4	28.3	21.5

Table 1: Task 4 strict validation results for clinical and social impact recognition

the disorder. A total of 10,734 tweets were collected and divided into three sets: 7,398 tweets for training, 389 for validation, and 1,947 for testing. (Klein et al., 2024)

3 Methodology

For Task 4, the main NLP task is named entity recognition (NER), for which we use a span-based encoder-only model that enumerates contiguous spans of tokens up to a certain length (we used 8) and classifies each of them as any of the allowed entity types. We used the Princeton University Relation Extraction (PURE) (Zhong and Chen, 2021) pipeline's entity model component, which is originally inspired by other prior efforts (Wadden et al., 2019). Span-level representations for each token are first derived from pre-trained language models, such as the BERT base uncased model (Devlin et al., 2019) and the ALBERT xx-large model (Lan et al., 2020), both English models. The span representation is the concatenation of these encoder-only model outputs for the first and last token embeddings along with an embedding for span length. A feed forward network processes these span representations to compute the probability distribution of entity types. As a variation, we also incorporated a bidirectional LSTM (BiLSTM) layer, with 150 units in each direction (totaling 300 units), between the span embeddings and the classification layers in the BERT and ALBERT models; this was shown to improve results in prior experiments and was also reported by Li et al. (2021).

The dataset for Task 4 was used with varying batch sizes and context windows. Here, a context window denotes the extent of textual context surrounding the target sentence that is being considered during the entity extraction process. We did some basic preprocessing of the messages such as lower casing and converting it into the format expected by our span based approach. For hyperparameters used in task 4, please refer to Table 5 in the Appendix.

For task 5, we fine-tuned pretrained encoder-only language models for tweet classification. Specifically, we used a DeBERTa v2 xlarge model (He et al., 2021), a DeBERTa v3 Large model, already fine-tuned on the Multi-NLI (MNLI) dataset (Manakul et al., 2023), and a RoBERTa (Liu et al., 2019) base model, previously fine-tuned on a general tweet dataset (Loureiro et al., 2022). All models were further fine-tuned on the task 5 dataset, which focuses on children's disorders such as ADHD, ASD, speech delays, and asthma. The best-performing models were obtained by systematically adjusting key hyperparameters such as the number of epochs, batch size, and learning rate. The specific hyperparameters used for the models are detailed in Table 8 in the Appendix. To combine the potential complementary predictive capabilities of each model, we integrated them into 3-model majority vote ensemble.

4 Results

4.1 Task 4 findings

To evaluate the impact of the additional BiLSTM layer modification, the models were assessed using the F1 score, both with and without it. Experiments were conducted across several batch sizes and a batch size of four was determined to be optimal. Table 1 presents the strict F1 scores on the validation dataset, and Table 2 presents the relaxed, strict, and token-level F1 scores obtained on the test dataset, utilizing two different context window sizes: 0 and 100 (indicated as CW0 and CW100). The first three entries in each table represent the results that were officially submitted. The final row is a post-evaluation entry. Relaxed and token-level F1 scores on the validation dataset are provided in Appendix Tables 6 and 7, respectively. For token-level F1 scores, the micro-average F1 is reported.

For the validation results in Table 1, we used strict F1 scores to evaluate the models. The ALBERT model, configured with a context window of

Method (# parameters)	Relaxed F1	Strict F1	Token-level F1
ALBERT CW0 (BiLSTM) (223M)	40.4	14.4	53.2
BERT CW100 (No BiLSTM) (110M)	46.2	17.1	53.1
BERT CW0 (No BiLSTM) (110M)	45.9	16.1	48.8
BERT CW0 (BiLSTM) (110M)	47.8	14.5	52.4

Table 2: Test results for Task 4 (clinical/social impact spotting) considering relaxed, strict, and token-level F1 scores

Method (# parameters)	Precision	Recall	F1 Score
DeBERTa v3 large (MNLI) (304M)	94.2	97.0	95.6
RoBERTa base (Twitter) (100M)	93.9	91.1	92.5
DeBERTa v2 XL (900M)	93.5	95.6	94.5
Ensemble method (1.3B)	93.5	95.6	94.5

Table 3: Task 5 validation results for classification of tweets with parental disclosure of childhood disorders

zero and augmented by a BiLSTM layer, reached an F1 score of 19.3. In contrast, extending the context window to 100 for the BERT model without the inclusion of a BiLSTM layer decreased performance, with the F1 dropping to 15.5. The BERT model, without a BiLSTM layer and with a context window of zero, recorded an F1 score of 18.6. Finally, the BERT model with a BiLSTM layer and a context window of zero achieved the highest F1 score of 21.5 among the methods tested. These findings suggest that adding a BiLSTM layer does improve performance in the examined scenarios.

For the test results presented in Table 2, relaxed F1 scores were used for comparison as specified by the SMM4H testing guidelines. The results reveal a marginal benefit from integrating BiLSTM layers. Specifically, the ALBERT model augmented with a BiLSTM layer achieved a score of 40.4, which is substantially lower than the scores achieved by both configurations of the BERT model without the BiLSTM layer, which scored 46.2 and 45.9 for context windows of 100 and zero, respectively. Furthermore, the BERT model configured with a BiLSTM layer and a zero context window registered a score of 47.8, surpassing the performance of its counterpart without the BiLSTM. These results suggest that although BiLSTM layers have the potential to enhance feature representation and temporal dependencies, their effectiveness is likely dependent on the specific model architectures and the contextual requirements of the task. Although the ALBERT architecture and training regimen were introduced to be more efficient and performant compared to BERT models, that did not turnout to be the case for this task. Our best test score (row 2 of Table 2) is also the top score in the shared task. It is important to note nontrivial variations in the ranking of best models (as per strict F1 scores) based on validation and test scores — potentially due to smaller datasets, validation results may not be strong indicators of what works best in the end.

4.2 Task 5 findings

Extensive testing was conducted to determine the optimal values for parameters such as learning rate, batch size, and number of epochs for each model used. Tables 3 (validation) and 4 (test) present the results obtained for the two DeBERTa models, the RoBERTa model, and the ensemble model that combines all three. In the validation phase, the DeBERTa v3 Large model, fine-tuned on the MNLI dataset, exhibited superior performance, achieving an F1-score of 95.6. The DeBERTa v2 XL and the ensemble models also performed notably well, each achieving an F1-score of 94.5. The RoBERTa Base model, fine-tuned on Twitter data, achieved an F1 score of 92.5.

Method	F1 Score
DeBERTa v3 large (MNLI)	92.4
RoBERTa base (Twitter)	89.5
DeBERTa v2 XL	94.8
Ensemble method	94.2

Table 4: Task 5 test results

In the testing phase, these models were applied to the test data for both internal evaluation and competition submissions. For the competition, we submitted results from the DeBERTa v3 Large model, which secured an F1-score of 92.4 with the test dataset. In the post-evaluation phase, results from the other models were analyzed and submitted. The DeBERTa v2 XL model achieved the highest performance with an F1-score of 94.8, followed by the ensemble model with an F1-score of 94.2 and the RoBERTa Base model with an F1-score of 89.5. The DeBERTa v3 and RoBERTa models dipped over 3% from validation to test F1-score. But the DeBERTa v2 XL and ensemble models more or less stayed the same potentially due to their larger model capacities. This aspect needs further investigation to see whether there is a generalizable explanation for this or if this is purely an artefact of the task 5 dataset.

5 Concluding Remarks

In this report, we reviewed our approaches to SMM4H 2024 tasks 4 and 5, which mostly focused on applying encoder-only language models to NER and text classification. We were officially informed of our first place in task 4 results and it appears our post-evaluation scores for task 5 are near the top. We conducted preliminary error analyses that mostly pointed to informal/casual language (that does not adhere to grammatical norms) as a prominent trait characterizing both false positives and false negatives. While the precision and recall are around the same range for task 5, for the NER task 4 we notice recall is at least ten points higher than precision; the difference was close to 12 points in row 2 (Table 1). Reducing this gap without compromising too much on precision is a promising general strategy we intend to pursue in the future. This could be done by lowering the output probability threshold as a starting point. More sophisticated changes to the loss function and potential post-processing strategies may be needed to obtain further improvements.

We did not attempt to use decoder-only large language models (LLMs) (e.g., Mistral and Llama) because in our lab's prior explorations (Gupta et al., 2023), with ample training data, encoder-only models always fared better for information extraction (IE) tasks. This was also observed by other researchers who focused on IE tasks that need language understanding capabilities more than generative skills. However, it is worth reassessing this with bigger LLMs (7B and more parameters) to see if they excel at supervised NER and classification tasks. Our working hypothesis is that LLMs may show benefits when dealing with shorter entity NER tasks (single or two token entities) but encoder-only models might still be the best option for handling longer entities that were more common in task 4. Additionally, we also plan to exploit even bigger LLMs (e.g, GPT-4) to generate synthetic examples that augment training data for both tasks to see if that benefits the overall performance. However, it is not clear how augmentation can be carried out for NER to retain longer named entities intact while changing the overall sentence with decent coherence and preserved meaning. Effective augmentation for classification may be relatively easier to achieve. These are some future directions we hope to pursue soon.

Acknowledgement

This work is supported by the NIH National Institute on Drug Abuse through grant R01DA057686. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

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. *arXiv preprint arXiv:2405.06145*.

Shashank Gupta, Xuguang Ai, and Ramakanth Kavuluru. 2023. Comparison of pipeline, sequence-to-sequence, and gpt models for end-to-end relation extraction: experiments with the rare disease use-case. *arXiv preprint arXiv:2311.13729*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

AZ Klein, JA Gutiérrez Gómez, LD Levine, and G Gonzalez-Hernandez. 2024. Using longitudinal twitter data for digital epidemiology of childhood health outcomes: An annotated data set and deep neural network classifiers. *J Med Internet Res*, 26:e50652.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Fei Li, ZhiChao Lin, Meishan Zhang, and Donghong Ji. 2021. A span-based model for joint overlapped and discontinuous named entity recognition. 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 4814–4828.

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. *Preprint*, arXiv:1907.11692.

Daniel Loureiro, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-collados. 2022. TimeLMs: Diachronic language models from Twitter. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 251–260, Dublin, Ireland. Association for Computational Linguistics.

Potsawee Manakul, Adian Liusie, and Mark John Francis Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *ArXiv*, abs/2303.08896.

David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. 2019. Entity, relation, and event extraction with contextualized span representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5784–5789.

Zexuan Zhong and Danqi Chen. 2021. A frustratingly easy approach for entity and relation extraction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 50–61, Online. Association for Computational Linguistics.

Appendix

Hyperparameter	**Value**
head_hidden_dim	150
width_embedding_dim	150
max_span_length	8
lstm_hidden_dim	150 (300 units with BiLSTM)
train_batch_size	4
learning_rate	1e-5
task_learning_rate	5e-4
context_window	0 and 100
warmup_proportion	0.1

Table 5: Task 4 hyperparameters and configurations

Method (# parameters)	Relaxed Precision	Relaxed Recall	Relaxed F1 Score
ALBERT CW0 (BiLSTM) (223M)	31.6	52.6	39.5
BERT CW100 (No BiLSTM) (110M)	24.0	50.0	32.5
BERT CW0 (No BiLSTM) (110M)	28.1	51.7	36.4
BERT CW0 (BiLSTM) (110M)	32.1	52.4	39.9

Table 6: Task 4 relaxed validation results for clinical and social impact recognition

Method (# parameters)	Token-level Precision	Token-level Recall	Token-level F1
ALBERT CW0 (BiLSTM) (223M)	45.6	33.0	38.3
BERT CW100 (No BiLSTM) (110M)	46.4	38.7	42.2
BERT CW0 (No BiLSTM) (110M)	49.4	34.2	40.5
BERT CW0 (BiLSTM) (110M)	48.6	35.3	40.9

Table 7: Task 4 token-level validation results for clinical and social impact recognition

Method (# parameters)	# Epochs	Learning Rate	Batch Size
DeBERTa v3 Large (MNLI) (304M)	3	1e-5	8
RoBERTa base (Twitter) (100M)	3	2e-5	8
DeBERTa v2 XL (900M)	3	1e-5	16

Table 8: Hyperparameters for the models used in Task 5