Extracting Social Determinants of Health from Clinical Notes using LLMs

Charles Jin[1], Leah Ning[1], Sifei Han, PhD[2], Fuchiang (Rich) Tsui, PhD, FAMIA[2,3]

[1]University of Pennsylvania, SEAS 2027, [2]Tsui Laboratory, Department of Biomedical and Health Informatics, Children's Hospital of Philadelphia, [3]Perelman School of Medicine, University of Pennsylvania

PURM Penn Undergraduate Research Mentoring Program

Introduction

This study aims to use large language models (LLMs) to extract social determinants of health (SDOH) from clinical notes. We hypothesized that with prompt engineering and fine-tuning, open-source LLMs could improve their performance in the SDOH extraction task.

Background

- **Social determinants of health (SDOH)** are non-medical factors that can have a *significant* impact on patient health outcomes (poverty, unemployment, substance abuse, etc.)
  - Account for 80% of a patient's overall health
  - Knowing patients' SDOH greatly improves risk predictions, but SDOH are only found in unstructured, narrative clinical notes.
- **Why LLMs?** Current state-of-the-art approaches include using the BERT model, but newer LLMs, with much higher number of parameters, have significantly out-performed BERT in text classification tasks.

Methodology - Setup

- **LLMs**: Llama-3-8b, Llama-3-70b, Mixtral-8x7b
- **Dataset**: 188 annotated social work notes pairs for inference (dev), 1316 for fine-tuning (train)

Patient Note:

```
SOCIAL HISTORY: The patient denies any
history of tobacco or alcohol use. She lives
alone. Her family is involved in her care.
```

Annotation File:

Triggers

```
T1  Tobacco 51 58   tobacco
T3  StatusTime 29 39    denies any
T2  Alcohol 62 73   alcohol use
T4  LivingStatus 80 85  lives
T5  TypeLiving 86 91    alone
T6  StatusTime 80 85    lives
```

SDOH events/labels

```
E1  Tobacco:T1 Status:T3
E2  Alcohol:T2 Status:T3
E3  LivingStatus:T4 Status:T6 Type:T5
```

Argument values

```
A1  StatusTimeVal T3 none
A2  TypeLivingVal T5 alone
A3  StatusTimeVal T6 current
```

Methodology - Prompt Engineering

SDOH extraction inference tasks:

- Note-level: Determine whether SDOH categories are present in each note
  - Compared results of 0-shot learning prompts and 5-shot, as well as different formatting of prompts
- Span-level: Determine spans/phrases for SDOH events, triggers, and argument values
  - 5-shot prompt with task outline, output format guidelines, and example input-output pairs
  - Modifications include additional span inputs and output format examples.

Methodology – Fine-tuning

Supervised fine-tuning:

- LoRA fine-tuning for the span-level task with Llama-factory
- Fine-tuning dataset: MIMIC train set, combined with best span-level prompt for corresponding models
- Epochs: 30, Learning rate: 5 x 10^5
- Computing resources: 2 A100 GPUs

Evaluation Metrics:

- Accuracy, Precision (positive predictions quality), Recall (sensitivity to positive instances), F1 Score (harmonic mean of precision and recall)

Results

Note-Level Task with Span-Level Prompts Evaluation Results:

	F1	Precision	Recall	Accuracy
Llama-3-70b (Prompt 1)				
Alcohol	0.9434 (0.9424, 0.9444)	0.9156 (0.914, 0.9172)	0.9735 (0.9725, 0.9744)	0.9049 (0.9034, 0.9065)
Drug	0.9046 (0.9031, 0.9062)	0.8585 (0.856, 0.8609)	0.9573 (0.9558, 0.9588)	0.8981 (0.8965, 0.8997)
Employment	0.9228 (0.9213, 0.9243)	0.8831 (0.8807, 0.8854)	0.9674 (0.966, 0.9688)	0.9269 (0.9255, 0.9283)
LivingStatus	0.8993 (0.8979, 0.9006)	0.8177 (0.8154, 0.8199)	1.0 (1.0, 1.0)	0.8626 (0.8609, 0.8643)
Tobacco	0.948 (0.9471, 0.949)	0.9166 (0.915, 0.9182)	0.9821 (0.9813, 0.9828)	0.9122 (0.9107, 0.9138)
Overall	0.9271 (0.9263, 0.9279)	0.8822 (0.8808, 0.8837)	0.9772 (0.9768, 0.9777)	0.7023 (0.7, 0.7047)

	F1	Precision	Recall	Accuracy
Mixtral-8x7b (Prompt 1)				
Alcohol	0.9125 (0.9113, 0.9138)	0.9251 (0.9236, 0.9266)	0.9011 (0.8993, 0.9029)	0.863 (0.8611, 0.8648)
Drug	0.8639 (0.8619, 0.8659)	0.8236 (0.8208, 0.8265)	0.9103 (0.9081, 0.9125)	0.863 (0.8612, 0.8649)
Employment	0.8772 (0.8753, 0.8791)	0.7826 (0.7796, 0.7855)	1.0 (1.0, 1.0)	0.8859 (0.8843, 0.8875)
LivingStatus	0.8712 (0.8696, 0.8728)	0.7727 (0.7702, 0.7751)	1.0 (1.0, 1.0)	0.8218 (0.8198, 0.8238)
Tobacco	0.9273 (0.9262, 0.9284)	0.9024 (0.9006, 0.9041)	0.9543 (0.953, 0.9555)	0.8846 (0.8829, 0.8863)
Overall	0.8958 (0.8948, 0.8967)	0.8482 (0.8467, 0.8498)	0.9495 (0.9486, 0.9504)	0.5998 (0.5973, 0.6024)

Span-Level Task Evaluation Results Progression:

	NT	NP	TP	Precision	Recall	F1
Mixtral-8x7bL	1239	1398	219	0.1567	0.1768	0.1661
Fine-tuned Mixtral-8x7bL	1443	1569	543	0.3461	0.3763	0.3606
Index-Corrected Mixtral-8x7bL	1443	1569	1141	**0.7272**	**0.7907**	0.7576

Key: NT=Number of Truths, NP=Number Predicted, TP=True Positives

Evaluation & Conclusion

- **Span-Level Evaluation:** We used the BRAT scoring Python package, which considers the event types, triggers, argument values, and span index locations to generate the evaluation metrics (by comparing to the gold-standard from human annotation).
- **Data processing:**
  - Cleaned inference output by removing output lines that don't match the specified output format.
  - Initially, evaluation scores were low due to LLMs outputting incorrect indices even though the spans they identified were correct.
  - Index correction: Searched through the input file for each identified trigger span and replaced the index the LLMs outputted with the index of the closest trigger span in the input file. Greatly improved performance.
- **Summary**: Our best result for the note-level task was from the Llama-3-70b model with an **F1 score of 0.93**, and our best result for the span-level task was from the index-corrected, finetuned Mixtral-8x7b model with an **F1 score of 0.76**.

Acknowledgements

We'd like to express our gratitude to our mentors, Sifei and Rich, for their invaluable guidance and support. We also want to thank the rest of the Tsui Lab members for their insightful questions and engaging conversations during our lab meetings.