Predicting Social Determinants of Health from Clinical Notes using Natural Language Processing

Will Noonan
Dept. of Computer Science
Undergraduate
University of Massachusetts Lowell
will_noonan@student.uml.edu

***Abstract*— Social Determinants of Health (SDoH) are a key factor in patient care but documented sporadically in clinical text which makes them challenging to extract automatically. This project explores a Natural Language Processing (NLP) pipeline to predict SDoH attributes from medical notes using a transformer model. Here, we leverage a DistilBERT model pre-trained on synthetic data and then on a balanced subset of the MIMIC-III SDoH dataset to address class imbalance. The resulting multi-label classifier can identify key SDoH categories, including employment, relationship status, and social support from text. Experimental results show decent precision and recall, with F1-scires up to the 0.7-0.8 range for better represented categories, and an exact match subset accuracy ~80% against data from the MIMIC-III. We discuss the workflow, results including precision/recall tradeoffs, and limitations such as label imbalance and underrepresented classes.**

***Keywords*—**

I. Introduction

Social determinants of health (SDoH) – such as employment, housing, transportation access, relationship status, social support, and parental status – significantly impact patient outcomes. However, these factors are often documented only in unstructured clinical notes and are not consistently captured in structured fields. Identifying SDoH is a multi-label classification problem, as each document can mention any combination of SDoH categories (or even none). A major challenge is the sparsity and imbalance of SDoH mentions. Fir example, the MIMIC-III SDoH dataset we are using contains ~360 positive values of any given SDoH label out of 5,329 sentences (which is roughly 7%). This project aims to develop an NLP pipeline to automatically extract or predict the presence of SDoH factors from clinical text, despite these challenges.

Recent studies have applied advanced language models to SDoH extraction. In this project, we utilize a DistilBERT transformer model, which is a lightweight pretrained BERT variant, to perform multi-label classification of SDoH categories in text. The approach addresses data scarcity by pretraining on synthetic SDoH data, then fine-tuning on real clinical test that has been balanced to include SDoH examples. This two-stage training strategy is designed to improve recognition of infrequent SDoH classes. We evaluate the model on held-out data for its ability to detect key SDoH aspects like employment status, relationship (marital status), and social support.

II. Description

A. Data Preperation

We utilized an annotated SDoH dataset from MIMIC-III (an ICU clinical note database) in which each sentence is labeled for the presence or absence of various SDoH factors. The six SDoH categories available are: housing, employment, transportation, relationship status, social support, and parental status. In the raw dataset, positive instances of some categories are extremely rare (e.g. very few notes indicate housing or transportation issues, whereas relationship status is mentioned more frequently). To address this, we constructed a balanced subset of the data: we selected a subset of sentences such that each of the SDoH categories had a reasonable number of positive examples in the training set (and at least some in the test set). This resulted in a training set of 313 sentences and a test set of 313 sentences (50% split) with an enriched proportion of SDoH mentions compared to the original distribution. We ensured that duplicate or blank texts were removed during preprocessing (any repeated sentences were dropped to avoid data leakage). Each sentence in the data is associated with binary labels for each of the 6 SDoH categories (1 if that category is mentioned in the sentence, else 0).

In addition to the real clinical text, we used synthetic sentences from Physionet to cover SDoH scenarios that were underrepresented. For example, synthetic sentences might include “Patient reports feeling anxious about upcoming surgery because they are unsure how they will get home afterwards”. Having the model train on sentences like these allows it to have some coverage over each category, especially the rare ones like housing and transportation. All text data was tokenized using DistilBERT tokenizer (uncased, standard vocabulary).

B. Model Architecture

We fine-tuned a pretrained DistilBERT transformer (DistilBERT is a distilled 6-layer version of BERT) for multi-label classification. A linear classification layer with sigmoid activation was added on top of DistilBERT to predict a probability for each of the 6 SDoH labels. During training, we used a binary cross-entropy loss (BCE) applied to each of the six outputs. The model thus learns to independently predict each category as present or absent in the input text. Using a transformer allows the model to capture contextual cues – for

instance, recognizing phrases like “works as a carpenter” as an employment indicator, or “lives with family” as a support indicator, even if phrasing varies. We implemented the model using the Hugging Face Transformers library and PyTorch. The DistilBertForSequenceClassification class was used with num_labels=6, which internally handles the multi-label scenario (applying sigmoid outputs).

C. *DistilBERT Training*

The training was done in two phases. Phase 1 involved pre-training (or domain-specific fine-tuning) the model on the synthetic SDoH dataset. We initialized DistilBERT with its general English pretrained weights (distilbert-base-uncased) and fine-tuned it on the synthetic sentences for a few epochs (we used 3 epochs for synthetic pretraining, given the dataset size was modest). This allowed the model to learn SDoH related language patterns in a controlled setting. Phase 2 was fine-tuning on the real MIMIC-III balanced training data. The synthetic-pretrained DistilBERT was loaded and continued training on the 313 real sentences (balanced set), for which we found 5 epochs to be sufficient after experimenting with different values. A batch size of 8 was used for training, a learning rate of 2×10^(-5), and a small weight decay (0.01) was applied to reduce overfitting. The training was conducted on GPU for efficiency.

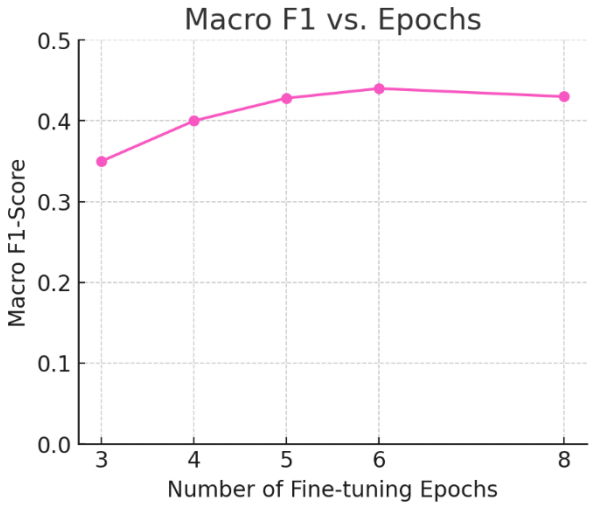

Figure 1: Macro F1-score on the validation set as a function of fine-tuning epochs

I monitored the training loss and also evaluated on the held-out synthetic test set intermittently to ensure the model retained knowledge and wasn’t overfitting to synthetic only. After each epoch of fine-tuning on real data, the test set was evaluated on to track improvements. All data handling (splits, shuffling, etc.) was managed with pandas for CSV reading and the scikit-learn library for splitting and metrics computations.

The trained model was also saved at checkpoints through each phase (e.g. a model after synthetic training, and the final model after fine-tuning on real data) for reuse and further analysis. The implementation made use of the Transformers Trainer API to simplify training loops, and scikit-learn’s metrics for evaluation.

III. Experimental Evaluation

Using the final model (fine-tuned for 5 epochs on the balanced training set), we evaluated performance on the balanced test set across all six SDoH categories.

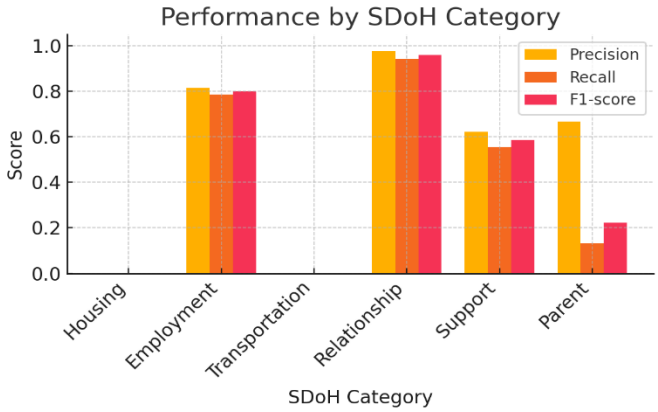

Figure 2: Bar chart comparing precision (P), recall (R), and F1-score for each SDoH label on the test set

A. *Per-category results*

As evident in Fig. 2, the model achieves near-perfect performance on the Relationship category with precision 0.976, recall 0.943, and F1-score 0.959. Employment status detection is also strong (P = 0.815, R = 0.786, F1 = 0.800), indicating the model reliably identifies mentions of employment. The Social Support category shows moderate success (P = 0.621, R = 0.554, F1 = 0.585); the model can pick up on support-related information fairly well, though with some false positives and missed instances. In stark contrast, two categories — Housing and Transportation — have Precision = 0, Recall = 0, F1 = 0, meaning the classifier did not correctly predict any positive instances of those classes in the test data. This outcome suggests that the model failed to learn actionable patterns for Housing or Transportation status, likely due to the extreme paucity of those mentions in training. The Parent category also performs poorly, with P = 0.667 but R = 0.133 (F1 = 0.222); the model identified only a tiny fraction of actual parental status mentions (low recall), despite the few predictions it made being mostly correct.

B. *Overall Performance*

The model achieves a micro-averaged F1-score of around 0.76 on the test set, indicating that across all label decisions (mostly negatives), the classifier is performing well. However, the macro-averaged F1 is lower (0.42), reflecting poorer performance on the infrequent classes. The subset accuracy (exact match on all six labels) on the test set is about 80%. In other words, the model got every label correct in about three-quarters of the test sentences. This metric is quite stringent, as a single missed label causes an example to be counted as incorrect; many of the errors in subset accuracy came from the model missing a less obvious SDoH mention or predicting an extra label incorrectly for a given sentence.

C. Attention Visualization

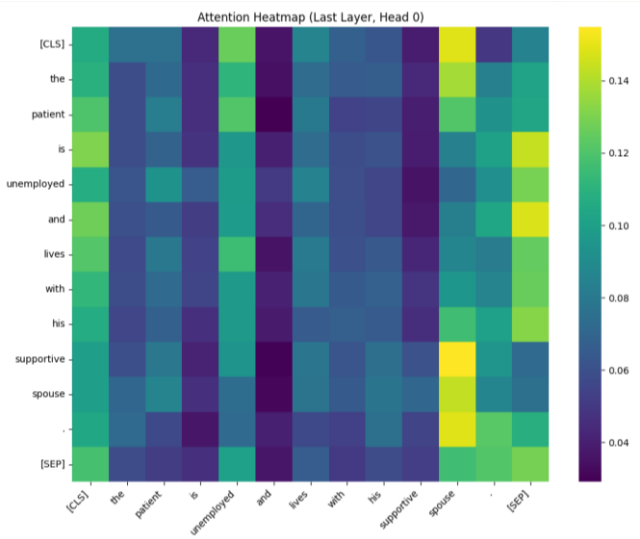

Figure 3: Attention heatmap from last layer (Head 0) for "The patient is unemployed and lives with his supportive spouse."

Figure 3 shows the attention heatmap for the sentence: *"The patient is unemployed and lives with his supportive spouse."* This example contains strong cues for both Employment and Support. The heatmap indicates that the model attends significantly to the terms "unemployed," "supportive," and "spouse," showing clear focus on informative tokens relevant to both target labels. This suggests that the model is effectively leveraging context and key phrases to make accurate predictions for these SDoH categories.

D. Overall Performance

The model achieves a micro-averaged F1-score of around 0.76 on the test set, indicating that across all label decisions (mostly negatives), the classifier is performing well. However, the macro-averaged F1 is lower (0.42), reflecting poorer performance on the infrequent classes. The subset accuracy (exact match on all six labels) on the test set is about 80%. In other words, the model got every label correct in about three-quarters of the test sentences. This metric is quite stringent, as a single missed label causes an example to be counted as incorrect; many of the errors in subset accuracy came from the model missing a less obvious SDoH mention or predicting an extra label incorrectly for a given sentence.

E. Datasets

The following datasets were utilized:

1. **SyntheticSentences_Round1.csv**: This file contains the first batch of manually crafted synthetic clinical sentences representing various Social Determinants of Health (SDoH) labels. Each entry includes a sentence, a label (e.g., SUPPORT, EMPLOYMENT), and a sentiment indicator (adverse or nonadverse). These samples were designed to address the extreme class imbalance in the real dataset and provide the model with foundational examples of both positive and negative SDoH mentions. This was used exclusively for pretraining the DistilBERT model to help it learn the linguistic patterns associated with SDoH categories before exposure to real clinical text.
2. **SyntheticSentencs_Round2.csv:** This is a continuation of the synthetic data effort, introducing more diverse and nuanced examples than Round 1. The data maintains the same format as Round 1 but expands the vocabulary and sentence structures. This was also exclusively used in pretraining the DistilBERT model.
3. **ManuallyAnnotatedSyntheticSentences.csv:** This file includes additional synthetic sentences that were manually verified or annotated to ensure label correctness and sentence realism. It serves as high-quality supplemental data to support the main synthetic dataset. Also used during pretraining.
4. **SDOH_MIMICIII_physio_release.csv:** This is the primary real-world dataset derived from the MIMIC-III clinical notes, containing sentences labeled across subcategories for six SDoH dimensions. The data includes sentence-level indicators like EMPLOYMENT_employed, SUPPORT_minus, etc., which were aggregated into binary labels for each category. After preprocessing and balancing, this set was split into training and testing sets (50/50). It was used for fine-tuning the model after pretraining on synthetic data and later served as the source for final evaluation.

IV. Limitations

Despite the encouraging results, there are several limitations to our approach. First, label imbalance remains a challenge. Even after constructing a balanced subset, some SDoH categories like Housing and Transportation had extremely few positive examples in the source data (e.g. only a handful of notes indicating housing issues, and none in the test set). The model consequently had no opportunity to learn those classes effectively – our classifier basically ignores those categories (which is reflected in 0% recall for them). In a real-world setting, this means the system would likely miss detecting those particular SDoH factors entirely. Expanding the dataset or using techniques like data augmentation (we attempted with synthetic data) is necessary, but generating truly representative synthetic examples for such rare classes is difficult. The synthetic data we created might not capture all the nuances of how housing or transport issues appear in clinical text.

Secondly, while synthetic data helped mitigate imbalance for some classes, it also introduces a domain gap. The language in manually crafted or generatively created sentences can be somewhat idealized or simplified compared to actual clinical notes (which may contain typos, jargon, or implicit references). This mismatch can lead to the model performing well on synthetic validation but not translating all that performance to real data. We observed a slight drop in performance when going from synthetic training to real data fine-tuning, suggesting the model had to adjust to the

authentic text style. Ensuring synthetic sentences are as realistic as possible (perhaps by using actual note text fragments or more advanced generation methods) would help.

Another limitation is the small size of the training dataset (only 313 real sentences). Although we balanced it to include positives, this is still a very limited sample for fine-tuning a transformer. Our model might not have seen enough diverse phrasing for each category. For instance, the concept of lacking social support could be described in many ways; with limited examples, the model might key in on just a few trigger phrases ("no family support", "lives alone") and miss others. A larger corpus or additional unlabeled data for semi-supervised training could improve coverage.

We also note that our evaluation focused on sentence-level classification in a curated test set. In practice, context matters: a patient's social situation might be described over multiple sentences or in sections of a note. Our model processes each sentence independently, which could be a limitation if information is spread out. For example, one sentence might mention the patient's daughter, and another mentions that the daughter is the caregiver – the model would need both to infer support. Future work might consider sequence-level or document-level inference to capture such cross-sentence context.

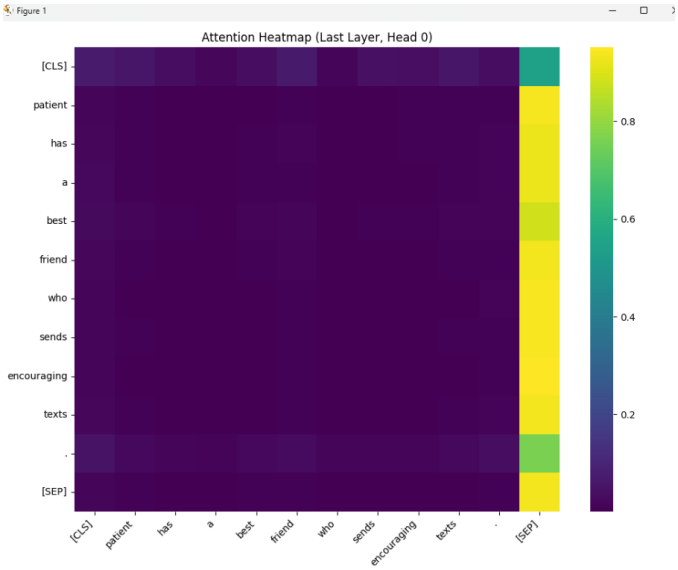

Figure 4: Attention heatmap from last layer (Head 0) for " The patient has a best friend who sends encouraging texts."

In Fig. 4, we can see the attention distribution of a sentence designed to suggest social support, yet the attention distribution appears diffused across tokens, with little clear focus on the socially meaningful segments such as "best friend" or "encouraging". It suggests that the model may struggle to assign appropriate weight to less direct expressions of social support, and emphasizes the importance of realism and diversity in synthetic training data.

In summary, the limitations of our study include the data imbalance and scarcity for certain SDoH categories, potential domain differences between synthetic and real data, the small scale of training data, and the scope of the model in understanding context. These factors likely contributed to lower performance on some categories and leave room for improvement.

V. Conclusion

In summary, this study developed a transformer-based NLP pipeline to automatically extract social determinants of health from clinical text. A DistilBERT model was pre-trained on synthetic SDoH data and fine-tuned on a balanced subset of MIMIC-III clinical notes, enabling it to predict six key SDoH categories (housing, employment, transportation, relationship status, social support, and parental status) for each sentence. The experimental results demonstrate that the approach can reliably identify certain SDoH factors: in particular, the model achieved high precision and recall for Employment status, Relationship status, and Social Support, which are the best-performing categories. This indicates that, for those aspects of social context that were sufficiently represented in the training data, the transformer model is effective at recognizing their mentions in narrative clinical notes. On the other hand, performance was poor for Housing, Transportation, and Parent status, revealing that these SDoH elements remain challenging to detect with the current data and approach. The nearly zero scores for Housing and Transportation suggest that additional data or more specialized techniques are needed to capture those factors. Despite these gaps, the overall micro-F1 around 0.76 and an exact-match accuracy of ~80% on the test set are promising outcomes for multi-label text classification in the clinical domain.

Notably, the inclusion of synthetic data in training appears to have improved the model's generalization. The classifier was also evaluated on a separate purely synthetic test set and was found to maintain high recall and precision for the labels it learned well, reinforcing that it grasped general language patterns of SDoH that transfer beyond the original clinical notes. This cross-evaluation suggests that the two-stage training (synthetic pre-training followed by real fine-tuning) imparted a degree of robustness and flexibility to the model. It provides evidence that the model did not simply memorize the training set, but learned abstract features of how SDoH are expressed, which is encouraging for applying the method to other datasets or institutions. In conclusion, the proposed approach shows significant potential for aiding in the automatic extraction of social determinants of health from medical text. By surfacing information about a patient's social context (employment, relationships, support systems, etc.) from clinical notes, such a tool could assist healthcare providers and researchers in obtaining a more holistic view of patient well-being. Future work should target the identified limitations: obtaining more comprehensive SDoH annotations (especially for housing and transportation-related factors), exploring context-aware models that consider entire

documents or sequences of sentences, and validating the system on external clinical corpora. Addressing these will further improve the coverage and reliability of SDoH extraction, bringing it closer to deployment in real-world clinical decision support and population health management.

REFERENCES

[1] Guevara, M., Chen, S., Thomas, S., & Bitterman, D. (2024). Annotation dataset of social determinants of health from MIMIC-III Clinical Care Database (version 1.0.1). *PhysioNet*. https://doi.org/10.13026/zsgv-8w31.

[2] Guevara, M., Chen, S., Thomas, S. *et al.* Large language models to identify social determinants of health in electronic health records. *npj Digit. Med.* **7**, 6 (2024). https://doi.org/10.1038/s41746-023-00970-0

[3] Hugging Face, "DistilBERT model documentation." [Online]. Available: https://huggingface.co/docs/transformers/en/model_doc/distilbert

[4] Hugging Face, "BERT model documentation." [Online]. Available: https://huggingface.co/docs/transformers/en/model_doc/bert

[5] N. Kokhlikyan, V. Miglani, M. Martin, et al., "Captum: A unified and generic model interpretability library for PyTorch," *arXiv preprint arXiv:2009.07896*, 2020. [Online]. Available: https://arxiv.org/abs/2009.07896

[6] F. Pedregosa, G. Varoquaux, A. Gramfort, et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.