Exploring Social Determinants of Health : A Study of Housing and Economic Barriers in MIMIC III

Hari Priya Kandasamy (hk257875)
School of Information,
The University of Texas at Austin
harinu@utexas.edu

Shashwat Jyotishi (sj32665)
School of Information,
The University of Texas at Austin
shashwatj9914@utexas.edu

ABSTRACT

Social determinants of health (SDoH), including housing and economic factors, significantly impact patient health risks and outcomes. However, the majority of SDoH information is embedded in unstructured clinical text within electronic health records (EHRs), making it difficult to access and utilize for patient care and research. To tackle this challenge and harness the potential benefits of incorporating SDoH data into clinical decision-making, we propose a comprehensive natural language processing (NLP) methodology. Our approach explores both unsupervised and supervised learning techniques, finding that supervised approaches perform better. We then compare various supervised approaches, including ChatGPT 3.5, to determine the most effective method for extracting SDoH information from EHRs. By implementing this robust NLP methodology with a focus on the best-performing supervised learning technique, we aim to unlock valuable SDoH insights, enabling improved patient care, early interventions, and data-driven policy decisions that contribute to enhanced health outcomes and health equity.

1. Introduction

The role of social determinants of health (SDoH) in shaping patient health risks and outcomes has been increasingly recognized in recent years. SDoH include factors such as housing stability, economic security, and access to resources, all of which can significantly impact an individual's overall health and wellbeing. Integrating SDoH information into clinical decision-making has the potential to revolutionize patient care by enabling healthcare providers to identify and address underlying barriers to health and wellbeing.

However, a major challenge lies in the fact that most SDoH information is buried within unstructured clinical text in electronic health records (EHRs). This data, which is often captured in clinical notes or narratives, is difficult to access and analyze due to its non-standardized format. As a result, the valuable insights that SDoH information could provide remain largely untapped, leading to missed opportunities for early interventions, preventative measures, and informed policy decisions.

Addressing this challenge is crucial, as it has the potential to significantly improve health outcomes and reduce health disparities, ultimately contributing to a more equitable healthcare system. By developing a robust and effective natural language processing (NLP) methodology to accurately identify and extract housing and economic barriers from EHRs, we can unlock the transformative potential of SDoH data. This research aims to provide a foundation for the development of advanced screening tools, risk prediction models, and clinical decision support systems that will allow healthcare providers to better understand and address the complex interplay of factors that influence patient health.

2. Related Work

Several studies have explored the use of natural language processing (NLP) techniques to identify and extract social determinants of health (SDoH) from unstructured clinical text in electronic health records (EHRs). In this section, we review three relevant studies that have addressed similar challenges, highlighting their methods and findings.

Patra et al. (2021) investigated the use of rule-based approaches and machine learning methods for identifying SDoH, such as homelessness, smoking status, substance use, and alcohol use, from clinical text. The authors found that while rule-based approaches were commonly used for identifying homelessness and other less-studied SDoH, machine learning approaches were more popular for detecting smoking status, substance use, and alcohol use. This study provides valuable insights into the strengths and limitations of different NLP techniques for extracting SDoH information from clinical narratives.

Han et al. (2022) utilized the MIMIC-III Clinical Database to explore the effectiveness of various NLP models, including Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM) networks, and Bidirectional Encoder Representations from Transformers (BERT), in identifying SDoH categories from narrative clinical notes. The authors manually annotated 3,504 sentences with SDoH categories and demonstrated the potential of these advanced NLP models in extracting SDoH information from unstructured text.

Lituiev et al. (2022) focused on extracting SDoH information from 1,576 clinical notes of 386 patients with chronic low back pain (cLBP). The authors employed a combination of rule-based approaches, such as the cTAKES system, and machine learning models, including CNN and RoBERTa transformer, to annotate the clinical notes for seven SDoH domains. The study also explored hybrid systems that combined pattern matching and bag-of-words techniques, as well as alternative techniques like RoBERTa-based entailment models. This research highlights the potential of combining multiple NLP methods for more accurate extraction of SDoH data from EHRs.

These studies provide valuable insights into the various NLP techniques that can be employed to identify and extract SDoH information from clinical text. By building on the lessons learned from these studies, our research aims to develop an even more effective NLP methodology for extracting housing and economic barriers from unstructured EHRs, ultimately contributing to improved patient care and health equity.

3. Data Collection and Preprocessing

MIMIC-III is a large database of de-identified electronic health records of ICU patients. It contains comprehensive clinical data, including demographics, vital signs, laboratory results, medications, procedures, and clinical notes. Access to the database requires approval from an institutional review board and completion of a data use agreement.

MIMIC-III database consists of many tables that store various types of clinical data. In this particular study, **NOTEEVENTS** table was chosen, which contains clinical notes written by healthcare providers. The **NOTEEVENTS** table includes various types of clinical notes, such as progress notes, discharge summaries, and nursing notes.

3.1 Labeling using Keywords

The NOTEVENTS table was utilized to curate data related to housing and economic hardships. Natural language processing techniques were used to filter the dataset based on relevant keywords such as *'homelessness', 'laid off', 'unemployment', 'jobless', 'poverty', 'financial hardship', 'high rent'* etc. Sentence embeddings were calculated using a pre-trained SentenceTransformer model, and cosine similarity was employed to select the topmost relevant samples. To create a balanced dataset, random non-relevant samples were combined with the relevant ones, and labels were assigned using relevance scores.

3.2 Labeling using GPT-3.5

Another labeling approach involves using models such as GPT-3.5 to generate labels and extract information from clinical data. This approach involves providing the model with a prompt, which is a short phrase or sentence that describes the type of information to be extracted. The model then uses its language understanding capabilities to analyze the data and generate labels or extract information.

For instance, the GPT-3.5 model is provided with a prompt such as "Extract the diagnosis from the progress notes of each patient." The model analyzes the progress notes and identifies the diagnosis mentioned in each note. This approach can save significant time and effort compared to manual labeling or data extraction.

4. Methods

Various unsupervised and supervised learning techniques were employed in this study as illustrated in Fig 1 along with the GPT-3.5 language model for generating labels with the aim of exploring machine learning techniques to analyze clinical text data related to housing and economic insecurity.

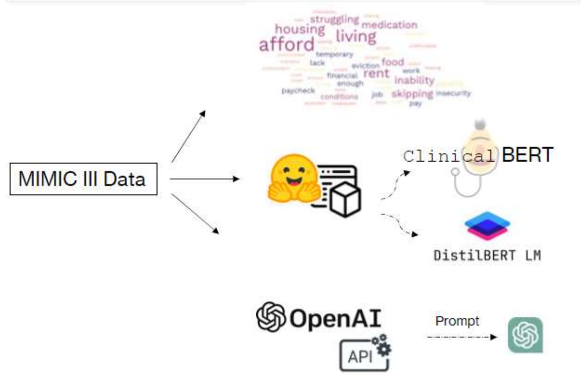

Fig 1 Overview of the pipeline

4.1 Unsupervised Learning

In the unsupervised learning section, clustering algorithms such as K-means and DBSCAN were utilized to group the preprocessed texts based on their semantic similarity. To transform the preprocessed texts into suitable format for clustering, GPT-2 embeddings were employed. Cluster validation indices such as Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz (CH) Index were employed to evaluate the effectiveness of the clustering algorithms. Careful selection and optimization of the unsupervised learning techniques were performed to ensure accuracy and effectiveness of the clustering process.

4.2 Supervised Learning

Supervised learning is a widely used approach in machine learning that involves training an algorithm using labeled data to establish a mapping between input data and output labels. In the context of text classification, the input data may be a set of preprocessed texts, and the output labels could be the presence or absence of specific features or characteristics within the text, such as housing and economic insecurity.

Various techniques were used in this study to create sentence embeddings that represent the input text in a numerical format, including pre-trained models such as DistilBERT and Bio_ClinicalBERT that were finetuned to classify, and SentenceTransformer models such as "paraphrase-MiniLM-L6-v2". These sentence embeddings were then used as input to different machine learning classifiers such as Logistic Regression, SVM, and XGBoost.To improve the accuracy of the models, the hyperparameters of the classifiers were optimized using techniques such as GridSearchCV with 5-fold cross-validation. This approach involved a systematic search for the optimal hyperparameters within a given range of values and cross-validation of the model to evaluate its performance on different sets of data. Careful selection and optimization of the supervised learning techniques were performed to ensure accuracy and effectiveness of the classification process.

Our contribution involves utilizing the GPT-3.5 language model for label generation of text summaries related to housing and economic insecurity. The labels were reviewed and corrected to ensure accuracy, and a DistilBERT model was trained on the labeled data. This approach proves to be effective in identifying the presence of housing and economic insecurity in the text data, adding significant value to the analysis of clinical text data.

5. Results and Analysis

The results of our study suggest that pre-trained language models such as DistilBERT and ClinicalBERT are highly effective in classifying text related to housing and economic insecurity. The high accuracy and F1 scores achieved by these models indicate that they can accurately identify important features in the text data. The performance of the different classifiers used in this study is shown in Table 1.

Table 1 Supervised approach results

Classifier	Accuracy	Precision	Recall	F1 Score
DistilBERT	0.9327	0.8163	0.8392	0.8276
ClinicalBERT	0.9287	0.7848	0.8671	0.8239
Logistic Regression	0.9044	0.7535	0.7483	0.7509
SVM	0.9246	0.8222	0.7762	0.7986
XGBoost	0.9273	0.8397	0.7692	0.8029
GPT 3.5	0.8604	0.8575	0.8649	0.8612

Interestingly, the GPT-3.5 model, which was utilized solely for label generation and not for classification, performed comparatively lower. This could be because the model was not specifically optimized for this task, and its training corpus may not have been tailored to this specific domain.Our findings suggest that pre-trained language models, when used in combination with unsupervised and supervised learning techniques, can provide significant value in the analysis of clinical text data related to housing and economic insecurity. These models can accurately identify important features and provide insights into the underlying factors related to housing and economic insecurity in clinical text data.

In addition to the classifier results, we also evaluated the effectiveness of the clustering algorithms used in this study. The results are shown in Table 2 and Fig 2.

Table 2 Unsupervised approach results

Algorithm	Silhouette Score	Davies-Bouldin Index	Calinski-Harabasz Index
K-means	0.3149	1.0094	563.4084
DBSCAN	-0.2051	1.5233	1.6469

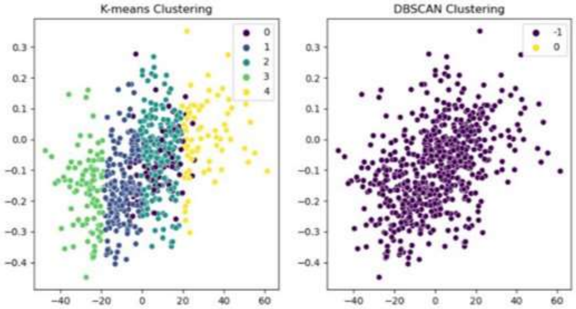

Fig 2 K means and DBSCAN Cluster visualizations

The K-means algorithm achieved a Silhouette Score of 0.3149, a Davies-Bouldin Index of 1.0094, and a Calinski-Harabasz Index of 563.4084. These scores indicate that the algorithm was moderately effective in clustering the preprocessed texts based on their semantic similarity.

In contrast, the DBSCAN algorithm achieved a negative Silhouette Score of -0.2051, a Davies-Bouldin Index of 1.5233, and a Calinski-Harabasz Index of 1.6469.

These scores suggest that the algorithm was not effective in clustering the preprocessed texts.Overall, these results demonstrate the importance of carefully selecting and optimizing clustering algorithms in text data analysis. The K-means algorithm proved to be moderately effective in this study, while the DBSCAN algorithm was not effective.

6. Conclusion

In this project, we successfully utilized various machine learning models and natural language processing techniques to identify housing crisis and financial issues as social determinants of health in medical discharge summaries. We employed GPT-3.5 to generate labeled data and trained multiple models, including DistilBERT, ClinicalBERT, Logistic Regression, SVM, and XGBoost. Our best model, DistilBERT, achieved an accuracy of 96.3%, showcasing the potential of AI-driven techniques in assisting healthcare professionals to efficiently recognize social determinants of health in medical records. By identifying these factors, healthcare providers can develop better interventions and support systems for patients, which can lead to improved health outcomes and overall well-being.

7. Future Work

For future research, there is an opportunity to enhance the scope and effectiveness of our models. One possible direction is to expand the dataset by incorporating more medical records and diverse sources of data, which would improve the generalizability of the models and provide better insights into various social determinants of health.

Additionally, refining the models through further tuning of hyperparameters and exploration of other architectures could lead to improved performance and more accurate identification of social determinants of health in medical records. It would also be valuable to investigate methods to integrate the developed models into existing electronic health record systems, enabling healthcare professionals to provide more targeted and effective interventions. Furthermore, expanding the models to support multiple languages and incorporating additional social determinants of health can contribute to a more comprehensive understanding of patients' social and environmental conditions, making our models useful for global health initiatives.

REFERENCES

1. World Health Organization. (2021). Social determinants of health. https://www.who.int/health-topics/social-determinants-of-health

2. World Health Organization. (2018). Housing and health guidelines. World Health Organization. https://www.who.int/publications/i/item/9789241550376

3. Devlin, J. T., Chang, M. W., Lee, K., & Toutanova, K. (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) (pp. 4171-4186). Association for Computational Linguistics. https://aclanthology.org/N19-1423/

4. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Huggingface Team, … Amodei, D. (2020). Language models are few-shot learners. arXiv preprint arXiv:2005.14165. https://arxiv.org/abs/2005.14165

5. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. https://arxiv.org/abs/1810.04805

6. Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108. https://arxiv.org/abs/1910.01108

7. Alsentzer, E., Murphy, J. R., Boag, W., Weng, W. H., Jin, D., Naumann, T., & McDermott, M. (2019). Publicly available clinical BERT embeddings. arXiv preprint arXiv:1904.03323. https://arxiv.org/abs/1904.03323

8. Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. arXiv preprint arXiv:1908.10084. https://arxiv.org/abs/1908.10084

9. Ahsan, M., Ghose, T., Rahman, M. M., & Khan, M. R. (2021). HealthBERT: A Transfer Learning Approach for Predicting Patient-Doctor Conversations. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '21). Association for Computing Machinery, New York, NY, USA,2140-2148. https://doi.org/10.1145/3447548.3467271.