Unsupervised Topic Modeling vs. Transportation

Extracting transportation-related Social Determinants of Health (SDoH) from MIMIC-III doctor's notes

Ajay Balaji
School of Information
University of Texas
Austin, Texas USA
ajay.balaji@utexas.edu

Noel Charlie
School of Information
University of Texas
Austin, Texas USA
noel.charlie25@utexas.edu

Haley Triem
School of Information
University of Texas
Austin, Texas USA
haleytriem@utexas.edu

ABSTRACT

This study aimed to address the issue of identifying transportation-related Social Determinants of Health (SDoH) in doctor's notes to improve transportation equity in healthcare. Using natural language processing, word embeddings, and unsupervised machine learning methods, the study analyzed the MIMIC-III dataset and evaluated the performance of two models: Latent Dirichlet Allocation (LDA) and BERTopic. Results showed that the identification of transportation related SDoH in unstructured notes is challenging due to language overlap and complexity in patient journeys. However, the study demonstrated the potential of these methods to inform further research in transportation equity in healthcare. The coherence score of the LDA model was 0.43, while BERTopic generated topics that were thematically similar to one another. The study highlights the need for improved identification of transportation related SDoH to address healthcare disparities.

CCS CONCEPTS

• Health care information systems • Health Informatics • Natural Language Processing • Unsupervised Learning and Clustering

KEYWORDS

Social Determinants of Health, SDoH, Transportation, Health Access, Word Embeddings, Unsupervised Topic Modeling, Latent Dirichlet Allocation (LDA), BERTopic, MIMIC-III

1 INTRODUCTION

Reliable transportation services are fundamental to healthy communities. However, "each year, 3.6 million people in the United States do not obtain medical care due to transportation issues" [1]. Although transportation is the third most cited barrier to accessing health services for older adults [2], it is often neglected.

Furthermore, studies show that ethnic minorities, women, older people, those with lower-income, those with less education, and those experiencing homelessness are disproportionately affected by transportation-related barriers to health access [3][4][5]. Environmental issues such as proximity to care [2] and access to a car as opposed to public transport [3] augment equity disparities amongst minority populations as patients struggle to get healthcare for issues perpetuated by systemic inequity.

These disparities have consequences. The most common barrier to follow-up care in a study on diabetes patients was lack of transportation [6]. Children whose families experience transportation barriers are far less likely to be brought in for medical concerns [7], and patients who have difficulty visiting the pharmacy "are [35%] less likely to fill prescriptions" [8].

Yet while transportation equity in health is an indisputable priority, it is often difficult to identify patients that might be experiencing transportation-related healthcare barriers. Electronic Healthcare Records (EHRs), while rich in vitals and chart events, patient diagnoses, and base demographics, often fail to include flagging for Social Determinants of Health (SDoH) that might prove critical to helping patients. Medical notes, written in natural language, are usually a good source for extraneous patient data, but are not practical for large-scale analysis of demographic needs, or even in predicting whether a patient might come across transportation-related barriers in the future.

Because of this, we wanted to utilize the vast amount of doctor's notes stored in MIMIC-III—one of the largest, publicly open datasets of de-identified patient data—to see if we could solve this issue. Through natural language processing, word embeddings, and unsupervised machine learning methods such as topic extraction, we sought to locate transportation related SDoH hidden in unstructured data.

2 RELATED WORK

Several studies influenced the direction of our work. In an overview of extracting social determinants of health using NLP, Patra et al. [2021] leveraged queries for SDoH keywords, creating "lexicons for each SDoH category [and] developing rule-based or supervised systems to locate clinical notes associated with" that category [9]. Due to these findings, we wondered if spaCy, scispaCy, and manually generated word embeddings were viable options when searching for SDoH keywords.

Another paper recommended using BERTopic as a means to natural topic modeling using class-based TF-IDF procedure [10]. It pointed out that a transformer-based language model paired with the power of TF-IDF embeddings should be able to cluster "topic representations" and generate "coherent topics" based off of those

clusters. Finally, another piece recommended using Latent Dirichlet Allocation (LDA), a "generative probabilistic model" that uses inference techniques useful for estimating Bayes parameters, which outputs the probability of random variables with unknown parameters [11]. In order to evaluate our models, we referenced Rosner et al.'s [2014] work regarding coherence scores, which "pair complex word scorings and apply them to topic scoring" [12]. The combination of these unsupervised methods (BERTopic and LDA) measured by coherence scores would be placed in conjunction with our base technique of using NLP libraries such as spaCy and scispaCy to generate word-embedding scores that evaluate similarities between chosen keywords and doctor's notes.

Finally, along with the above methodological papers, we relied heavily on sociological papers mentioned in the introduction to inform our knowledge of transportation related SDoH as a whole and help us better understand which keywords to target. Social Determinants of Health are nuanced issues, in which it is best to marry computational methods with a deep understanding of the underlying web of personal factors that affect a patients' health.

3 METHODOLOGY & EXPERIMENTS

As informed by our research, we approached our problem using two main methodologies: word embeddings and unsupervised topic modeling. Using word embeddings felt like a natural first step because there was potential for relationships between transportation words to be identified systemically, rather than predictively. This technique then could be used as a baseline to evaluate our unsupervised models, because word embeddings are commonly used to manually model topics and word relationships across a corpus. Unsupervised learning techniques were a sensible second phase to the project, because there is no way to feasibly flag a large corpus of documents accurately and quickly, given that there is a discrepancy between medically-related transportation words and SDoH-related transportation words (i.e. "Patient should not drive for three days" vs. "Patient has no way to drive to the hospital"). Because of a lack of an accurately flagged corpus of transportation-related documents, supervised learning was not the most viable method to tackle this problem.

3.1 Data

We utilized the MIMIC-III clinical database, "a large, freely available database comprising deidentified health-related data associated with over forty thousand patients" [13]. For this project, we focused on the "NOTEEVENTS" table, which contains 2,083,180 naturally worded text notes for patients [14]. In order to prepare our dataset, basic data preprocessing tasks for text cleaning were used. These methods include removing stop words, lemmatizing the text, and stemming the text. Notably, we removed de-identified information, coded in brackets, as well as common words found in every text like "Age", "Gender," etc. Finally, the text was tokenized, in order to be fed into our model.

3.2 Environment

The environment setup was relatively simple. Word embeddings required installing spaCy and scispaCy, as well as language models "en_core_web_lg" and "en_core_sci_lg." TF-IDF and Word2Vec libraries were used to generate relationships between words. The unsupervised portion required packages such as NLTK, Gensim, Scikit-learn, and BERTopic. All experiments were run with Python, using Google CoLab.

3.3 Word Embeddings

In order to gain insight into our corpus and develop a base model to compare unsupervised methods, we used word embeddings to generate cosine likelihood scores of subsequent documents containing transportation-related content. We employed four sub-methods: spaCy, scispaCy, custom embeddings generated off of a large corpus of doctor's notes, and custom embeddings generated off of known transportation-related doctor's notes.

For all models, we chose target words for the model to base its search off of, aka keywords. We generated target embeddings using each sub-method outlined above. For spaCy and scispaCy, we used pre-built models, and for custom embeddings, we utilized TF-IDF scores to create Word2Vec embeddings. Then, for all methods, we located centroids to map relationships between words. Later, we iterated through documents to calculate cosine similarity scores, or, the likelihood that a document contains transportation-related SDoH. Finally, we were able to attach cosine similarity scores to each document and pull the most "likely" transportation documents for examination.

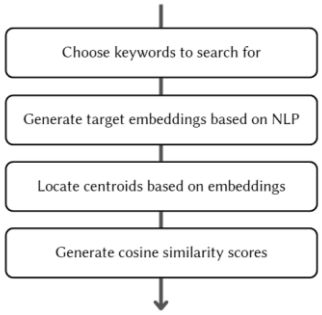

Figure 1: Method for generating embeddings.

3.4 Unsupervised Topic Modeling

One of the unsupervised learning methods commonly used in Natural Language Processing is topic modeling and extraction. Topic modeling is the process of extracting similar words and grouping them into topics (akin to clustering) based on the probability distributions of the topics generated and the words in the document corpus.

a. Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) was one of the first topic modeling algorithms used in Information Retrieval systems to display relevant results on search queries. This algorithm is a generative probabilistic algorithm that initially assumes that each

document in the corpus is a mixture of a small number of topics. There are 4 stages in this algorithm:

1. Initialization: randomly assigning each word in a document a topic
2. Topic distribution: computing probability distribution of topics for each document
3. Word distribution: computing probability distribution of words in each topic
4. Iteration: iteratively assigning each word a topic based on the probability distribution of words and topics

The Latent Dirichlet Allocation algorithm is a pre-built function from the Gensim library. We wrote a wrapper function to accommodate all parameters passed to the LDA function. To find the optimal number of topics, we ran the algorithm with a different number of topics and generated a coherence score each time, picking the ideal number of topics as the one with the highest coherence score. For our dataset and use case, the ideal number of topics was 12.

```
lda_model = genism.models.LdaModel(doc_term_matrix,
                                   num_topics = 12,
                                   id2word = dictionary,
                                   passes = 10,
                                   chunksize=10,
                                   update_every=1,
                                   alpha='auto',
                                   per_word_topics=True,
                                   random_state=42)
```

Figure 2: Method for training an LDA Model.

b. BERTopic

BERTopic is another topic extraction algorithm, based on the BERT (Bidirectional Encoder Representations from Transformers) architecture to generate word embeddings for each document in the corpus. Once these embeddings are generated, they are passed to clustering algorithms such as the k-means algorithm, with the number of clusters in this case determined using elbow method or other methods. The result of this clustering algorithm are topics, which BERTopic can then generate meaningful names for. One of the advantages of BERTopic over LDA is that BERTopic can handle complex topics and large numbers of documents. Furthermore, not predefining the number of topics means that it can find the optimum number of topics on its own. These advantages yield a highly flexible model that provides more accurate results.

BERTopic is available to use as a library in Python from Pypi. Once we had installed the library, we passed the processed text as the input using the vectorizer model. The vectorizer model was used to generate dense vector embeddings of the documents in the corpus using BERT transformers available in HuggingFace. It then generated embeddings of a fixed size using the BERT transformer by converting them to subwords, tokenizing these subwords, and feeding them to the BERT model. The output was a set of vectors of word embeddings which could be used to perform other tasks in various use cases such as Information Retrieval.

```
Bertopic_model=BERTopic(vectorizer_model=vectorizer_model,
                        Language='english',
                        calculate_probabilities=True,
                        verbose=True)

topics, probs = berttopic_model.fit_transform(notes)
```

Figure 3: Method for training BERTopic.

4 RESULTS

4.1 Embeddings

When reading through the top ten cosine-scored notes for each model (spaCy, scispaCy, custom embeddings, and custom transportation embeddings), we noted any times the model *positively* identified transportation need, identified *potential* transportation need, and actively *misidentified* transportation need. This is where the issues of tracking transportation-related social determinants of health became especially apparent: language surrounding transportation needs often overlaps with language surrounding questions of mobility.

For example, a patient might "not have means to drive to the hospital," but also "might not be able to drive for five days post-operation". While the denotations of the word "drive" are the same in both examples, the implications behind it are very different. This is further complicated when considering that a large-scale database like MIMIC-III contains a network of tables that represent the patients' multiple stays in the hospital, time series input and output events, and patient locations across different departments during their stay. A patient might be "transferred to the Oncology division," but also might "require assistance transferring from a smaller clinic to a larger research clinic," two drastically different meanings of the word "transfer." Innumerable input and output variables for medication as it travels through the body are essentially indiscernible.

spaCy	scispaCy
"take taxicab to ICU" † "no bicycling allowed"- "may not drive"- "vehicle accident"- "rehab placement"*	"social history homeless"* "cab voucher" † "suicide by train"- "public transportation" † "drive every day" †
Word2Vec: entire corpus	**Word2Vec: transportation corpus**
"await rehab placement"* "make go home"- "transfer floor care"- "home schedule"- "home follow up"-	"transfer neurology"- "bring ambulance"* "extend care facility"* "currently live rehab"* "send home stable"-

Figure 4: Examples of positively identified (green †), misidentified (red -), and uncertain (orange *) transportation needs extracted from highly scored cosine-related documents.

In the top ten cosine-scored documents, our results were as follows, with scispaCy positively identifying the most transportation-related social needs, at a low two out of ten documents identified. Having two out of the ten "chosen" needs in the whole corpus is admittedly

not a good score, however this method has potential to inform further research.

	Correctly Identified	Potentially Correct	Incorrectly Identified	Uncertain Correctness
spaCy	1	2	6	1
scispaCy	2	3	4	1
Custom	0	1	6	3
Transportation	0	2	4	4

Figure 5: Identification of transportation need (columns) within the top ten cosine-similar documents between different methods of embedding (rows). Correct example: "Patient needs help driving to hospital." Incorrect example: "Patient should not drive 3 days post op."

4.2 Topic Modeling

For topic modeling, there were not many evaluation metrics that we could use to evaluate how our model performed. For LDA, we identified a metric called the coherence score which represents the degree of coherence and interpretability within topics and documents by using similarity metrics. We obtained a value of 0.43 for our LDA model—The closer the coherent score is to 1, the better the extracted topics are, mathematically. BERTopic, on the other hand, was easier to evaluate. By eyeballing the keywords generated in each topic and figuring out whether these words are similar to one another, we were able to see if topic clusters were viable for extracting necessary information. Three of our six topics contained "transportation," giving insight into which words are potentially being grouped with transportation by the model.

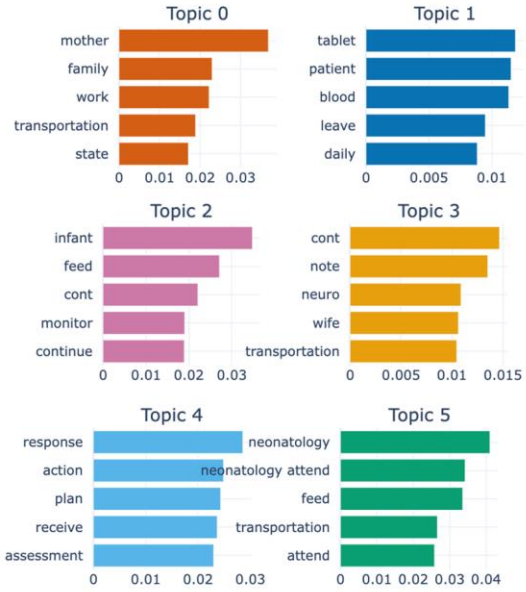

Figure 6: BERTopic generated topics.

5 CONCLUSION

While embeddings have potential to be useful *aspects* of other methods to locate transportation-related SDoH, issues arose given a disparity between medically related transportation words and SDoH-related transportation words. Custom embeddings were significantly less accurate in locating documents than spaCy and scispaCy models, likely due to the large corpus in which spaCy and scispaCy are trained.

Unsupervised learning models such as LDA and BERTopic were a much better option, and identified topics that were related to each other, with BERTopic pulling transportation-related topics from three out of six generated word scores. While we recognize that supervised learning has great potential in this field, we had no means to accurately identify which patients definitively needed transportation assistance without input from a medical professional. Our future research would prioritize contextualizing transportation-related words, and entail creating models that understand which words implicate a patient as requiring help.

Briefly, we entertained using ChatGPT as a means to flag documents in our corpus as related to transportation SDoH or not. After running some tests, however, we noticed GPT-3.5 was unreliable, as it threw errors and hallucinations that would be tricky to fact check given a large corpus annotated by ChatGPT. Given the advancements of GPT-4, and its increased knowledge of medical terminology [15], harnessing its power could prove successful for flagging transportation-related social determinants of health.

Beyond the scope of computationally solving the issue, our study made us consider larger social implications. Given that the main limitations behind uncovering transportation related SDoH were due to lack of labeled data, systems should be developed that place this issue as priority and preemptively flag data. Metadata such as SDoH flags as parts of patient portals are potential avenues to rectify a lack of information on Social Determinants of Health. While this is a long-term and admittedly idealistic solution, reframing the healthcare landscape to one where we prioritize SDoH as much as physical determinants of health could lead to a future in which disease is preventatively handled and peoples' social lives are as protected as their medical ones.

ACKNOWLEDGEMENTS

This paper was produced for the course *AI in Healthcare* taught by Dr. Ying Ding at the School of Information, University of Texas at Austin.

REFERENCES

[1] American Hospital Association, 2017. Social Determinants of Health Series: Transportation and the Role of Hospitals. *AHA/HRET* (Nov, 2017). Website: https://www.aha.org/ahahret-guides/2017-11-15-social-determinants-health-series-transportation-and-role-hospitals

[2] Samina T. Syed, Ben S. Gerber, and Lisa K. Sharp, 2013. Traveling Towards Disease: Transportation Barriers to Health Care Access. *J Community Health* 38, 5 (Oct, 2013), 976-993. DOI: https://doi.org/10.1007%2Fs10900-013-9681-1

[3] Diana Silver, Jan Blustein, and Beth C. Weitzman, 2010. Transportation to Clinic: Findings from a Pilot Clinic-Based Survey of Low-Income Suburbanites. *Journal of Immigrant and Minority Health* 14 (Apr. 2012), 350-355. DOI: https://doi.org/10.1007/s10903-010-9410-0.

[4] Donna L. Washington et al., 2011. Access to Care for Women Veterans: Delayed Healthcare and Unmet Need. *J Gen Intern Med* 26, 2 (Nov, 2011), 655-661. DOI: https://doi.org/10.1007%2Fs11606-011-1772-z.

[5] L. G. Branch and K. T. Nemeth, 1985. When elders fail to visit physicians. *Med Care* 23, 11 (Nov, 1985), 1265-1275. DOI: https://doi.org/10.1097/00005650-198511000-00005.

[6] Kate Wheeler et al., 2007. Inpatient to Outpatient Transfer of Diabetes Care: perceptions of Barriers to Postdischarge Followup in Urban African American Patients. *Ethnicity & Disease* 17, 2 (Spr. 2007), 238-243. Link: https://www.jstor.org/stable/48667072.

[7] Glenn Flores, Milagros Abreu, Mary Anne Olivar, et al., (1998). Access Barriers to Health Care for Latino Children. *Arch Pediatric Adolescent Med.* 152, 11 (Nov 1998), 1119-1125. DOI: 10.1001/archpedi.152.11.1119.

[8] Sunil Kripalani, Laura E. Henderson, Terry A. Jacobson, Viola Vaccarino, 2008. Medication use among inner-city patients after hospital discharge: patient-reported barriers and solutions. *Mayo Clinic Proc.* 83, 5 (May, 2008), 529-535. DOI: 10.4065/83.5.529.

[9] Braja G. Patra et al., 2021. Extracting social determinants of health from electronic health records using natural language processing: a systematic review. *Journal of the American Medical Informatics Association* 28, 12 (Dec. 2021), 2716-2727. DOI: https://doi.org/10.1093/jamia/ocab170.

[10] Maarten Grootendorst, 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv* (Mar. 2022). DOI: https://doi.org/10.48550/arXiv.2203.05794.

[11] David M. Blei, Andrew Y. Ng, Michael I. Jordan, 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* (Jan. 2003). .pdf: https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf.

[12] Frank Rosner et al., 2014. Evaluating topic coherence measures. *arXiv* (Mar. 2014). DOI: https://doi.org/10.48550/arXiv.1403.6397.

[13] Alistair Johnson, Tom Pollard, Roger Mark, 2016. MIMIC-III Clinical Database. *PhysioNet* ver. 1.4 (Sept. 2016). Link: https://physionet.org/content/mimiciii/1.4/.

[14] MIT-LCP, 2023. NOTEEVENTS Documentation. *MIMIC* (2023). Link: https://mimic.mit.edu/docs/iii/tables/noteevents/.

[15] OpenAI, 2023. GPT-4 (Mar. 2023). Link: https://openai.com/research/gpt-4.