Max Swinton
BAIS:6100
Final Project
5/3/2025

Exploratory Analysis of Electronic Health Records

1 Introduction

In this paper, we present an analysis of a patient clinical notes dataset aimed at understanding the nuances of clinical record language in its structured and unstructured format. The dataset used for this study is detailed in Kumar et al. Creation of a new Longitudinal Corpus of Clinical Narratives [2015]'. With our analysis, we aim to answer the question: “How can we parse the raw text of electronic health records to gather useful clinical information, find common visit themes, and understand trends in doctor sentiment over time?”

To answer this question, we complete several tasks, namely generating frequency distributions of vital signs and extracting social history records from the free-text portions of the records, and conducting various analyses on these extracted data points. Specifically, we identify patients with blood pressure and social history entries, perform topic analysis on social history data, and analyze provider sentiment across visits.

The dataset for this paper comes from Kumar et al, and comprises 1304 de-identified records of 296 diabetic patients contained in a semi-structured XML format. From Kumar et al. “The corpus contains three cohorts: patients who have a diagnosis of coronary artery disease (CAD) in their first record, and continue to have it in subsequent records; patients who do not have a diagnosis of CAD in the first record, but develop it by the last record; patients who do not have a diagnosis of CAD in any Record.” For the purposes of this paper, we focus on text mining and information extraction regardless of diagnosis, but will specify when the CAD diagnoses are relevant for our analysis.

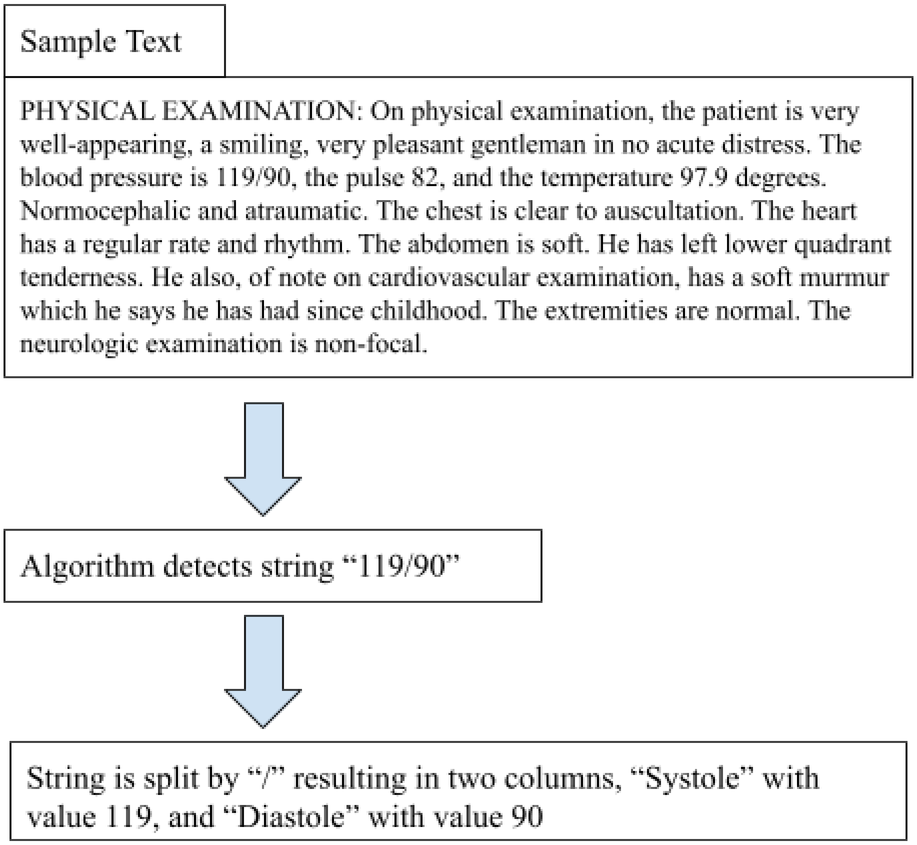
2 Blood Pressure Extraction

The most difficult step in natural language processing is, without a doubt, processing the natural language. Each person has their own unique grammar, and doctors will have their own unique notes. The goal of this section is to extract and analyze vital sign information from the text portions of the files in the dataset. While the medical history and medication information is available in the xml tags and can be easily extracted with relative accuracy, the physical examination data, such as blood pressure can only be found in the free text portion of the file; notes handwritten by each doctor. For this section, we focus on extracting the blood pressure, as other vital signs see too much variability in their labels or are not present in a significant number of records.

For extracting vital signs we use regular expression patterns for digit identification, and identify blood pressure measurements contained in the free text of patient records. We then compute various frequency distributions, and finally test the accuracy of our extraction methods.

Some doctors in the dataset write the blood pressure measurements in a “physical examination” section, others in an “exam” section, while most simply place it wherever they see fit in their notes. Blood pressure may be labeled “bp”, or “blood pressure”, or “BP”, or it may be misspelled. Short of manually annotating every note in the dataset, it would seem nearly impossible that blood pressure data can be reliably extracted from these personal notes. However, one heuristic that is common through nearly all of the notes is that the blood pressure is nearly always written in the form X/Y, where both X and Y are integers.

This is where we start with this analysis. Rather than looking for the text “blood pressure” we instead attempt to identify the location in the text where there are two numbers separated by “/”. We extract this string, split it by the slash mark, and obtain numeric columns, one with the systolic blood pressure, and one with the diastolic blood pressure.



Sample Text

PHYSICAL EXAMINATION: On physical examination, the patient is very well-appearing, a smiling, very pleasant gentleman in no acute distress. The blood pressure is 119/90, the pulse 82, and the temperature 97.9 degrees. Normocephalic and atraumatic. The chest is clear to auscultation. The heart has a regular rate and rhythm. The abdomen is soft. He has left lower quadrant tenderness. He also, of note on cardiovascular examination, has a soft murmur which he says he has had since childhood. The extremities are normal. The neurologic examination is non-focal.

Algorithm detects string “119/90”

String is split by “/” resulting in two columns, “Systole” with value 119, and “Diastole” with value 90

Figure 1: Sample of text parsing algorithm to obtain blood pressure measurements

According to this algorithm, there are 1053 records out of 1304 total records that contain a blood pressure measurement. After removing non-conforming rows where blood pressure is written but not measured numerically, we get 764 observations. The observations are then classified according to the standards of the American Heart Association, where a reading is “Normal”, “Elevated”, “Hypertension 1”, “Hypertension 2”, or “Hypertensive Crisis”. The results are summarized below:

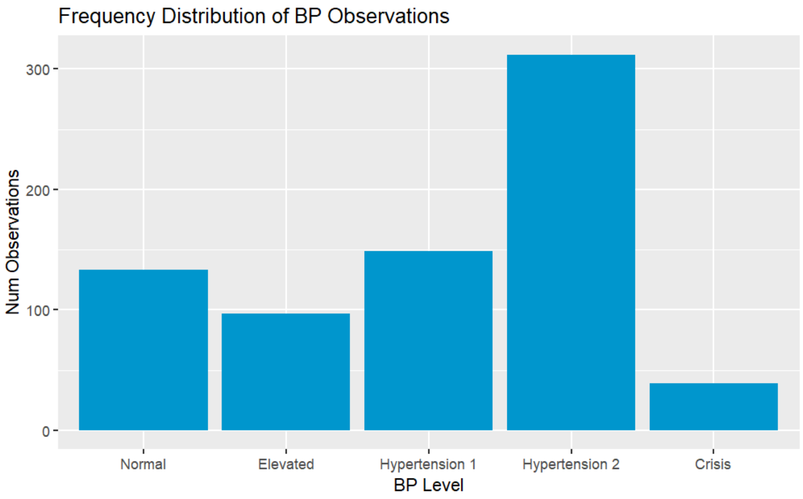
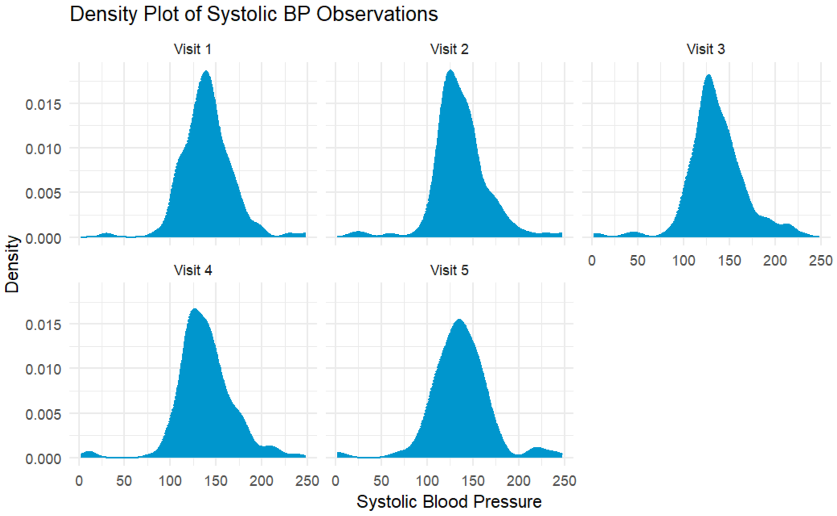


Figure 2: Visualization of Blood Pressure Measurements

The most striking observations initially are that very few observations appear to be within the normal range of blood pressure readings. In fact, a majority of observations are early hypertensive or worse. However, these readings must be taken in context, as this dataset is a collection of records of patients with diabetes, with some either suffering from, or developing, coronary artery syndrome (CAD). As is described in Naha S, Gardner MJ, Khangura D, et al. Hypertension in Diabetes [2021], patients with diabetes are much more likely to suffer from high blood pressure, and if not treated effectively, are very likely to develop further coronary diseases. Thus, the nature of this dataset means that these numbers should be expected, and are not a representative sample of the whole human population.

Additionally, this data is longitudinal, thus most patients have multiple blood pressure measurements included in this dataset. By visualizing all the records, we are seeing patients in different stages of their treatment, perhaps a better visualization would be to sort these observations by visit number.



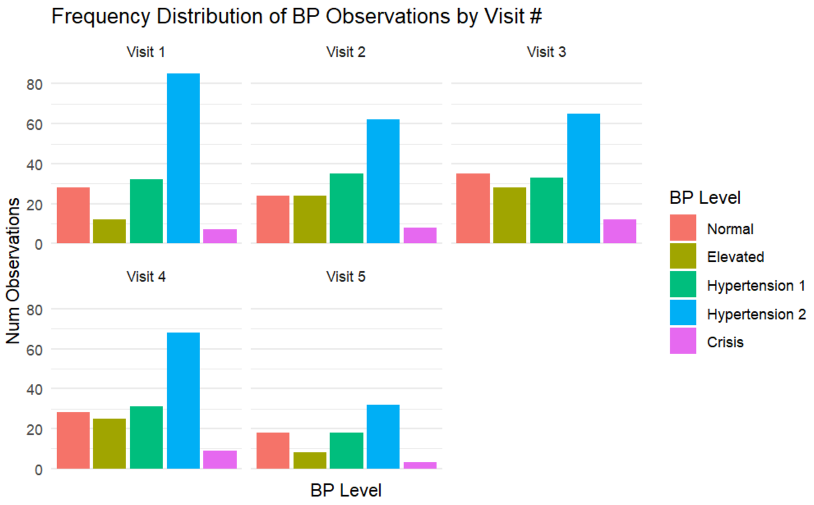


Figure 3: BP Measurements Faceted by Visit

By sorting this way, we can see that many patients first come in with hypertension, with most seeing a marked improvement in their blood pressure levels. Measurements trend toward a normal distribution as visit number increases, however the average blood pressure is still outside the "Normal" range by the final recorded visit.

2.1 Testing

After obtaining the 764 blood pressure measurements, we want to test the results to assess the ability of the algorithm to correctly identify blood pressure data. We randomly select 50 observations from the full corpus of 1304 records to be manually annotated for blood pressure data. The full corpus is chosen for the sample so as to account for possible false negatives. Records may contain multiple blood pressure measurements, in these cases, the chosen measurement is the blood pressure on admit.

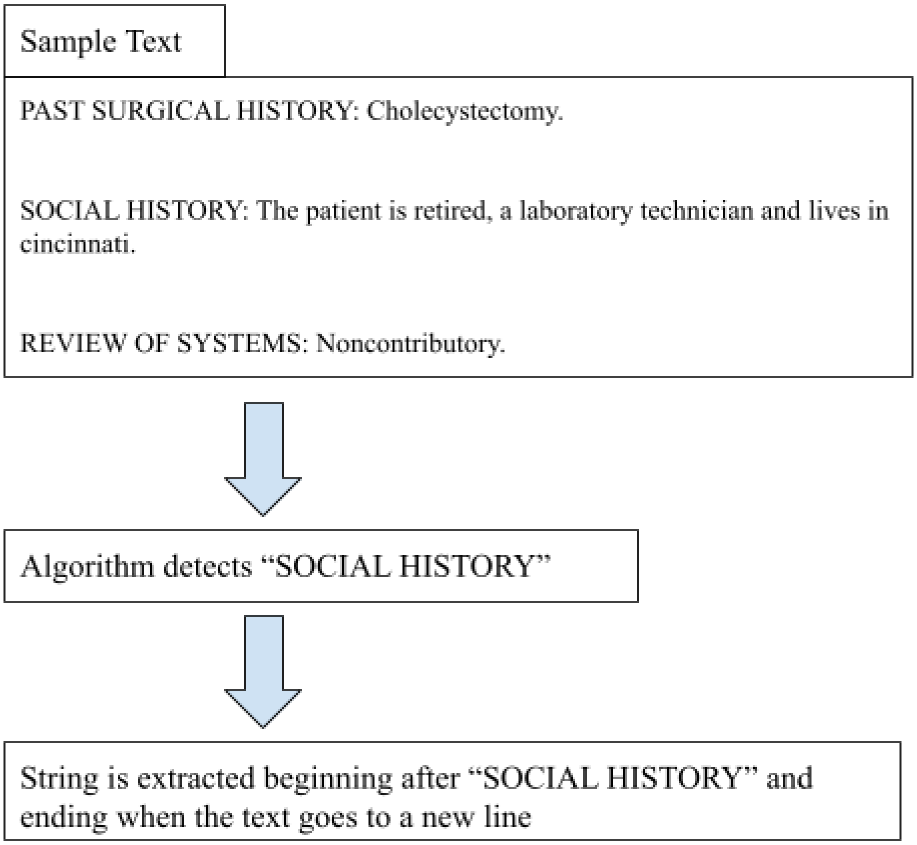
Precision	Recall	F1-Score	Balanced Accuracy
1.00	0.78	0.877	0.89

Table 1: Results of BP Algorithm Against Test Set

We find that this algorithm is very accurate at extracting the correct blood pressure measurement when it detects that a measurement is present. However, it struggles with broad detection of such measurements, often finding no blood pressure reading when in fact one exists in the record. In testing, we found that the algorithm is quite robust in finding a blood pressure measurement so long as text relevant to "blood pressure" is present, but will fail unsurprisingly when a record does not contain any text around the measurement. For example, record 325-04 has the following text: "PHYSICAL EXAM VITALS: 159/93, 98.9, 78, 20, 98%." A human evaluator will easily spot that the blood pressure is the first measurement in this string, however it is not supported in either direction by text referring to this as a blood pressure result. This phenomenon was common enough in the dataset to have an effect on the results of the testing process, with roughly 10% of the test set containing this form of the data. The second phenomenon is when multiple blood pressure results are presented in one record. We found that less than 5% of the data has multiple results, thus it did not warrant extra consideration in the algorithm, however at the small test set size, this can have an outsized impact. Finally, 64% of the records in the test set contained a blood pressure measurement, a similar result to the findings from the algorithmic approach, which detected a blood pressure measurement in 58.5% of the 1304 records.

3 Social History

This section focuses on free-text extraction of social history data from the corpus of 1304 medical record notes. Similar to Section 3, the social history data is not contained within the xml tags of the records, but is instead within the uncurated text section, thus a heuristic is required to identify this data. In examining the records, we found that the most common instances of social history data follow either the string "Social History" or "SH". We first identify these instances, and extract the subsequent text until the file progresses to a new line, indicating that the section has concluded.



Sample Text

PAST SURGICAL HISTORY: Cholecystectomy.

SOCIAL HISTORY: The patient is retired, a laboratory technician and lives in cincinnati.

REVIEW OF SYSTEMS: Noncontributory.

Algorithm detects "SOCIAL HISTORY"

String is extracted beginning after "SOCIAL HISTORY" and ending when the text goes to a new line

Figure 4: Sample Procedure of Social History Algorithm

This proved to be the most effective method to gather the widest breadth of data, without incurring a large penalty of false negative. The algorithm detects 762 records containing social history data, with 281 unique patients represented.

3.1 Social History Topic Modeling

Using the data gathered from the above social history extraction, we then combine the results into one unified record for each patient. In this section, we use these combined records to perform topic modeling using Latent Dirichlet Allocation (LDA).

LDA will take a vectorized corpus of records and attempt to group documents into a number of topics that cover the range of text and are distinct from each other. Each document is placed into one topic, then the words are sorted by relevance. For this work, we take the combined social history data for each patient as one record and create a Document-Topic Matrix. We then assess the model both on its within-topic similarity and between-topic dissimilarity.

In order to determine the optimal number of topics to generate for LDA, we assess a range of topics, from 2 to 10, and record the perplexity scores for each model. Perplexity score attempts to measure the within-cluster semantic similarity of the LDA output, the results are summarized below.

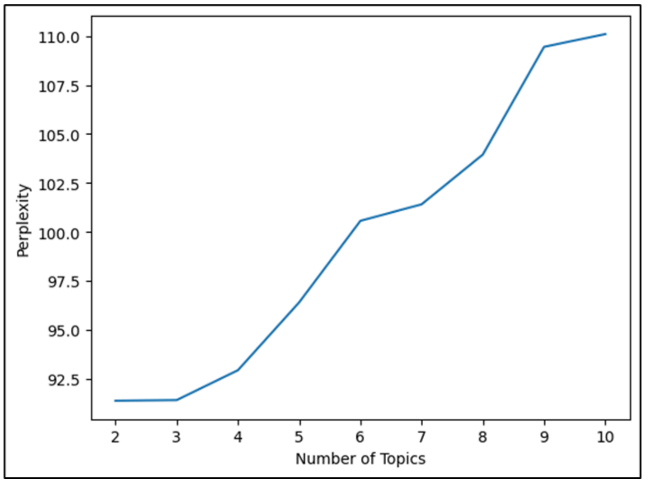
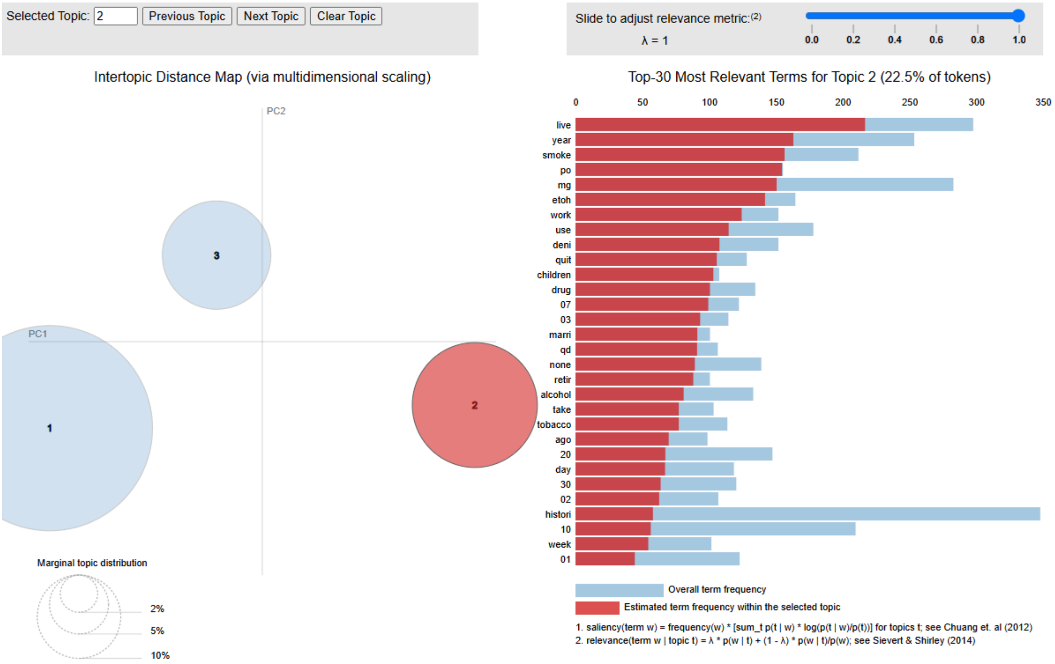


Figure 5: Reported coherence scores from LDA

We choose to use an LDA model with only three topics, as the model shows signs of overfitting with an increase in topics. Generally, LDA will perform better with more topics, however this is understandable in context. The social history records are mostly consistent across patients and doctors, often only including information about patients’ occupation and history of tobacco/drug use. Thus, if the source documents lack variance, as is the case with this dataset, then LDA will perform best when generating fewer topics.

To better understand the results from LDA, we can use the “pyLDAvis” library in python, which allows us to visualize the results generated by LDA on a coordinate plane using vector representations of the topics. We expect to see, with perfect topic generation, maximal coverage of the vector space by topic circles, and distinction between topics.



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 6: Topics Generated from LDA

In this case, with only three topics it is unlikely for the entire vector space to be covered, however we do see significant distinction between the topics. From Figure 6 we can see that the top 30 words in Topic 2 account for 22.5% of terms in the corpus, with the most relevant words being, "live," "year," and "smoke," pointing to common questions that doctors may ask. Another very common term in many of these records is "etoh." This is a common shorthand for ethyl alcohol, which is found in alcoholic beverages such as beer, wine, and liquor. The term is used by doctors to reference alcohol use, a common question asked of patients in this dataset. In the same vein, doctors often inquire into a patient's drug use, thus the LDA algorithm has detected "smoking" as being very relevant to this topic.

Lastly, to assess the performance of the LDA topic modeling, we compute the within-cluster and between-cluster similarity using a cosine similarity measure. Much like a perplexity score, cosine similarity attempts to measure the semantic similarity of words in a document. In this context we would hope to see high cosine similarity within each topic, which indicates well-defined clusters, and low similarity between topics, indicating distinct cluster formation.

	Topic 1	Topic 2	Topic 3	Average
Within-Cluster similarity	0.915	**0.937**	0.904	0.919
Between-Cluster similarity	0.353	0.414	0.330	0.366

Table 2: Results of LDA Topic Modeling

The model used for this topic modeling, from `scikit-learn`, generates word embeddings derived from a training corpus of English-language words and phrases meant for general use and sufficient coverage of everyday language. For many corpora, as with this implementation, the model is not specifically trained on medical language terminology. This can reduce performance on topics involving such language, as is used for this analysis of social history data. Nonetheless, Figure 7 indicates strong positive results from the topic modeling with LDA. The within-cluster similarity is quite high, indicating the topics are well-defined. However, this corpus is relatively small, only containing short social history entries for 281 patients, future research may find that a sufficiently large pre-training corpus of medical literature and social histories will increase performance beyond what is seen in Table 2.

With only three topics, it should be expected that within-cluster similarity would be high for a sufficiently large training corpus. This would also be the expectation when computing the between-cluster similarity scores. Fewer topics reduces the likelihood of overlap, which is what can be seen in both Figure 6 and Table 2. A similarity score of 0 would indicate two orthogonal vectors, and the result for the between-topic similarity score indicates that the vectors are well-defined and sufficiently distinct.

4 Sentiment Analysis of Text Records

The last section of this analysis focuses again on the free-text portion of patient records, with a goal of assessing trends in doctor sentiment across visits. To accomplish this goal, we implement a base sentiment analyzer from the HuggingFace library in python. HuggingFace provides off-the-shelf transformers trained for specific classification purposes, including a model for sentiment analysis. The HuggingFace analyzer will assess each record in the corpus and assign a label of either “Positive” or “Negative” to define the overall sentiment of the record.

Additionally, the model will calculate a “Sentiment Score” to evaluate the strength of the sentiment in the record. The sentiment score ranges from 0 to 1, where a higher score indicates stronger sentiment. For example, a record classified as positive with a sentiment score of 95% would mean that the model assessed the record to be 95% positive sentiment. Below, we visualize the trends in sentiment labels and sentiment scores across visit numbers.

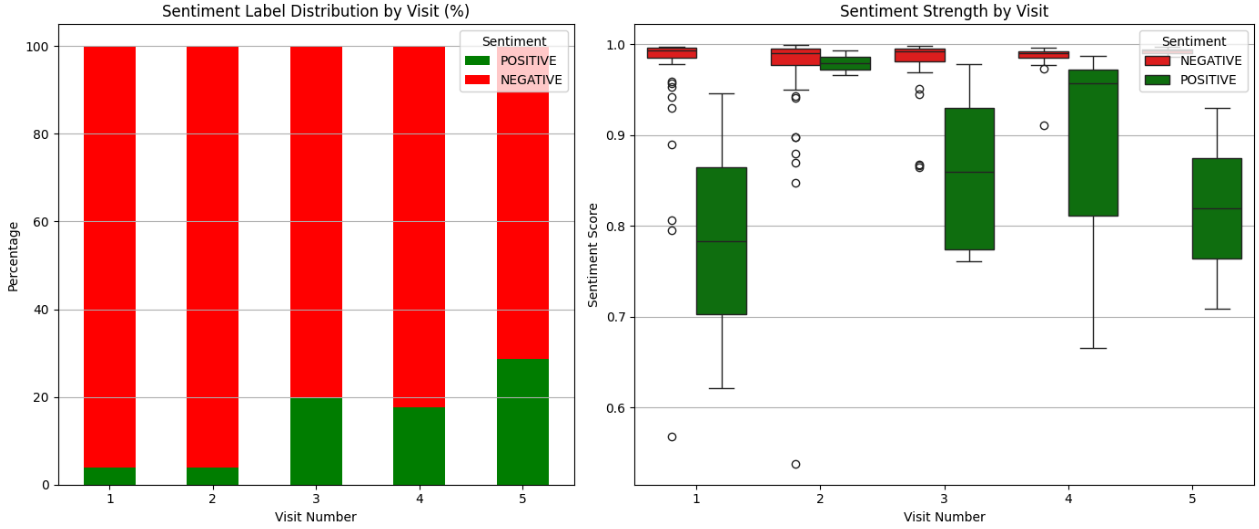


Figure 7: Results of Sentiment Analyzer

In the first plot of Figure 7, a stacked bar chart represents the percentage of records in each visit which are either positive or negative. We see that the sentiment is overwhelmingly (over 90%) negative for patients in their first visit, with an increase in positive sentiment as visits increase. This is understandable in the context of the dataset, as a patient's first visit will typically be when they are showing complications with diabetes, while the following visits are only check-ups. Thus, we can expect that the initial visit will have the most "negative" sentiment, as this is when patients are yet to be treated, and presumably experiencing their most severe symptoms.

In the second plot, we look at the distribution of sentiment scores across visits. For each visit, there are two boxplots for the distributions of negative and positive sentiment scores, respectively. From this, we can see that, when a record is classified as negative, it is very *strongly* negative. This continues across visits, only decreasing in variance over time. Positively classified records, however, show high variance in their sentiment scores. This could be due to a small sample size, as the percentage of records classified as positive is below 10% in the first two visits and only increasing to 30% by the final visit. With such a wide spread in the positive sentiment scores, we are unable to make conclusions about a significant difference in positive sentiment over time.

5 Limitations

The most significant limitation of this research is in the nature of the dataset. While the vague structural requirements of the records is meant to represent real-world data, with only 1304 records, and varying degrees of specificity and structure in each record, it can be very difficult to make broad-scale inference about patients, nor are we able to test unsupervised deep-learning methods that require larger datasets.

Additionally, because this patient data is public, it is highly de-identified and lacks structured demographic or geographic information. This can make it difficult to make accurate inferences for the patient population, or identify any trends that can be applied during patient care.

6Conclusion

Our frequency distributions of vital signs highlight common trends and variations in medical data across the dataset. Extracting social history entries allowed us to quantify the prevalence of this information within patient records, with topic modeling to identify the most common lifestyle traits of patients with diabetes. Lastly, with sentiment analysis, we were able to assess trends in doctor sentiment over time. Overall, this project was enlightening, and helped to practice working with semi-structured data in a practical setting.