**Final Project: Healthcare Reviews**

**Center for Information Systems and Technology, Claremont Graduate University**

**IST 332: Natural Language Processing**

**Dr. Hengwei Zhang**

**Ahmed, Shih Min, Olga, Soren**

**Dec 12, 2025**

[Link to GitHub](https://github.com/slaskin1/NLP_Final_Project)

[Link to Google Drive](https://drive.google.com/drive/folders/16LDkqbbAEDUEG0tUv0ABJvZM-t5PqPur?usp=sharing)

Table of Contents

[Part I: Introduction & Research Question 4](#_Toc216446669)

[**1. Research Question** 4](#_Toc216446670)

[**2. Motivation and Expected Contribution** 5](#_Toc216446671)

[**3. Relevance of the Study** 6](#_Toc216446672)

[Part II: Corpus Creation (Data Collection) 6](#_Toc216446673)

[Part III: Text Preprocessing 8](#_Toc216446674)

[Part IV: Data Understanding 11](#_Toc216446675)

[**1. Data Quality Verification** 11](#_Toc216446676)

[**2. Exploratory Review Analysis** 12](#_Toc216446677)

[**4. Data Preparation for Downstream Tasks** 14](#_Toc216446678)

[Part V. Sentiment Analysis 15](#_Toc216446679)

[Part VI: Topic Modeling 16](#_Toc216446680)

[Part VII. Supervised Learning 18](#_Toc216446681)

[VIII. Deployment Plan 19](#_Toc216446682)

**List of Figures**

[Figure 1: Lexical Diversity Distribution 11](#_Toc216446918)

[Figure 2: A Summary Table for Exploratory Review Analysis 12](#_Toc216446919)

[Figure 3: Star Rating Distribution 12](#_Toc216446920)

[Figure 4: Correlation between Polarity & Review Rating 15](#_Toc216446921)

[Figure 6: Supervised Learning Results 18](#_Toc216446922)

**List of Tables**

[Table 1: Metadata Summary 6](#_Toc216446924)

[Table 2: Average tokens per reviews & vocabulary size 9](#_Toc216446925)

[Table 3: Star Rating Distribution 12](#_Toc216446926)

[Table 4: A summary table of Sentiment Analysis Results 14](#_Toc216446927)

[Table 5: Coherence Score 16](#_Toc216446928)

[Table 6: Top 10 Words for Each Topic 17](#_Toc216446929)

# **Part I: Introduction & Research Question**

Accessing urgent care services is often associated with high stress, uncertainty, and time sensitivity. Patients typically visit urgent care when they are sick, injured, or dealing with sudden health concerns, which heightens emotions and amplifies expectations about service quality, wait times, and staff responsiveness. Because of this inherently stressful context, patient experiences tend to be polarized—individuals who receive fast, attentive care often leave extremely positive feedback, while those who encounter delays or poor communication frequently respond with strong dissatisfaction. This dynamic leads to a review landscape dominated by 1-star and 5-star ratings, with relatively few moderate evaluations.

Within this environment, user-generated online reviews offer a rich, publicly available source of information that captures both the emotional intensity and the nuanced details of real patient experiences. These reviews contain quantitative indicators (such as star ratings) and qualitative narratives that describe perceptions of wait times, staff professionalism, medical competence, and overall service quality.

This project examines whether natural language processing (NLP) techniques can be applied to these textual reviews to predict the quality of care delivered by urgent care centers. By analyzing patient narratives from urgent care clinics across California, the study seeks to uncover linguistic and emotional patterns that correspond to high- and low-quality experiences.

**1. Research Question**

Primary Research Question (Review-level prediction)

How well can sentiment analysis and NLP techniques applied to Google Maps patient reviews predict medical center (urgent care focus) quality-of-care ratings?

Predict the individual review\_rating (1–5) from review\_text

* patient reviews = review\_text
* quality-of-care ratings = review\_rating
* What linguistic, emotional, and thematic patterns in patient reviews are most strongly associated with high (5-star) and low (1-star) urgent care ratings?

In other words, the sentiment in the text explains why this particular patient gave 1 vs. 5 stars.

Secondary Research Question (Facility-level prediction)

Can online sentiment predict the overall averaged quality of selected urgent care?

* Sentiment scores per user → predict review\_rating

Doing both makes the research much stronger than doing only the primary one.

**2. Motivation and Expected Contribution**

The project aims to contribute:

* Evidence of whether review text alone can reliably predict star ratings in a setting known for polarized experiences.
* Identification of language patterns that correspond to positive elements (e.g., quick service, friendly staff) and negative elements (e.g., long waits, lack of communication).
* Practical insights that urgent care clinics can use to identify service gaps, improve patient care, and understand expectations during stressful healthcare encounters.
* A demonstration of the usefulness of public online reviews for healthcare quality assessment.

**3. Relevance of the Study**

With increasing dependence on online platforms for selecting healthcare providers, understanding the factors that drive patient satisfaction is more important than ever. The urgent care environment - emotionally charged, time-sensitive, and highly variable - creates a natural testbed for evaluating the power of NLP to interpret patient sentiment. By uncovering patterns across thousands of reviews, this study supports both patient decision-making and provider efforts to improve urgent care service delivery.

# **Part II: Corpus Creation (Data Collection)**

The dataset for this project was collected from Google Places API, which provides public information on businesses. Google Places was selected because it is a free public source that provides structured access to a large number of user-generated reviews. It also includes consistent metadata and clear documentation. The data includes patient reviews for medical centers located in multiple cities across Southern California, including such cities as Los Angeles and San Diego.

The dataset was collected using a Python script that accessed several Google APIs to gather medical care reviews across California. First, each city name was converted into geographic coordinates using the Geocoding API, and these coordinates were then used as the center point for Places searches. For each city, the script used the Places Text Search API to search for clinics using terms like “urgent care,” “walk-in clinic,” “medical clinic,” and “emergency clinic.”

All data used in this project comes from public Google Maps business listings accessed through official APIs. The analysis focuses only on review text and ratings, and no attempt is made to identify or track any individuals. The dataset is used strictly for the project purposes.

After combining all search queries and all listed locations in California, the final dataset consists of 12091 raw reviews.

The table below shows the summary of metadata of our dataset

Table 1: Metadata Summary

|  |  |
| --- | --- |
| **Feature** | **Description** |
| **Place ID** | The unique identifier |
| **Place Name** | Business name that is shown in Google maps |
| **Place\_address** | The address of the urgent care |
| **Place\_avg\_rating** | The average rating for each urgent care |
| **Place\_user\_ratings\_total** | The sum of ratings from all reviewers for each urgent care |
| **Place\_types** | Categories assigned by Google that describe the type of business (e.g., doctor, health, establishment). |
| **Review\_rating** | Numeric rating from 1 to 5 given by the reviewer |
| **Review\_text** | Free-text patient feedback. |
| **review\_relative\_time\_description** | shows how long ago the review was written, using natural wording like “2 months ago” or “3 days ago.” |

# **Part III: Text Preprocessing**

The dataset contained **12,091 raw reviews**, of which **11,658 included usable text**. The following preprocessing steps were applied to standardize and clean the corpus before analysis:

* Removal of empty and missing reviews (433 removed)
* Lowercasing all text
* Contraction expansion
* Expansion of medical abbreviations (e.g., “dr” → “doctor”) using a custom abbreviation dictionary.
* Removing the doctors’ names
* Tokenization using spaCy
* Removal of URLs, symbols, and punctuation
* Stopword removal
* Removal of tokens containing digits and tokens shorter than two characters
* Lemmatization.
* Reconstruction of cleaned text for downstream tasks

These preprocessing steps ensure that the review text is clean, consistent, and suitable for sentiment analysis, topic modeling, and supervised learning. First, empty and missing reviews were removed to eliminate unusable entries. All text was then lowercased to standardize tokens and ensure that words with different capitalization forms are treated consistently. Contractions were expanded to preserve negation and reduce ambiguity, and medical abbreviations (e.g., *“dr” → “doctor”*) were expanded using a custom dictionary to improve semantic clarity. Doctor names were removed because they do not contribute meaningful information to NLP tasks and would otherwise introduce unnecessary noise.

The text was then tokenized using spaCy, allowing subsequent cleaning steps to operate at the word level. URLs, symbols, and punctuation were removed to reduce noise commonly found in user-generated content. Stopwords were removed to focus analysis on more meaningful content words, and tokens containing digits or fewer than two characters were removed because they provide little interpretive value in this context. Lemmatization was applied to normalize words to their base forms, improving vocabulary consistency. Finally, cleaned tokens were reconstructed into processed text suitable for downstream modeling tasks.

Together, these steps reduce noise, enhance vocabulary quality, and produce a well-structured corpus that supports accurate and efficient NLP analysis.

* **Summary Statistics**
  + ***Vocabulary Size***

The processed corpus contains **15,798 unique terms**, indicating that the cleaning pipeline effectively preserved meaningful linguistic content while reducing noise such as punctuation, stopwords, and personal names. The vocabulary size also reflects the impact of abbreviation expansion, which replaced shortened medical terms (e.g., *dr → doctor*) with their full forms, improving consistency without inflating the vocabulary artificially.

* + ***Average Tokens per Review***

After all preprocessing steps, reviews contain an average of **approximately** **42 tokens**. This confirms that the reviews are generally short to medium in length, consistent with typical user-generated feedback found on Google Maps. The average token count remains stable despite normalization, demonstrating that steps like abbreviation expansion and name removal did not artificially inflate review length.

Table 2: Average tokens per reviews & vocabulary size

|  |  |
| --- | --- |
| **Metric** | **Value** |
| Vocabulary Size | 15,798 |
| Average Tokens per Review | 41.32 |

* + ***Most Frequent Words***

The cleaned corpus reveals frequent terms such as **doctor**, **care**, **time**, **staff**, **wait**, and **appointment**, which reflect common themes in patient feedback experience with healthcare providers.

Top 20 Most Frequent Words

1. **doctor –** 11,046
2. **care –** 9,828
3. **go –** 6,512
4. **time –** 5,908
5. **get –** 5,716
6. **take –** 4,692
7. **wait –** 4,377
8. **make –** 4,017
9. **see –** 3,918
10. **patient –** 3,824
11. **appointment –** 3,756
12. **would –** 3,694
13. **need –** 3,607
14. **come –** 3,457
15. **urgent –** 3,442
16. **call –** 3,410
17. **say –** 3,296
18. **experience –** 3,199
19. **tell –** 2,867
20. **medical –** 2,763

# **Part IV: Data Understanding**

This section focuses on evaluating the structure and quality of the dataset and understanding key characteristics of the review text. These insights help determine whether the data is suitable for sentiment analysis, topic modeling, and supervised learning.

### **1. Data Quality Verification**

Several checks were conducted to ensure the dataset was complete and reliable:

* **Missing data:** Reviews with missing text were removed.
* **Review length outliers:** Reviews containing no meaningful content (e.g., 1 word) were identified and removed.

After cleaning, **11,658 valid reviews** were retained for analysis.

### **2. Exploratory Review Analysis**

#### **Review Length and Lexical Diversity**

* After preprocessing, reviews contained **approximately 42 tokens on average**, indicating that most reviews provide enough information for NLP analysis.
* Most reviews show a relatively high lexical diversity, usually between **0.70 and 0.95**. This makes sense because the cleaned reviews are short and mainly contain meaningful words, so there isn’t much repetition. The distribution is right-skewed, with many reviews having mostly unique tokens after preprocessing. Only a small number of reviews show lower diversity, which usually happens when the review is very short or repeats the same terms. Overall, this indicates that the reviews provide enough variety in the wording to support meaningful NLP analysis.

A graph of a number of different colored lines

AI-generated content may be incorrect.

Figure 1: Lexical Diversity Distribution

A summary table was generated showing raw review length, cleaned review length, and lexical diversity for each review. The following figure shows the results for the first 5 rows

A screenshot of a black screen

AI-generated content may be incorrect.

Figure 2: A Summary Table for Exploratory Review Analysis

#### **3. Star Rating Distribution**

Table 3: Star Rating Distribution

|  |  |
| --- | --- |
| **Start Rating** | **Number of Reviews** |
| 1 | 3792 |
| 2 | 401 |
| 3 | 269 |
| 4 | 378 |
| 5 | 6818 |

A bar graph with blue rectangles

AI-generated content may be incorrect.

Figure 3: Star Rating Distribution

This pattern shows that users are more likely to leave feedback when they have very positive or very negative experiences.

### **4. Data Preparation for Downstream Tasks**

The cleaned data was prepared specifically for each type of analysis:

Sentiment analysiswas performed using only the essential cleaning steps to keep the text as close as possible to the original wording. We expanded contractions, converted medical abbreviations to their full forms (e.g., “dr” → “doctor”), lowercased the text, and removed doctor names. These steps help the sentiment model correctly interpret key terms without losing important context. Beyond this, we avoided heavy preprocessing so that emotional cues, phrasing, and natural expressions in the reviews remain intact. This approach supports more accurate polarity scoring because the model can analyze the full sentiment expressed by the reviewer.

**Topic modeling** relied on the fully cleaned version of each review to ensure clear and meaningful patterns. This included tokenization, lemmatization, abbreviation expansion, removal of doctor names, and reconstruction into cleaned text strings. Using this deeper level of preprocessing reduces noise, improves vocabulary consistency, and removes irrelevant identifiers.

**Supervised learning** used the same fully cleaned text as topic modeling to keep the input features consistent across the workflow. The cleaned reviews were converted into numerical representations such as TF-IDF and Bag-of-Words, allowing machine learning models to learn from the language patterns in the text. The **review\_rating** field served as the target variable for training and evaluating the classifiers. This setup helps the models capture the relationship between how patients describe their experiences and the rating they assign.

# **Part V. Sentiment Analysis**

To evaluate the emotional tone of the reviews, we implemented a multi-step sentiment analysis pipeline that operates on progressively cleaned versions of the review text. TextBlob was used to compute polarity (positive–negative orientation). So, reviews with language such as amazing, kind, or best received high positive polarity scores. In contrast, reviews discussing issues such as wait times, rude staff, or billing problems received negative polarity values. Subjectivity (objective–subjective tone); this metric helped distinguish reviews grounded in factual descriptions (procedures, appointment dates, etc) versus reviews expressing personal feelings, judgements, and narratives. while VADER was used to generate a compound sentiment score for each review. These models are appropriate for short, informal text and are commonly used in review analysis. The output provided three different sentiment indicators to be used for hypothesis testing.

This table shows summary results of sentiment analysis for first 5 rows

Table 4: A summary table of Sentiment Analysis Results

A screenshot of a computer

AI-generated content may be incorrect.

**Formulate and test a short hypothesis:**

***- H0 (Null Hypothesis):***There is no relationship between sentiment polarity and review ratings.

***- H1 (Alternative Hypothesis):***Higher polarity (more positive sentiment) is associated with higher review ratings.

**Test the Hypothesis (Using Correlation)**

The Correlation between polarity and review\_rating were used to test the hypothesis. The correlation between polarity and review rating is **0.6768**, which indicates a strong positive relationship. This means that higher textual sentiment (more positive polarity) is generally associated with higher numerical ratings. In other words, when users write more positive reviews, they also tend to give higher star ratings.

The following figure shows the correlation between polarity and review\_raring:

A screenshot of a black and white screen

AI-generated content may be incorrect.

Figure 4: Correlation between Polarity & Review Rating

This strong correlation provides evidence that the sentiment analysis model is capturing meaningful emotional signals from the text. According to Evans (1996), correlations between 0.60 and 0.79 are considered strong, this supports the hypothesis that sentiment aligns with review ratings in this dataset. The consistency between text sentiment and ratings increases confidence in the validity of the sentiment extraction method

# **Part VI: Topic Modeling**

LDA with a bag-of-words representation was chosen because it offers stronger interpretability and more coherent topics. It performs well on short, noisy review text and is able to capture meaningful patterns even when context is limited. LDA also models word co-occurrence more effectively than LSA, making it a better fit for our use case. Overall, it aligns well with our preprocessing pipeline and supports consistent extraction of key themes in patient feedback. For this project, three LDA models were trained with 3, 5, and 7 topics, and each model was saved for later evaluation and comparison. The results show that the model with 3 topics performed the best and achieved the highest coherence score, indicating that it captured the most meaningful and interpretable themes in the review corpus. This finding was also consistent with our manual review, where the top terms were easy to interpret and aligned with how we would naturally group the themes in the patient reviews.

The following table displays the coherence score for each trained model:

Table 5: Coherence Score

|  |  |
| --- | --- |
| **Number of Topics** | **Coherence (c\_v)** |
| 3 topics | 0.4027 |
| 5 topics | 0.3423 |
| 7 topics | 0.3517 |

The three-topic LDA model highlighting the key themes patients focus on when evaluating urgent-care clinics. Topic 1, *"Quality of Care & Professionalism"*, captures how patients describe their overall experience, including the quality of treatment, professionalism of staff, and whether they would recommend the clinic. Topic 2, *“Communication & Staff Interaction Issues,”* reflects concerns specifically related to communication, patients frequently mention unclear explanations, difficulty getting answers, unreturned calls, and other interaction-related frustrations. Topic 3, *“Wait Times & Appointment Delays,”* focuses on operational issues such as long waits, scheduling problems, and delays in being seen.

Together, these three topics provide a structured understanding of patient feedback and directly support our research question by showing that urgent-care experiences are shaped primarily by care quality, communication effectiveness, and service efficiency.

The table below shows the top 10 words for each topic:

Table 6: Top 10 Words for Each Topic

|  |  |
| --- | --- |
| **Topic** | **Top 10 words** |
| Quality of Care & Professionalism | care, doctor, experience, time, great, recommend, medical, take, patient, friendly |
| Communication & Staff Interaction Issues | doctor, go, tell, get, say, call, would, patient, ask, give |
| Wait Times & Operational Delays | wait, time, go, get, appointment, see, urgent\_care, doctor, take, hour |

# **Part VII. Supervised Learning**

Logistic regression models are chosen to determine the outputs whether they are featured as a positive or negative review. It is more straightforward to understand a review with its tendency in a binary classification than multiple classified groups. Therefore, the positive reviews are defined as those which have a 4 or 5 star rating, and the negative reviews are thereby defined as the 1 to 3 star-rating reviews.

TF-IDF and Bag-of-Word are mainly used as the features with or without a truncated SVD. Since then, four cases of supervised models have been determined, including 1. TF-IDF + Logistic Regression, 2. TF-IDF + SVD + Logistic Regression, 3. BoW + Logistic Regression, and 4. Bow + SVD + Logistic Regression.

The snippet below indicates the final performance, where all models have a pretty high  F1-scores. Noteworthy, the TF-IDF outperforms Bag-of-Word in text classification. This is clear evidence that demonstrates IDF weighting and length normalization remains superior to raw term counts. In addition, logistic regression models without truncated SVD perform slightly better than those included truncated SVD. This might be because L2-regularized Logistic Regression already performs powerful implicit dimensionality reduction and feature selection. However, Truncated SVD, by keeping only the top k components, could introduce irreversible information loss, even if a narrow range of total variance. For a linear model like Logistic Regression, it can already exploit the full high-dimensional sparse signal efficiently, thereby this loss typically hurts performance slightly with futile effort.

In conclusion, TF-IDF + Logistic regression is the best model with the highest weighted F1 score among four models.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 5: Supervised Learning Results

# **VIII. Deployment Plan**

To conclude the project, this section describes how the outlined NLP workflow developed in this project could be integrated into a real world application that would support urgent care quality monitoring and decision making, as well as give insight to stakeholders looking to develop new health centers. The goal is to build automated pipelines that monitor patient reviews and gleans interpretable insights for healthcare administrators. The tool would also be used to help develop new healthcare centers and discuss business priorities for peak customer satisfaction.

**1. Target Application Overview**

The envisioned deployment would be a web based application support dashboard for healthcare administrators. The system would:

Collect Google Maps reviews and ratings from participating urgent care centers in desired locations (our project covers California but depending on interest the dataset could be expanded).

Apply the preprocessing pipeline, sentiment analysis, and topic modeling to highlight emerging complaint themes (like staff, wait times, parking, billing), trends in sentiment and average rating over time, and facilities whose predicted sentiment diverges from their star ratings. It would also be able to provide alerts and visualizations to help administrators prioritize improvements. The system is intended as decision support, not meant to replace human judgement.

**2. Automation of Data Collection and Preprocessing**

A predetermined time (weekly, monthly, quarterly, etc) would refresh Google Places/Maps API to retrieve new reviews since last run based on timestamps and refresh lists of urgent care facilities. Newly collected reviews and metadata would be stored with clear separations between raw and processed data. The same text cleaning pipeline outlined in Part III would be implemented and run automatically on each new batch of reviews.

**3. Model Serving and Prediction Workflow**

Once reviews are preprocessed, the system would run trained models on them such as sentiment analysis, topic modeling (LDA), and supervised rating prediction. This would all be compiled into a web dashboard showing sentiment and average rating trends over time, top positive and negative topics for each facility, and outlier facilities. Users would also be able to set up alerts to notify teams when sentiment drops sharply, or volume of negative topics exceeds a predefined threshold. All models and pipelines are containerized and deployed on a given cloud platform to enable scalability and maintenance.

**4. Model Retraining & Updating**

To keep the system aligned with evolving language and expectations, there would also be a model lifecycle management plan. Every 6 months the model would be retrained on recent data using time based splits and comparing new models against current models on test sets. These models would be monitored based on distribution of star ratings over time and distribution of textual features. A model registry would be used to track versions, training periods, and evaluation metrics. New models would be deployed but only be used to score data, not impact dashboards, and only be implemented to dashboards after it was determined the data produced was valid.

**5. Monitoring Performance & Ethics Over Time**

Evaluation metrics would be continuously updated on recent reviews and can be tracked by facility and region. The dataset does not contain demographic information about reviewers or patients and all reviews are public. Fairness considerations are therefore based on institutional patterns rather than individual attributes. This can be done by monitoring error rates across regions or facility types (including chains, like Kaiser in California). Stakeholders would also keep in mind the inherent bias in review samples, since there is a sample bias in the customers who leave feedback (negative and positive) compared to ones who do not. If fairness issues are detected, retraining models by region and providing confidence intervals could be helpful.

In preprocessing names of specific staff are removed and the dashboard presents aggregated results at facility level at the lowest grain, never individual level reviews for patent staff confidentiality. Raw text can only be accessed by high level security clearances. Participating clinics should be informing all their staff that an NLP based monitoring tool is used and know what insights would be generated. Documentation should be available if requested and describe data sources, preprocessing steps, model types, known limitations, and fairness/bias monitoring procedures. It should also be stressed this is a tool meant to augment, not replace, human judgement and a human in the loop review process ensures critical decisions are made by humans, not the algorithmic outputs.

**IX. References**

Evans, D. A., Ginther-Webster, K., Hart, M., Lefferts, R. G., & Monarch, I. (1991). *Automatic indexing using selective NLP and first-order thesauri.* In *Proceedings of the Intelligent Multimedia Information Retrieval Systems & Management (RIAO ’91)* (pp. 624–643). Elsevier.