NLP
Final Project Proposal
October 29, 2025
Team Members: Anvita Suresh (as1693), Arushi Singh (as1685), Vihaan Manchanda (vm180)

Business Case: Medical abbreviations are a leading cause of medication errors and miscommunication in healthcare. Many doctors use shorthand abbreviations in their clinical notes, which can be misinterpreted by other healthcare providers, pharmacists, and automated clinical decision support systems, leading to incorrect treatments and adverse patient outcomes. This project addresses the problem of disambiguating commonly confused medical abbreviations, such as "MS" (Multiple Sclerosis vs. Mitral Stenosis vs. Mental Status vs. Morphine Sulfate), using contextual information from clinical notes.

Implementation: We will implement a classification-based approach that uses n-gram context features from surrounding words to predict the correct expansion for each ambiguous abbreviation. The model will be trained and evaluated on both synthetic examples generated from expansion-specific templates and real clinical notes from the MIMIC-III database. This applies multi-class classification and n-gram modeling to a novel healthcare problem where disambiguation has direct patient safety implications. This is a Word Sense Disambiguation (WSD) problem.

Solution Outline:

1. Data Collection & Preparation

- Primary source: MIMIC-III clinical discharge summaries and progress notes
- Target abbreviations: Focus on 3-4 highly ambiguous abbreviations (MS, PT, CA)
- Gold standard creation: Semi-automatic labeling by identifying sentences where expansions appear near abbreviations (e.g., "multiple sclerosis (MS)"), supplemented with manual annotation of ambiguous cases
- Dataset size: ~100-200 labeled examples per abbreviation meaning

2. Feature Extraction

- Context window: Extract words within ±5 positions of the target abbreviation
- Feature representation: Bag-of-words model using word counts from context window
- Additional features: Bigrams and trigrams to capture phrasal patterns

3. Model Development

- Baseline: Most-frequent-sense classifier (always predicts the most common expansion)
- Primary model: Multinomial Naive Bayes classifier with n-gram features
- Alternative model (if time permits): Logistic Regression for comparison
- Separate classifier trained for each abbreviation

4. Evaluation Strategy

- **Synthetic data (Step 3a):** Generate 200-500 template-based examples per expansion that align with model assumptions. Evaluate accuracy, per-class precision/recall, and confusion matrices. Expected performance: >90% accuracy.
- **Real data (Step 3b):** Train on 70% of labeled clinical notes, test on 30%. Evaluate using same metrics. Provide qualitative analysis with 10-15 success cases and 10-15 failure cases, explaining why the model succeeds or fails based on context patterns.

5. Expected Outcomes

- Demonstrate that simple contextual features can effectively disambiguate medical abbreviations
- Identify which context words are most discriminative for each expansion
- Analyze limitations (insufficient context, rare expansions, domain-specific usage)
- Discuss practical deployment considerations for clinical decision support systems