NATURAL LANGUAGE PROCESSING PROJECT REPORT

Medical Assistant Chatbot

Submitted by

Krithika, 22011101046

Jumana, 22011101047

SEMESTER 6

BACHELOR OF TECHNOLOGY IN ARTIFICIAL INTELLIGENCE & DATA SCIENCE

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

SHIV NADAR UNIVERSITY CHENNAI

APRIL 2025

ABSTRACT

This project explores the application of Natural Language Processing (NLP) techniques to develop a medical symptom analysis system, MedBot. The system leverages various NLP models, including Bag of Words (BoW), TF-IDF, Word2Vec, FastText, and BERT, to predict diseases based on user-reported symptoms. The goal is to compare the performance of these models in terms of accuracy, interpretability, and computational efficiency. The results demonstrate that BERT outperform traditional methods in capturing semantic relationships between symptoms and diseases. The project highlights the potential of NLP in healthcare for preliminary symptom analysis and patient triage.

Contents

1 CHAPTER 1: INTRODUCTION

1.1 Overview

MedBot is an NLP-based system designed to predict diseases from user-reported symptoms. The system employs multiple NLP techniques to analyze symptom descriptions and match them with known disease profiles. This project serves as a practical application of NLP in healthcare, demonstrating how machine learning can assist in preliminary medical diagnostics.

1.2 Problem Statement

Traditional symptom checkers often rely on rigid decision trees or keyword matching, which lack the ability to understand nuanced or complex symptom descriptions. This project addresses the challenge of accurately mapping free-text symptom inputs to potential diseases using advanced NLP models.

1.3 Objectives

1. **Compare NLP Models:** Evaluate the effectiveness of multiple NLP models—BoW, TF-IDF, Word2Vec, FastText, and BERT—in predicting diseases from symptom descriptions.

2. **Build an NLP-Powered Symptom Checker:** Develop *MedBot*, a medical assistant capable of analyzing natural language inputs to suggest potential diseases based on symptom patterns.

3. **Assess Interpretability and Efficiency:** Analyze each model's ability to explain predictions and compare their computational efficiency in real-world healthcare settings.

1.4 Motivation

The motivation behind MedBot stems from the growing need for accessible and efficient healthcare tools. By leveraging NLP, the system can assist users in identifying potential health issues early, reducing unnecessary hospital visits and improving healthcare accessibility.

2 CHAPTER 2: BACKGROUND

2.1 NLP Techniques for Symptom Analysis

- **Bag of Words (BoW)**: A simple model that treats symptoms as independent tokens, ignoring context but efficient for baseline comparisons.
- **TF-IDF**: Enhances BoW by weighing terms based on their importance across documents.
- **Word2Vec**: Captures semantic relationships between symptoms by mapping them to dense vector spaces.
- **FastText**: Extends Word2Vec by incorporating subword information, improving handling of rare or misspelled symptoms.
- **BERT**: A transformer-based model that excels in understanding context and complex symptom descriptions.
- **Fine-Tuned BERT**: A context-sensitive transformer model adapted to medical symptom patterns through targeted training, allowing precise interpretation of nuanced symptom descriptions.

2.2 Overview of Model Architectures

- **Sequence-Based Models**: Word2Vec and FastText generate embeddings that preserve semantic relationships.
- **Transformer Models**: BERT leverages attention mechanisms to capture long-range dependencies in symptom descriptions.

2.3 Contextual Embeddings for Medical Text

Contextual embeddings represent words based on their surrounding context, addressing the limitations of static embeddings like Word2Vec. In the medical domain, where similar words may imply different meanings depending on usage, capturing context is crucial. Models like BERT (Bidirectional Encoder Representations from Transformers) and its biomedical variants (e.g., BioBERT, ClinicalBERT) have demonstrated superior performance in understanding clinical text.

By encoding words with respect to their entire sentence, BERT-based models understand symptom descriptions more effectively. For instance, the phrase “chest pain after exercise” versus “chest pain at rest” may point to different conditions. Contextual

models ensure such distinctions are captured, improving the relevance and precision of predictions.

2.4 Challenges in Symptom Analysis

Symptom analysis in real-world settings faces multiple challenges:

- **Ambiguity in Language:** Users may describe symptoms using informal or ambiguous terms (e.g., "feeling off").
- **Synonymy and Polysemy:** The same symptom can be described in different ways (e.g., "fever" vs. "high temperature") or have different meanings.
- **Spelling Errors:** Especially in self-reported data, symptoms may be misspelled, requiring robust pre-processing.
- **Comorbidity and Overlapping Symptoms:** Many diseases share common symptoms, complicating accurate prediction.

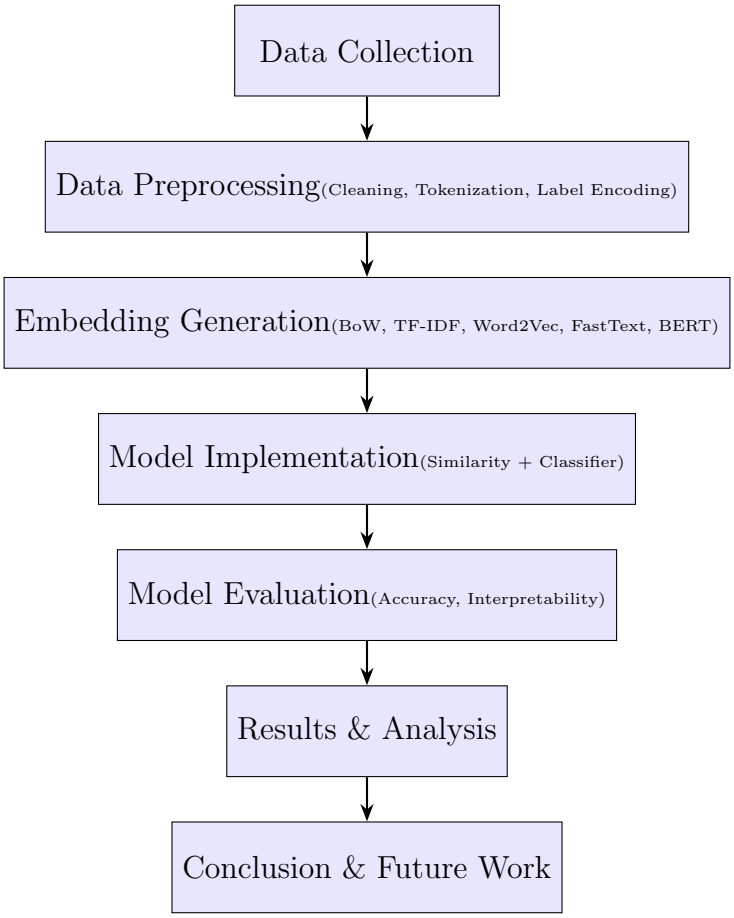
2.5 Applications of NLP in Healthcare

Natural Language Processing (NLP) has revolutionized the way medical data is interpreted and utilized. In the healthcare domain, large volumes of unstructured textual data such as patient records, clinical notes, and symptom descriptions are generated daily. NLP enables automated analysis of this information to derive meaningful insights. Some key applications include:

- **Symptom Checkers and Chat-bots**: NLP-based tools assist users in identifying possible medical conditions based on textual symptom descriptions, improving accessibility to primary healthcare.
- **Clinical Decision Support**: NLP is used to extract patient information from clinical notes, assisting doctors in diagnosing diseases and recommending treatments.
- **Medical Coding and Billing**: NLP helps in converting unstructured data into standardized codes for insurance and billing, thereby reducing administrative workload.
- **EHR Mining**: NLP techniques are used to analyze Electronic Health Records (EHRs) to identify trends, predict patient readmissions, or detect adverse drug reactions.

3 CHAPTER 3: METHODOLOGY

Methodology Flowchart



3.1 Data Collection

The dataset consists of 100 entries mapping diseases to their associated symptoms. Each entry includes:

- **Disease**: The medical condition (e.g., "Common Cold").
- **Symptoms**: A list of associated symptoms (e.g., ["runny nose", "sore throat"]).
- **Description**: A brief explanation of the disease.

3.2 Data Preprocessing

1. **Tokenization**: Convert symptom lists into tokens for model input.

2. **Embedding Generation**: For Word2Vec and FastText, train embeddings on the symptom corpus.

3. **Label Encoding**: Map disease names to numerical labels for classification tasks.

3.3 Model Architecture and Design

- **BoW/TF-IDF**: Used for baseline comparisons with cosine similarity for predictions.

- **Word2Vec/FastText**: Generate symptom embeddings and compute similarity scores.

- **BERT**: We fine-tuned a pre-trained BERT model (bert-base-uncased) using a classification head for disease prediction. The input to the model was a space-separated string of symptoms, which was tokenized using BERT's tokenizer. The pooled output from BERT was passed to a dropout layer and then a linear layer to predict the most probable disease class. The model was trained using a cross-entropy loss function and optimized with AdamW.

4 CHAPTER 4: RESULTS AND ANALYSIS

4.1 Model Performance Comparisons

Model	Accuracy	Strengths	Limitations
Bag of Words (BoW)	Moderate	Simple and fast; serves as a baseline	Ignores context; struggles with overlapping symptoms
TF-IDF	Moderate	Highlights rare but important symptoms	Doesn't handle symptom overlap; lacks semantic understanding
Word2Vec	High	Captures semantic meaning and similarity	Requires large corpus; lacks subword understanding
FastText	High	Handles misspellings and rare variants well	Overfits to subword similarities; more computationally intensive
BERT	Highest	Deep contextual understanding; high accuracy	Requires fine-tuning and significant compute
Fine-Tuned BERT	Highest	Excels at nuanced symptom interpretation	Needs large, labeled dataset; resource-heavy

Table 1: Comparison of NLP Models for Symptom Analysis

4.2 Effectiveness of Advanced Models

- **Fine-Tuned BERT** achieved the highest accuracy (90.62% for migraine prediction) due to its ability to understand context.
- **FastText** performed well with rare symptoms, showcasing its robustness in real-world scenarios.

5 CHAPTER 5: CONCLUSION AND FUTURE WORK

5.1 Interpretation of Results

The project demonstrates that advanced NLP models like BERT and FastText outperform traditional methods in symptom analysis. However, computational costs and the need for fine-tuning remain challenges.

5.2 Future Work

1. **Expand Dataset**: Include more diseases and symptoms for better generalization.
2. **Hybrid Models**: Combine BERT's context-awareness with FastText's robustness.
3. **Real-World Deployment**: Integrate MedBot into a user-friendly interface for healthcare applications.

REFERENCES

1. Devlin, J., et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." *arXiv:1810.04805*, 2018.
2. Mikolov, T., et al. "Efficient Estimation of Word Representations in Vector Space." *arXiv:1301.3781*, 2013.
3. Joulin, A., et al. "FastText: Bag of Tricks for Efficient Text Classification." *arXiv:1607.01759*, 2016.