Project Report

Computational Linguistics for Indian Languages
(CS689A)

Project Title: AI-Driven Medical Diagnosis and Drug Suggestion

Submitted By:
Sanjeev Kumar (231110044)
Govind Sharma (231110015)
Prashik Ganer (231110037)

Course Instructor:
Prof. Arnab Bhattacharya
CSE Department
IIT Kanpur

Problem Statement:

We aim to develop a model that predicts the medical condition based on patient feedback. After predicting the condition, the model will suggest the top 3 medicines from the dataset for that specific condition.

Our Solution:

We have trained and evaluated multi-class classification models to predict medical conditions from patient feedback. The best-performing model will be selected for further use.

Dataset Description:

The dataset, sourced from the UC Irvine Machine Learning Repository, contains patient reviews on specific drugs, medical conditions, and a 10-star satisfaction rating. It is a tab-separated file with six fields:

1. **drugName**: Name of the reviewed drug.
2. **condition**: The medical condition of the reviewer.
3. **review**: Detailed feedback on the drug's usage, including positive and negative effects.
4. **rating**: Patient's satisfaction rating on a 1-10 scale.
5. **date**: Date of the review.
6. **usefulCount**: Number of people who found the review helpful.

Pipeline:

Our workflow is divided into two categories: one uses raw feedback, and the other uses pre-processed feedback for training and testing models.

To prepare the text data for machine learning, we converted it into vectors using Bag of Words (BoW) and TF-IDF methods. We also extended our evaluation by considering both uni-grams and bi-grams in the vectorization process.

Data Pre-Processing:

To evaluate the impact of removing insignificant words, we cleaned the feedbacks before evaluating the classification models. The cleaning process included:

1. **Punctuation Removal**: Removed punctuation marks using regular expressions.
2. **Special Characters/Numbers Removal**: Removed non-alphabetic content.
3. **Lowercasing**: Converted all text to lowercase for consistency.
4. **Stopword Removal**: Removed common English stopwords.
5. **Lemmatization**: Reduced different word inflections to their root forms.

Vectorizers:

1. CountVectorizer
2. TFIDFVectorizer

Classification Models:

We have used below models from scikit learn library:

1. Decision Tree
2. Naïve Bayes

3. Passive Aggressive
4. Support Vector Machine
5. Random Forest

Accuracy of Different Models

Models	Cleaned				Uncleaned			
	BoW		TF-IDF		BoW		TF-IDF	
	uni-gram	bi-gram	uni-gram	bi-gram	uni-gram	bi-gram	uni-gram	bi-gram
Decision Tree	0.825	0.831	0.822	0.822	0.997	0.997	0.997	0.997
Naïve Baye's	0.756	0.663	0.533	0.352	0.728	0.757	0.532	0.357
Passive Aggressive	0.805	0.879	0.843	0.898	0.928	0.996	0.972	0.997
SVM	0.816	0.827	0.866	0.876	0.873	0.857	0.883	0.896
Random Forest	0.875	0.866	0.875	0.862	0.902	0.923	0.917	0.904

Result Analysis

After evaluating different models on 116,963 training and 39,117 test instances, the Decision Tree model on uncleaned data proved to be the most suitable due to its highest accuracy and simplicity. This model will be used in the Web App to predict medical conditions and suggest the top 3 corresponding medicines.