Natural Language Processing - Assignment

# Q.No.1: Customer Support Chatbot for E-commerce

## 1. Syntax

- Sentence: 'I want to return my order.'

- Syntactic Structure:

- Subject (I)

- Verb phrase (want to return)

- Object (my order)

- Problem: Complex subordinate clauses make intent detection difficult.

## 2. Semantics

- Semantic Error: Misunderstanding 'return' as financial instead of product return.

- Solution: Use contextual embeddings (BERT) and fine-tune for e-commerce intent detection.

## 3. Pragmatics

- Pragmatic Error: Asking for order number before acknowledging refund.

- Improvement: Acknowledge user request first, then ask for details.

## 4. Discourse

- Failure Impact: Breaks conversation flow.

- Solution: Maintain context tracking in multi-turn conversations.

# Q.No.2: NLP System for Medical Diagnosis

## 1. NLP Evolution Over the Past Decade

- Past: Rule-based.

- Now: Transformer models (e.g., BioBERT).

- Impact: Better understanding and scalability.

## 2. Future Directions

- Domain-specific pretraining.

- Knowledge graph integration.

- Explainable AI (XAI).

- Few-shot learning.

## 3. Challenges and Solutions (Next 3–5 Years)

- Challenges: Jargon complexity, data privacy.

- Solutions: Synthetic data, federated learning, cross-disciplinary efforts.

# Q.No.3: Sentiment Analysis on Social Media

## 1. Preprocessing Steps

- Tokenization, noise removal (hashtags, mentions, URLs), emoji handling.

## 2. SpaCy Code Snippet

import spacy

nlp = spacy.load("en\_core\_web\_sm")

def preprocess(text):

doc = nlp(text)

tokens = [token.text.lower() for token in doc if not token.is\_stop and token.is\_alpha]

return tokens

sample\_text = "I love this! 😍 It's awesome @user #happy"

print(preprocess(sample\_text))

## Why These Steps Matter

- Lowercase: Dimensionality reduction.

- Stopword removal: Focus on important terms.

- Tokenization: Prepare data for model.

# Q.No.4: News Recommendation System

## 1. Why Remove Stopwords

- Reduces noise; improves keyword matching.

## 2. Pros and Cons of Removing Punctuation

- Pros: Cleaner text.

- Cons: Loss of sentiment cues.

## 3. Handling Negation

- Preserve negations ('not', 'never') and use bigrams ('not\_good').

# Q.No.5: Email Categorizer (Spam/Ham Classification)

## 1. Preprocessing Steps

- Correct misspellings, remove noise, handle concatenated words.

## 2. Stemming and Lemmatization

- Reduces sparsity, improves model performance.

## 3. Word Embeddings

- Use embeddings to capture relationships (e.g., 'scam' and 'fraud').