NATURAL LANGUAGE PROCESSING (NLP) – DETAILED NOTES

1. INTRODUCTION TO NLP

Natural Language Processing (NLP) is a branch of AI that enables computers to understand, interpret, and generate human language.

Applications include chatbots, translation, sentiment analysis, search engines, and spam detection.

2. LEVELS OF NLP

Phonology – Study of sounds in language.

Morphology – Word structure (root, prefix, suffix).

Syntax – Grammar and sentence structure.

Semantics – Literal meaning of words/sentences.

Pragmatics – Meaning based on context.

Discourse – Understanding text across multiple sentences.

3. CORE NLP TASKS

• Tokenization – Breaking text into words/sentences.

• Stopword Removal – Removing common words like “the”, “is”.

• Stemming – Reducing words to root form (play, playing $\rightarrow$ play).

• Lemmatization – Converting words to dictionary form (better $\rightarrow$ good).

• POS Tagging – Identifying noun, verb, adjective.

• Named Entity Recognition – Identifying person, place, organization.

• Sentiment Analysis – Positive, negative, neutral.

• Language Modelling – Predicting next word.

4. APPROACHES TO NLP

Rule-Based NLP – Manual rules and grammars.

Statistical NLP – Machine learning models: Naive Bayes, SVM.

Deep Learning – RNN, LSTM, GRU.

Transformers – Modern models using self-attention (BERT, GPT).

5. WORD REPRESENTATION TECHNIQUES

One-hot Encoding – Sparse binary vectors with no semantic meaning.

Word Embeddings – Dense vectors with meaning (Word2Vec, GloVe).

Contextual Embeddings – Meaning changes with context (BERT, GPT).

6. EVALUATION METRICS

Accuracy – Overall correctness.

Precision, Recall, F1 – Used for classification evaluation.

BLEU – Machine translation evaluation.

ROUGE – Text summarization evaluation.

7. CHALLENGES IN NLP

Ambiguity – Multiple meanings of words.

Sarcasm – Hard for machines to detect.

Code-mixed Language – Example: Hinglish.

Low-resource Languages – Limited datasets.

Bias – Training data bias leads to biased models.

8. MODERN TRENDS IN NLP

Foundation Models – Large-scale language models (GPT, Llama).

Multimodal Models – Text + image + audio.

Prompt Engineering – Designing inputs for AI.

RAG (Retrieval-Augmented Generation) – Search + generation.

LoRA – Lightweight fine-tuning.

SUMMARY

NLP enables computers to process human language using multiple linguistic levels, machine learning, deep learning, and transformer-based models.