Artificial Intelligence NLP

---

1. Introduction to NLP

Natural Language Processing (NLP) is the field of Artificial Intelligence that focuses on enabling computers to **read, understand, interpret, and generate human language** (text or speech).

Humans communicate using natural language (English, Hindi, etc.), but computers understand only numbers.
NLP acts as a **bridge** between human language and machine understanding.

---

2. What is NLP?

Definition:
NLP is a subfield of AI that helps machines understand, analyze, and generate human language in a meaningful way.

In simple terms:
NLP = **Computer understanding + Human language + Machine learning / Deep learning algorithms**

Example:

- Siri/Alexa understanding your speech
- Google showing autocomplete suggestions
- ChatGPT generating answers
- Email apps detecting spam

---

3. Need of NLP

Why do we need NLP?

a. Humans communicate in natural language

- Machines can't understand Hindi/English directly

- NLP converts natural language → structured data → machine understanding

b. To extract useful information from text

- Social media posts
- Customer reviews
- Emails, messages, documents

c. Automation of text-heavy tasks

- Summarizing content
- Translating languages
- Auto-correct & auto-suggestions

d. Unlock value from unstructured data

Nearly **80% of global data is unstructured text—**NLP helps analyze it.

---

4. Real-World Applications of NLP

1 Chatbots & Virtual Assistants

- ChatGPT
- WhatsApp chatbots
- AI support agents

2 Machine Translation

- Google Translate
- Amazon Translate

3 Sentiment Analysis

- Brand monitoring

- Customer feedback classification
- Product review analysis

4 Speech Recognition

- Siri, Alexa, Google Assistant
- YouTube subtitles

5 Text Summarization

- News summarizers
- Research paper summarization tools

6 Spam Detection

- Email spam filters
- Fraud detection using text patterns

7 Search Engines

- Google search
- E-commerce product search
- Semantic search (meaning-based search)

8 Document Processing

- OCR + NLP for automated invoice, ID, KYC processing

---

5. Common NLP Tasks

1. Tokenization

Splitting text into words, sentences, or subwords.
Example:
“ChatGPT is awesome” → [ChatGPT, is, awesome]

2. Lemmatization/Stemming

Converting words to their root form.

- running → run
- better → good (lemma)

3. POS Tagging

Identifying noun/verb/adjective etc.

4. Named Entity Recognition (NER)

Detecting entities like:

- Person → “Narendra Modi”
- Location → “Delhi”
- Organization → “Google”

5. Text Classification

- Spam vs not spam
- Positive vs negative sentiment
- Support ticket categorization

6. Machine Translation

English → Hindi, Hindi → English

7. Text Generation

- ChatGPT
- Story generators
- Autocomplete

8. Question Answering

- Chatbots

- Search assistants
- Medical QA systems

9. Summarization

Extractive / abstractive summarization.

---

6. Approaches Used For NLP

There are three major approaches:

A. Rule-Based NLP (Old Generation)

Uses hand-written rules.
Example:
If the sentence contains “not good” → Negative sentiment.

Pros:

- Transparent
- Good for simple tasks
  Cons:
- Not scalable
- Hard to maintain
- Fails on complex cases

B. Traditional ML-Based NLP

Uses machine learning + manual features.

Algorithms:

- Naïve Bayes
- SVM

- Logistic Regression
- CRF
- Hidden Markov Models

Manual Features:

- Bag of Words
- TF-IDF
- N-grams

Pros:

- Works well on small data
- Interpretable
  Cons:
- Heavy feature engineering
- Struggles with context

C. Deep Learning-Based NLP (Modern NLP)

Uses neural networks to automatically learn features.

Models:

- RNN
- LSTM
- GRU
- CNN for text
- Transformer (latest and best)

Why are Transformers the best?
Because they understand long-range context using **attention mechanisms**.

D. Large Language Models (LLMs) – Current State-of-the-Art

Examples:

- GPT-4
- GPT-5
- LLaMA
- Gemini
- Claude
- Mistral

Uses massive datasets + transformer architecture.

---

7. Challenges in NLP

1. Ambiguity

Words with multiple meanings:

- "Apple" = fruit / company
- "Bank" = river bank / money bank

2. Sarcasm Detection

"I love working 12 hours a day 😒"

3. Context Understanding

Pronouns:

- "Ravi met Vijay. He was angry."
  Who is "he"?

4. Low-Resource Languages

Hindi, Bengali, Marathi—less data compared to English.

5. Slang & Informal Text

- “u” instead of “you”
- Emojis
- Social media shorthand

6. Data Privacy

Chat & conversation data must be handled securely.

7. Bias in Models

Models may learn biased patterns from the training data.

---

8. Assignment (46:25)

Assignment: Build a Mini NLP Pipeline (Hands-on)
You must complete the following tasks using **Python + any NLP library (NLTK, spaCy, or Transformers)**.

Part 1: Preprocessing

Take any paragraph (5–7 sentences) and perform:

1. Tokenization
2. Stopword removal
3. Lemmatization

Output must be shown step-by-step.

Part 2: Text Classification (Simple)

Train a small model on sample data:

Text	Label
“I love this product”	Positive

“Worst experience ever”	Negative
“Not bad”	Neutral
“Very happy with the service”	Positive

Use **Naive Bayes** or **Logistic Regression**.

Predict sentiment for:

- “The product quality is amazing”
- “I hate the delay”

Part 3: Build a Simple NER

Use a spaCy small model.
Extract entities from:
“Google CEO Sundar Pichai visited India on Monday.”

Output should show PERSON, ORG, LOC, DATE.

Part 4: Short Answer

Explain 4 lines each:

1. What is tokenization?
2. Difference between stemming & lemmatization
3. Why are LLMs better than older NLP models?
4. What is word embedding?