Introduction to Natural Language Processing (NLP)

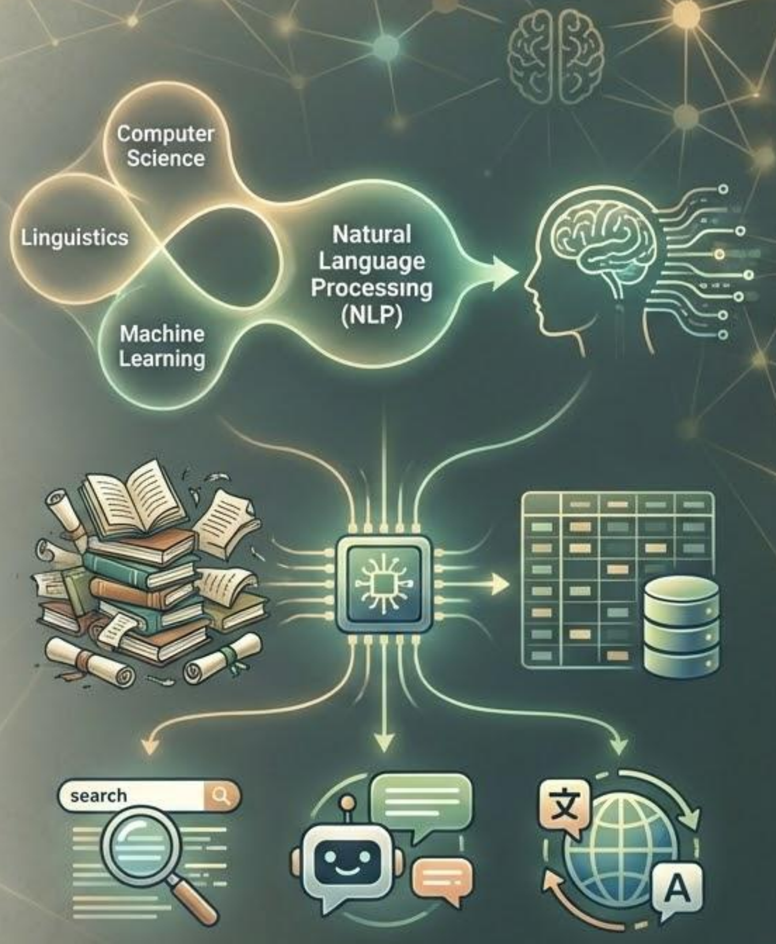
Dr. Abhishek Grover
Assistant Professor
Communication and Computer Engineering, LNMIIT

References

- Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models, Daniel Jurafsky and James H. Martin.
- Natural Language Processing with Python, Steven Bird, Ewan Klein and Edward Looper.
- Neural Network Methods for Natural Language Processing, Yoav Goldberg.
- Natural Language Processing with Transformers, Lewis Tunstall, Leandro von Werra, and Thomas Wolf.
- Web Links

Definition

- Natural Language Processing (NLP) is a field of AI that enables computers to understand, interpret, and generate human language
- It lies at the intersection of computer science, linguistics, and machine learning
- Core objective: convert unstructured language data into a machine-understandable form
- Widely used in real-world systems like search engines, chatbots, and translators



Core Components

● **Lexical Analysis**	Processing at word level (tokenization, normalization)
● **Syntactic Analysis**	Analyzing grammatical structure using parsing techniques
● **Semantic Analysis**	Extracting meaning from words and sentences
● **Discourse Analysis**	Understanding relationships across sentences
● **Pragmatic Analysis**	Interpreting meaning based on context and intent
● **Language Modeling**	Learning probability distributions over word sequences

NLP Tasks

Text Preprocessing
Tokenization, stemming, lemmatization, stop-word removal

Text Classification
Sentiment Analysis, Spam detection, topic classification

Sequence Labelling
Part-of-speech tagging, named entity recognition

Information Extraction
Entity-relation extraction

Machine Translation
Translation between languages

Text Generation
Summarization, chatbot

NLP Tasks

EASY

- Spell checking, keyword-based Information retrieval and topic modeling

MEDIUM

- Text classification,
- Information Extraction, Text summarization

HARD

- Question-answering,
- Machine Translation,
- Conversational agent

Evolution of NLP

Rule-based NLP (1950s-1980s)

Handcrafted linguistic rules

Statistical NLP (late 1980s-1990s)

Probabilistic models such as n-grams and HMMs

Classical Machine Learning (1990s-2010)

Feature-Engineered Models (SVM, Naive Bayes)

Word Embedding (2013-2016)

Dense semantic representation (Word2Vec, GloVe)

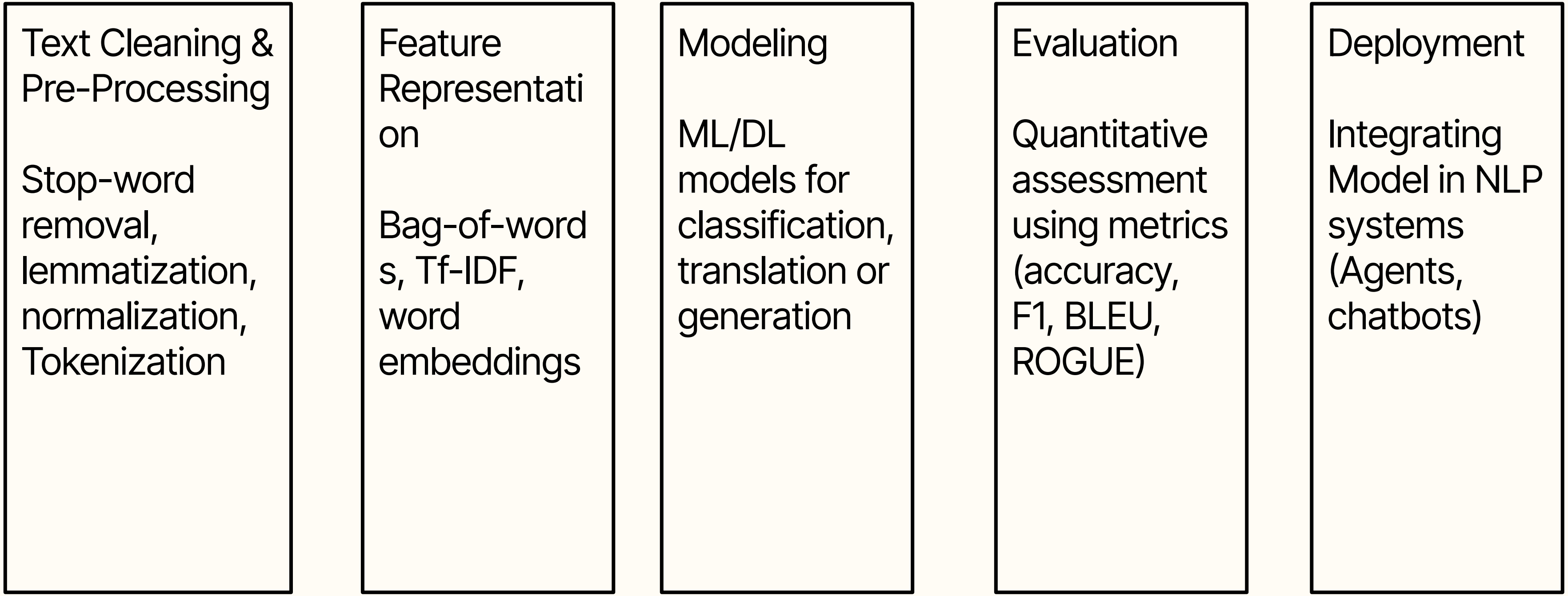
Deep Learning (2014-2018)

RNNs, LSTMs, CNNs for sequence modeling

Transformer (2018-present)

Attention-based models like BERT, GPT

NLP Pipeline



Text Cleaning & Pre-Processing

Stop-word removal, lemmatization, normalization, Tokenization

Feature Representation

Bag-of-words, Tf-IDF, word embeddings

Modeling

ML/DL models for classification, translation or generation

Evaluation

Quantitative assessment using metrics (accuracy, F1, BLEU, ROGUE)

Deployment

Integrating Model in NLP systems (Agents, chatbots)

Text Pre-processing

- Removal of noise such as HTML tags, URLs, emojis, and special characters
- Stop-word removal to reduce non-informative words
- Normalization: lowercasing, handling contractions, spelling correction
- Stemming or lemmatization to reduce words to base form
- Tokenization: splitting text into words or subwords
- Handling missing, ambiguous, or out-of-vocabulary tokens

Feature Representation

- Converts text into numerical form suitable for machine learning models

Count-Based Methods	**Frequency-based Methods**	**Distributed Representations**	**Contextual Embeddings**
Bag-of-words	TF-IDF	Word2Vec, GloVe	BERT, GPT

- Choice of representation impacts model performance and complexity

Modeling

- Converts text into numerical form suitable for machine learning models

Classical ML Models	Sequence Models	Deep Learning Models	Transformer Models
Naive Bayes, Logistic Regression	HMM, RNN, LSTM	CNN, attention-based networks	BERT, GPT

- Choice of representation impacts model performance and complexity

Evaluation

- Various metrics are used to measure model performance
- The type of metrics used depends on application
- Classification metrics: accuracy, precision, recall, F1-score
- Generation metrics: BLEU (for language translation), ROGUE (for text generation)
- Usually validation and test datasets are used for evaluation
- It helps in analyzing weakness of models and the developer may go back to previous stage for performance enhancement.

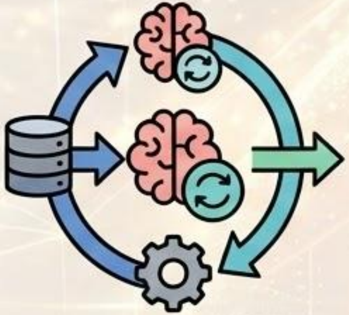
Deployment

Integrating Trained Models

Into real-world applications

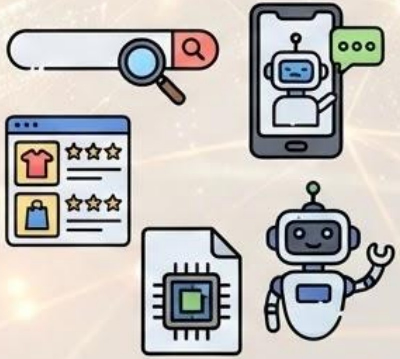
Exposing via API / Embedding

Models in software systems



Periodic Retraining & Updates

With new data



Real-world Applications

Search-Engines, Virtual assistants, recommendation systems, document-processing systems, Agents