Article Link: https://www.dhiwise.com/post/natural-language-processing-techniques-for-developers

What is NLP?

1. NLP (Natural Language Processing) interlinks human language and machines, enabling applications like chatbots, sentiment analysis, and search enhancements.
2. It integrates linguistics, machine learning, and AI to process, understand, and generate natural language effectively.

1. Text Preprocessing:

1. Tokenization: Splits text into meaningful units like words, sub-words, or characters.
2. Normalization: Converts text to uniform format (lowercasing, removing punctuation, special characters).
3. Stopword Removal: Eliminates common words like “and”, “the”, “is” to reduce noise while retaining meaning.
4. Stemming & Lemmatization: Reduces words to their root form (“running” becomes “run”)

2. Classical Text Representation: BoW & TF-IDF

1. Bag-of-Words (BoW): Represents documents as word-count vectors, ignoring order and context.
2. TF-IDF: Upgrades BoW by weighting words by importance across documents and downweights common terms like “the”.
3. While effective for classification tasks, these methods don’t retain sematic meaning.

3. Semantic Learning: Word Embeddings

1. Embeddings capture meaning by mapping words into dense vectors where similar words are close together.
2. Techniques:
   1. Word2Vec: Neural embedding using context-based prediction.
   2. GloVe: Pretrained embeddings using global co-occurrence statistics.
   3. fastText: Extends embedding to handle rare words via character n-grams.
3. Modern models (BERT, GPT) provide contextual embeddings that adjust depending on sentence context.

4. Transformer Models:

1. Transformer uses self-attention mechanisms to process text in parallel, understanding long-range dependencies.
2. Two main types:
   1. Encoders (BERT): Optimal for understanding tasks like classification or Q&A.
   2. Decoders (GPT): Designed for text generation.
3. These models can be pretrained on massive data and fine-tuned for specific tasks via frameworks like Hugging Face.

5. Real-World Use Cases

1. Chatbots / Virtual Assistants: User intent interpretation and conversational responses.
2. Text Classification: Spam detection, sentiment analysis, and topic categorization.
3. Information Extraction
4. Semantic Search
5. Voice Assistants & Social Monitoring: Speech-to-text pipelines and sentiment tracking on social media.