Natural Language Processing (NLP)

BROWN COLOUR : It is used to define the terminologies in NLP .
YELLOW MARK : It is used to **Highlight the words .**
DEEP BLUE : It is used to name the topic .
RED COLOUR : It is used for QUESTIONS .

What is NLP?

Natural Language Processing (NLP) is a field of **Artificial Intelligence (AI)** that focuses on the interaction between computers and humans through natural language. The ultimate goal of NLP is to enable computers or Machines to **understand, interpret, generate, and manipulate human language** in a valuable way.

💡 **Think of NLP as a bridge between human language and computers! .**

NLP helps developers to organize knowledge for performing tasks such as translation, automatic summarization, Named Entity Recognition (NER), speech recognition, relationship extraction, and topic segmentation.

📌 NLP Terminologies To Learn :

✅**Corpus:** A collection of text data used for training NLP models. (paragraph)
✅**Document:** When we have Any Kind of Sentences.
✅**Vocabulary :** All The Unique words present in the corpus .
✅**Tokens:** Individual words, sentences, or subwords extracted from text.
✅**Tokenization:** The process of splitting text into smaller units (words or sentences).
✅**Stopwords:** Common words like "the", "is", "and" that are removed to reduce noise.
✅**Unique Words (Vocabulary):** The set of distinct words present in a corpus.

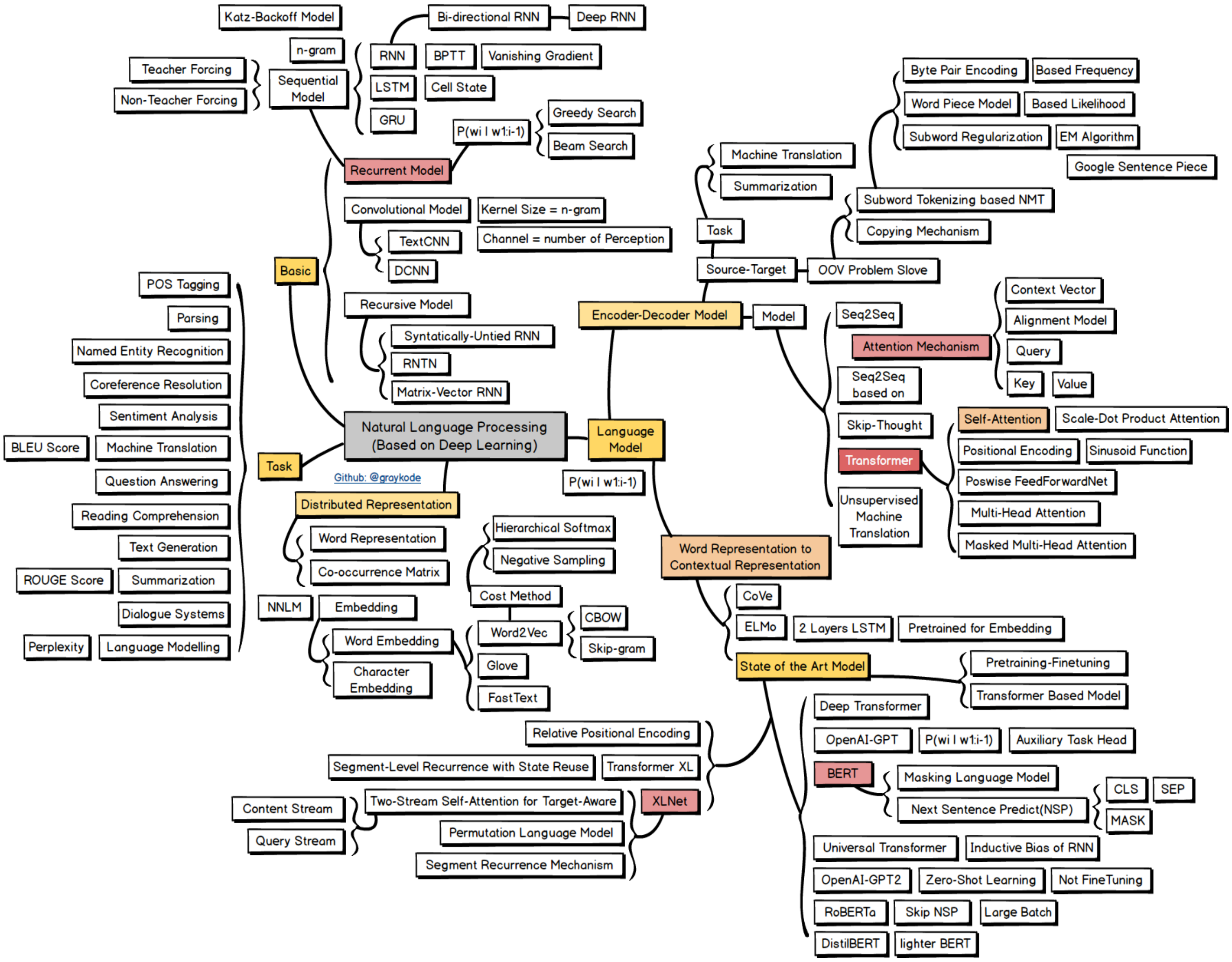
✅**Part-of-Speech (POS) Tagging:** Labeling words as nouns, verbs, adjectives, etc.
✅ **Named Entity Recognition (NER):** Identifying entities like names, dates, locations in text

✅ **Text Classification:** Categorizing text into predefined labels (e.g., spam detection).
✅ **Sentiment Analysis:** Determining whether text is positive, negative, or neutral.
✅**Dialects :** A particular Form of Language which is peculiar to a specific region or Social Group

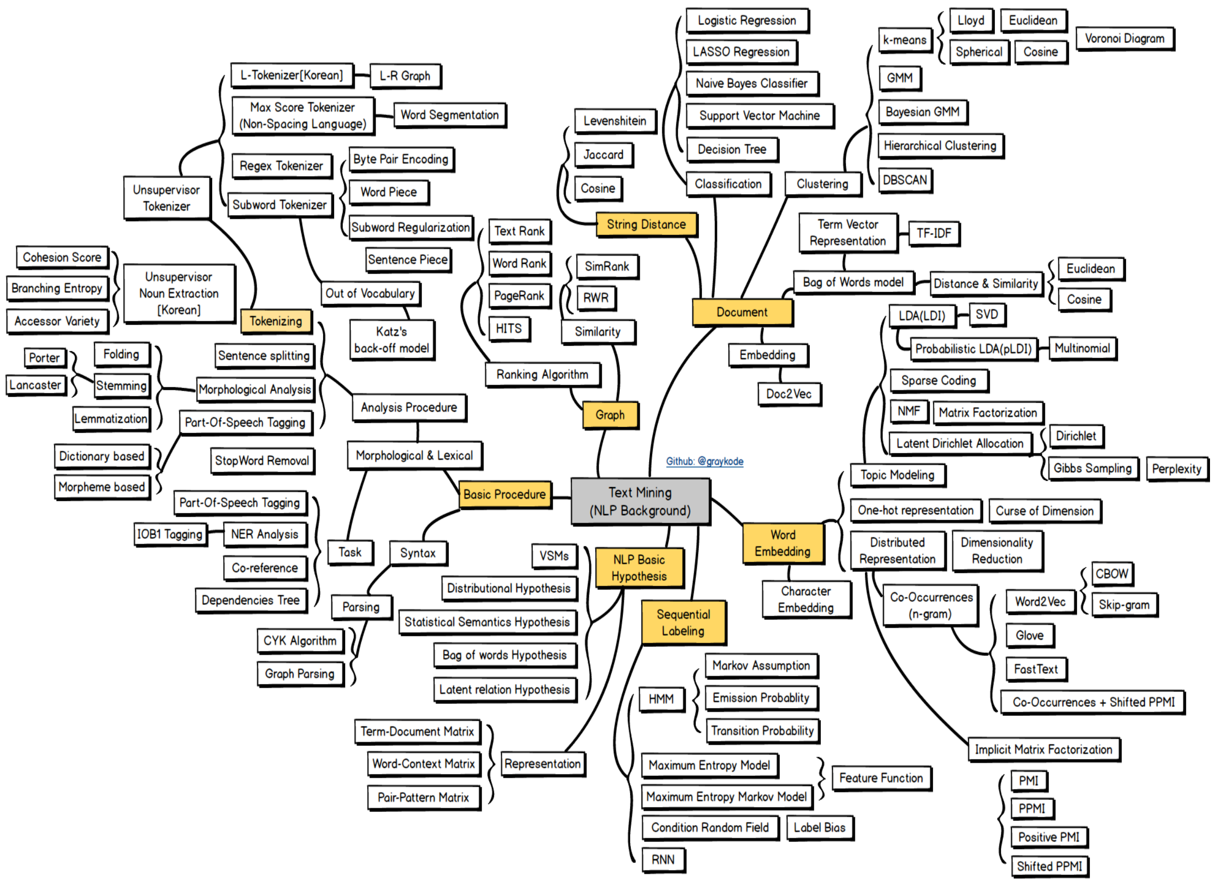
Key Components of NLP

1. Natural Language Understanding (NLU):
   - This involves comprehending the meaning behind human language. It focuses on tasks like sentiment analysis, entity recognition, and intent detection.
2. Natural Language Generation (NLG):
   - This refers to the process of generating human-like text from structured data. Applications include automated report generation and chatbots.
3. Speech Recognition:
   - Converting spoken language into text. This is crucial for applications like virtual assistants (e.g., Siri, Alexa).
4. Text-to-Speech (TTS):
   - The reverse process of speech recognition, where text is converted into spoken words.

NLP ROADMAP :


Natural Language Processing (Based on Deep Learning)
Github: @graykode
Basic
Recurrent Model
Katz-Backoff Model
n-gram
Teacher Forcing
Non-Teacher Forcing
Sequential Model
Bi-directional RNN
Deep RNN
RNN
BPTT
Vanishing Gradient
LSTM
Cell State
GRU
P(wi I w1:i-1)
Greedy Search
Beam Search
Convolutional Model
Kernel Size = n-gram
Channel = number of Perception
TextCNN
DCNN
Recursive Model
Syntatically-Untied RNN
RNTN
Matrix-Vector RNN
Task
POS Tagging
Parsing
Named Entity Recognition
Coreference Resolution
Sentiment Analysis
BLEU Score
Machine Translation
Question Answering
Reading Comprehension
Text Generation
ROUGE Score
Summarization
Dialogue Systems
Perplexity
Language Modelling
Distributed Representation
Word Representation
Co-occurrence Matrix
NNLM
Embedding
Word Embedding
Character Embedding
Hierarchical Softmax
Negative Sampling
Cost Method
Word2Vec
CBOW
Skip-gram
Glove
FastText
Language Model
P(wi I w1:i-1)
Encoder-Decoder Model
Task
Machine Translation
Summarization
Source-Target
OOV Problem Slove
Byte Pair Encoding
Based Frequency
Word Piece Model
Based Likelihood
Subword Regularization
EM Algorithm
Google Sentence Piece
Subword Tokenizing based NMT
Copying Mechanism
Model
Seq2Seq
Attention Mechanism
Context Vector
Alignment Model
Query
Key
Value
Seq2Seq based on
Skip-Thought
Self-Attention
Scale-Dot Product Attention
Transformer
Positional Encoding
Sinusoid Function
Poswise FeedForwardNet
Multi-Head Attention
Masked Multi-Head Attention
Unsupervised Machine Translation
Word Representation to Contextual Representation
CoVe
ELMo
2 Layers LSTM
Pretrained for Embedding
State of the Art Model
Pretraining-Finetuning
Transformer Based Model
Deep Transformer
OpenAI-GPT
P(wi I w1:i-1)
Auxiliary Task Head
BERT
Masking Language Model
Next Sentence Predict(NSP)
CLS
SEP
MASK
Universal Transformer
Inductive Bias of RNN
OpenAI-GPT2
Zero-Shot Learning
Not FineTuning
RoBERTa
Skip NSP
Large Batch
DistilBERT
lighter BERT
Relative Positional Encoding
Segment-Level Recurrence with State Reuse
Transformer XL
Content Stream
Query Stream
Two-Stream Self-Attention for Target-Aware
XLNet
Permutation Language Model
Segment Recurrence Mechanism

Text Mining :



Why Does Natural Language Processing (NLP) Matter ?

One of the essential things in the life of a human being is communication. We must communicate with others to deliver information, express our emotions, present ideas, and much more. The key to communication is language.

We need a common language to communicate, which both ends of the conversation can understand. Doing this is possible for humans, but it might seem a bit difficult if we talk about communicating with a computer system or the computer system communicating with us.It helps the computer system understand the literal meaning and recognize the sentiments, tone, opinions, thoughts, and other components that construct a proper conversation.

NLP is an integral part of everyday life and becoming more so as language technology is applied to diverse fields like Conversational agents such as Amazon's Alexa and Apple's Siri utilize NLP to listen to user queries and find answers. Google uses NLP to improve its search engine results, and social networks like Facebook use it to detect and filter hate speech.

What is NLP Used For ?

Humans communicate in natural language (English, Spanish, Hindi, etc.), but computers understand only numbers. **NLP helps machines process and analyze human language** so they can perform various language-related tasks.

Real-World NLP Applications

Chatbots & Virtual Assistants	Siri, Alexa, ChatGPT
Machine Translation	Google Translate
Spam Filtering	Gmail Spam Detection
Sentiment Analysis	Analyzing customer reviews
Speech Recognition	Voice Commands (Google Assistant)
Text Summarization	News Summaries
Named Entity Recognition (NER)	Identifying names, locations in text
Text Classification	Categorizing news articles

🔹 Example:

- When you ask Alexa, *"What's the weather today?"* – NLP helps it understand and respond.
- Gmail's spam filter automatically detects spam emails using NLP.

Why is NLP Challenging ?

Why Natural Language Processing Is Challenging

Human language stands out for many reasons. It is purposefully designed to express the speaker's or writer's intended meaning. While it is an intricate system, it's impressive that young children can pick it up quickly.

One of the most remarkable aspects of human language is its reliance on symbols. As noted by Chris Manning, a machine learning professor at Stanford, language operates as a discrete, symbolic, and categorical signaling system. This means that the same idea can be expressed in various ways—through speech, gestures, signs, and more. The brain encodes language through a continuous pattern of activation, where these symbols are transmitted via ongoing signals of sound and vision.

Understanding human language poses significant challenges because of its inherent complexity. For instance, there are countless ways to structure words in a sentence. Additionally, words can have multiple meanings, and it's often the surrounding context that enables accurate interpretation. Every language is distinct and, often, ambiguous. A prime example of this is the newspaper headline "The Pope's baby steps on gays." This sentence is open to multiple interpretations, highlighting the difficulties faced in natural language processing.

- Ambiguity: Human language is often ambiguous, with words having multiple meanings depending on context.
- Variability: People express the same idea in various ways, which makes it difficult for machines to understand.
- Complexity: Languages have complex rules regarding grammar and syntax that must be understood for effective processing.

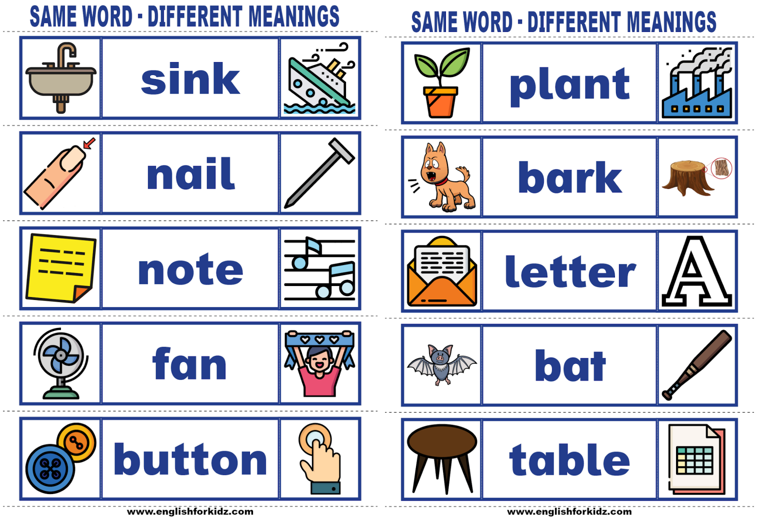
Natural Language Processing (NLP) is inherently challenging due to the complexity and variability of human language. Here are some of the primary reasons why NLP is difficult:

1. Ambiguity : Human language is often ambiguous, with words having multiple meanings depending on context.

- **Lexical Ambiguity**: A single word can have multiple meanings depending on the context. For example, the word "bank" could mean a financial institution, the side of a river, or a place where things are stored (e.g., a data bank).
- **Syntactic Ambiguity**: Syntactic ambiguity occurs when a sentence can be interpreted in more than one way because of its grammatical structure. The **same sequence of words** can be parsed differently, leading to multiple interpretations. could mean either:
  - I used a telescope to see the man.
  - The man I saw was carrying a telescope.
- **Semantic Ambiguity**: Semantic ambiguity occurs when a single word or phrase has multiple meanings, and without additional context, it is unclear which meaning is intended. This ambiguity arises due to the **multiple meanings** of words.

"He went to the **bank**."

- It could mean he went to a **financial institution** or the **edge of a river**.
- **The ambiguity is in the word "bank"**, which has two distinct meanings, and we need more context to understand which meaning is intended.



2. Contextual Understanding

- Human language relies heavily on **context** to derive meaning. Understanding a sentence often requires knowledge about previous sentences or background knowledge that isn't directly stated.
- Contextual understanding is the ability to interpret the meaning of words, phrases, or sentences based on the surrounding context. It helps resolve ambiguities by considering the **larger context** in which something is said.
    - **Context 1: A Timepiece**
      Sentence: "I bought a new watch yesterday."
      Meaning: Here, "watch" refers to a small, portable device used for telling time, worn on the wrist.
    - **Context 2: To Observe or View**
      Sentence: "I watch a movie every weekend."
      Meaning: In this sentence, "watch" means to view or observe something, like a movie or TV show.

    - **Context 1:** River Bank
      Sentence: "We sat by the bank of the river and enjoyed the view."
      Meaning: Here, **"bank"** refers to the side or edge of a river (a geographical feature).

    - **Context 2: Financial Institution**
      Sentence: "I need to go to the bank to withdraw some money."
      Meaning: In this case, **"bank"** refers to a financial institution where people keep their money, like a savings or checking account.

homonyms
bat
trunk
park
bark
bank
rock
saw
row

3. Subtlety of Language

- It refers to the nuanced and often indirect ways in which people express ideas, emotions, or intentions. It involves the subtle differences in meaning that can be conveyed through **tone**, **word choice**, **emotion**,**sarcasm**, **humor**, and even **cultural context**. These subtleties can make language complex and difficult for machines to fully understand, as they rely not just on the **literal meaning** of words but on their **underlying implications**.
    - **"A fish out of water":** Refers to someone who feels uncomfortable or out of place, but its literal meaning refers to a fish being outside its natural habitat, which doesn't make sense if taken literally.
- Recognizing these subtleties is often outside the scope of traditional NLP models, making tasks like **sentiment analysis** more complicated.

4. Syntax and Grammar Variability

- There are many ways to structure sentences that still convey the same meaning. In English, word order typically follows the **Subject-Verb-Object (SVO)** pattern. However, variations in syntax can still maintain grammatical correctness.
    - **Sentence 1**: "The cat chased the mouse."
    - **Sentence 2**: "The mouse was chased by the cat."

    Both sentences have the same meaning (a cat chasing a mouse), but the **word order** is different. Sentence 2 uses the **passive voice** structure, which swaps the subject and object around.

- These variations are grammatically correct in different contexts but are difficult for algorithms to process effectively.
- **Languages** themselves have highly diverse syntax and grammar rules. What works in one language may not apply in another (e.g., subject-object-verb order in English vs. subject-verb-object in Japanese).

5. Multilinguality and Dialects

Multilinguality refers to the ability to understand and use multiple languages. It encompasses the use of two or more languages by individuals, communities, or systems. In the context of Natural Language Processing (NLP), multilinguality involves processing and understanding texts in different languages or dealing with translation between languages.NLP systems trained on one language may not work well in another language due to differences in grammar, vocabulary, and syntax.

- Dialects can differ based on the **geographic region** where they are spoken.
- Additionally, dialects and regional variations within a language pose a challenge. For example, American English differs from British English in terms of spelling, slang, and idioms. Similarly, various dialects of Spanish or Arabic differ substantially.
- In fact, **multilingual NLP** requires models that can generalize across many languages, which is a difficult task due to these linguistic differences.

6. Data Scarcity and Labeling

- To train effective NLP models, large amounts of labeled data are often required. However, in many cases, especially for less common languages , large datasets are not readily available.
- **Labeling** data is also time-consuming and expensive, especially when it involves things like sentiment analysis, where human annotators have to understand and tag subtle emotions or opinions in text.

7. Text Variability

- **Synonymy**: Words with similar meanings can be used interchangeably in sentences. For example, “happy” and “joyful” convey the same idea, but an NLP system needs to understand that they are related to one another.
- **Paraphrasing**: People Express the same idea or concept in different ways. For instance, "Can you help me?" , "I need your help." or "Can you lend me a hand?" . In All the above cases the **intent** (asking for help) remains the same, but the **phrasing** is different. The machine needs to understand that these different forms essentially convey the same meaning.
- **Context-dependent Words and Phrases**: Many words change meaning depending on the surrounding words or context. Words like "bark" or "match" have different meanings based on the context in which they appear.

8. Long-term Dependencies

- In many NLP tasks, it’s crucial to understand **long-term dependencies** in text, such as understanding how previous sentences or even paragraphs relate to the current one.

Earlier models like **RNNs** and **LSTMs** had limitations in capturing these long-term dependencies effectively, though **transformer-based models** like **BERT** and **GPT** perform much better in this regard.

9. Idiomatic Expressions and Phrasal Verbs

- Languages often use **idiomatic expressions** and **phrasal verbs** that don't make sense when taken literally. For example:
  - "Kick the bucket" (meaning to die) is an idiom that a model must learn to recognize as a specific phrase.
  - "Get over" can mean "to recover from" or "to overcome" in different contexts, and distinguishing between those meanings can be difficult for machines.

10. Irony and Sarcasm

- Understanding irony and sarcasm is a massive challenge. Humans typically rely on tone, facial expressions, and context, but these are not directly available in text.
- For example, "That was just perfect!" could be positive or negative depending on the tone and context. A deep understanding of the situation and cultural norms is often needed to interpret this correctly, and it's difficult for AI to infer this information accurately.

11. Entity Recognition

- Identifying **named entities** like **people's names**, **organizations**, **locations**, **dates**, etc., is a challenging task because names and references may not follow strict conventions, and the same entity can be referred to in many ways (e.g., "U.S." vs. "America").
- Furthermore, some words can be ambiguous, like "Apple" (the company) vs. "apple" (the fruit), depending on the context.

12. Language Evolution

- Language constantly evolves with new words, phrases, and meanings entering the vocabulary. NLP systems need to adapt to these changes to stay relevant. Slang, cultural references, and internet-specific language (e.g., memes, hashtags, and emojis) add another layer of complexity.

Conclusion: Why NLP Is Hard

Natural language is highly complex, full of ambiguity, nuance, and subtlety. Understanding context, recognizing cultural or idiomatic references, managing linguistic diversity (across dialects, languages, and even historical usage), and dealing with inconsistent grammar and structure are some of the challenges that make NLP difficult.

While progress has been made, especially with **deep learning models** and **transformer-based architectures** (like BERT and GPT), there are still significant challenges in achieving human-level understanding and interpretation of language. The field of NLP continues to advance, but perfect, all-encompassing models are still a long way off.

📌 Benefits of NLP

- Improved Communication: Enhances interaction between humans and machines.
- Automation: Reduces manual effort in data analysis and customer service.
- Insights from Data: Extracts valuable information from vast amounts of unstructured text data.

📌 History of NLP: Classical NLP vs. Deep Learning-based NLP

Understanding the **history of NLP** will give you a solid foundation to see how NLP evolved from rule-based methods to the latest **Deep Learning-based NLP** models like BERT and GPT.

NLP has **evolved over time** through different approaches:

Era	Approach	Key Characteristics
1950s-1980s	Rule-Based NLP	Manually written rules, symbolic AI
1980s-1990s	Statistical NLP	Probability-based models, Hidden Markov Models (HMMs)
2000s-2014	Traditional ML-based NLP	Feature engineering + ML models (SVM, Naïve Bayes, Decision Trees)
2014-Present	Deep Learning-based NLP	RNNs, LSTMs, Transformers (BERT, GPT)

🔹 1. Classical NLP (Rule-Based & Statistical Approaches)

🏛️ 1950s-1980s: Rule-Based NLP (Symbolic AI)

- **Approach:** Linguists manually created rules for grammar and sentence structure.

- **Example:** "If a sentence has 'I' + 'am' → it's a present tense sentence."
- **Challenges:**
  - Rules do not generalize well for large datasets.
  - Could not handle ambiguous meanings.

Example: Early **machine translation systems** were purely rule-based.

📊 1980s-1990s: Statistical NLP (Probability-Based Models)

- **Approach:** Used **probabilities and statistics** instead of hardcoded rules.
- **Key Models:**
  Hidden Markov Models (HMMs) – Used for **POS tagging, Speech Recognition**.
  n-gram models – Used for **language modeling** (predicting next words).

🔹 **Example:** Spam detection using **Naïve Bayes classifier** (Probability-based classification).

📌 Limitations of Statistical NLP:

- Needed **lots of labeled data**.
- Could not **capture deep contextual meanings** of words.

🔹 2. Machine Learning-Based NLP (2000s-2014)

🔹 Shift from Rule-Based to Machine Learning Models

With more **data** and **computational power**, NLP moved to **feature-based machine learning models**.

Key ML Models in NLP

Naïve Bayes Classifier → Used for spam filtering.
Support Vector Machines (SVMs) → Used for text classification.
Decision Trees & Random Forests → Used for sentiment analysis.

📌 Limitations:

- Required **manual feature engineering** (e.g., TF-IDF, Bag of Words).
- Could not handle **sequential dependencies** (word order).

🔹 **3. Deep Learning-Based NLP (2014–Present)**

🔹 **The Deep Learning Revolution in NLP** started when researchers **replaced feature engineering with deep learning models**.

2014-2016: RNNs & LSTMs (First Wave of Deep NLP)

- **Recurrent Neural Networks (RNNs)** → Used for **sequence-based NLP** (Text Generation, Speech Recognition).
- **Long Short-Term Memory (LSTMs)** → Solved **short-term memory issues** of RNNs.

🔹 **Example:** Google Translate switched from **Statistical NLP → LSTMs** in 2016.

2017-Present: Transformer Models (Second Wave of Deep NLP)

2017 – Transformers (Attention Is All You Need)

- **Transformers** replaced RNNs & LSTMs.
- **Self-Attention Mechanism** improved NLP tasks like translation & summarization.

2018 – BERT (Bidirectional Encoder Representations from Transformers)

- **BERT** understands word **context better** than previous models.
- **Example:** "I went to the bank to deposit money" vs. "I sat on the river bank."

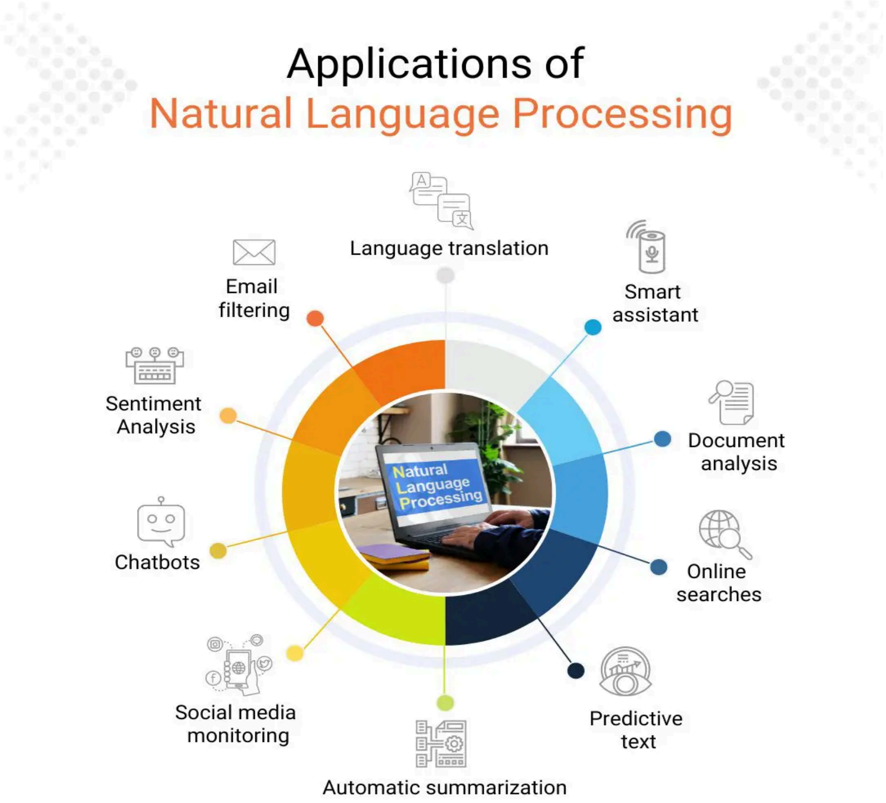
2020-Present – GPT Models (Generative NLP)

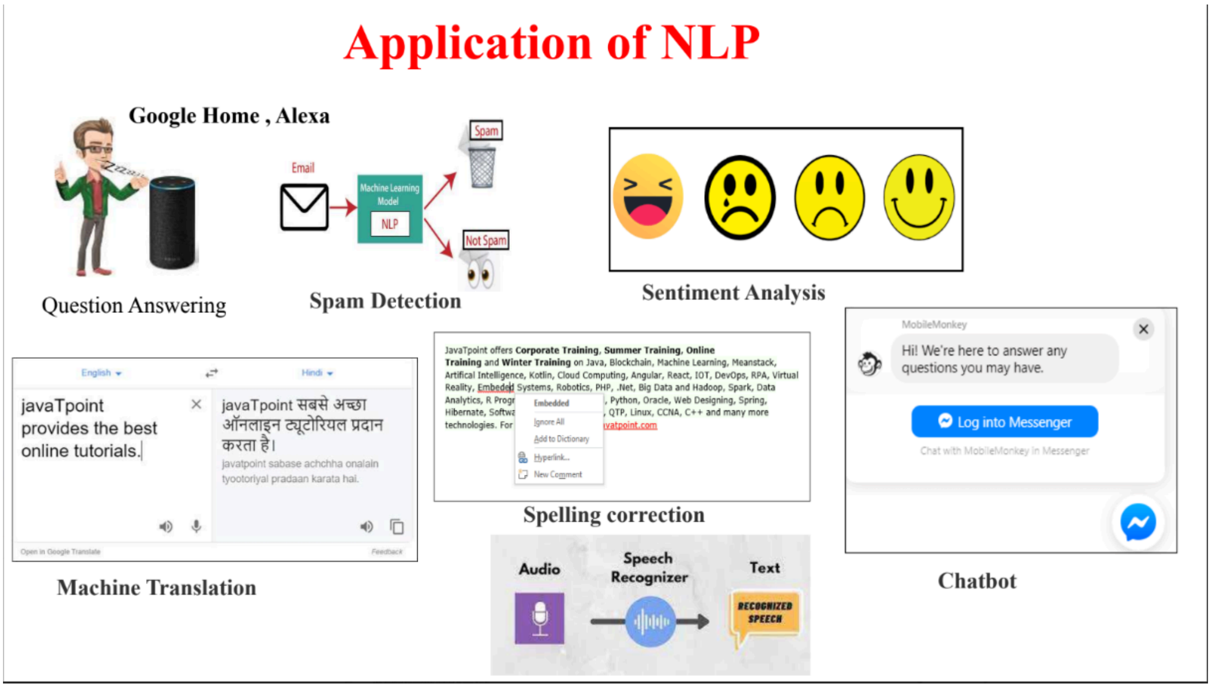
- **GPT-3, GPT-4** can generate **human-like** text.
- Used in **chatbots, content writing, question answering, and more**.

📌 Summary: NLP Evolution

Era	Method	Models Used	Limitations
1950s-1980s	Rule-Based NLP	Handcrafted rules	Hard to scale
1980s-1990s	Statistical NLP	HMMs, n-grams	Needed large data
2000s-2014	Machine Learning	Naïve Bayes, SVMs	Manual feature engineering
2014-2017	Deep Learning (RNNs, LSTMs)	RNNs, LSTMs	Hard to handle long text
2017-Present	Transformers	BERT, GPT, T5	Requires massive computing

📌 Applications of NLP





🔹 1. Text Processing & Information Retrieval

Search Engines:

- NLP helps search engines like **Google, Bing, and Yahoo** understand user queries and rank relevant results.
- Example: **Google Search Autocomplete, Voice Search, Semantic Search.**

Text Summarization:

- Generates short, meaningful summaries of long articles or documents.
- Example: **AI-powered news summarization (Google News, Summari AI).**

Optical Character Recognition (OCR):

- Converts scanned text/images into machine-readable text.
- Example: **Google Lens, Adobe Acrobat OCR, CamScanner.**

🔹 2. Conversational AI & Chatbots

Chatbots & Virtual Assistants:

- AI-driven chatbots for answering queries and providing customer support.
- Example: **ChatGPT, Google Assistant, Siri, Alexa, Cortana.**

AI Customer Support:

- Businesses use NLP-powered bots to automate responses and solve customer issues.
- Example: **Banking Chatbots (HDFC EVA, Capital One Eno).**

Voice Assistants:

- NLP allows assistants to process and respond to voice commands.
- Example: **Amazon Alexa, Google Home, Apple Siri.**

Conversational Agents for Mental Health:

- AI-driven chatbots help users manage mental health by providing counseling.
- Example: **Woebot, Wysa (Mental Health AI Chatbots).**

🔹 3. Text Understanding & Classification

Sentiment Analysis:

- Determines whether a text is positive, negative, or neutral.
- Example: **Social media monitoring, Product reviews analysis (Amazon, Twitter, Facebook).**

Spam Detection:

- Filters out spam emails using NLP techniques.
- Example: **Gmail Spam Filters, Outlook Spam Detection.**

Text Classification & Categorization:

- Categorizes text into predefined topics (e.g., News, Sports, Politics).
- Example: **Google News article categorization.**

Hate Speech & Fake News Detection:

- Identifies toxic or misleading content.
- Example: **Facebook, Twitter AI content moderation.**

🔹 4. Language Translation & Generation

Machine Translation:

- Automatically translates text between languages.
- Example: **Google Translate, DeepL, Microsoft Translator.**

Text Generation (NLG - Natural Language Generation):

- AI models generate human-like text for various applications.
- Example: **ChatGPT, GPT-4, Copy.ai for content writing.**

Speech-to-Text (Automatic Speech Recognition - ASR):

- Converts spoken language into written text.
- Example: **Google Speech-to-Text, YouTube Auto-Captions.**

Text-to-Speech (TTS):

- Converts text into natural-sounding speech.
- Example: **Google Assistant’s Voice, Apple’s VoiceOver, Amazon Polly.**

Paraphrasing & Rewriting Tools:

- AI-based tools rewrite sentences while preserving meaning.
- Example: **QuillBot, Grammarly Paraphraser.**

🔹 5. Healthcare & Medical NLP

Clinical Text Analysis:

- Extracts insights from medical records and prescriptions.
- Example: **IBM Watson Health, Google's Med-PaLM.**

Medical Chatbots:

- AI-powered bots assist in diagnosing symptoms.
- Example: **Ada Health, Babylon Health.**

Drug Discovery & Research Analysis:

- NLP helps analyze medical literature for new treatments.

- Example: **DeepMind’s AlphaFold for protein structure prediction.**

Healthcare Documentation Automation:

- Converts doctor-patient conversations into medical records.
- Example: **Nuance Dragon Medical AI.**

🔹 6. Legal & Finance NLP Applications

Contract & Legal Document Analysis:

- AI extracts important clauses and insights from legal documents.
- Example: **ROSS Intelligence (Legal AI).**

Automated Financial Reports Generation:

- AI summarizes financial data into reports.
- Example: **Bloomberg GPT (AI in finance).**

Fraud Detection:

- NLP analyzes transactions and detects anomalies.
- Example: **PayPal’s AI-driven fraud detection system.**

🔹 7. E-Commerce & Retail Applications

Product Recommendations:

- NLP-based recommendation engines personalize user experience.
- Example: **Amazon, Netflix, and Spotify recommendations.**

Review Analysis & Feedback Processing:

- AI extracts insights from product reviews and feedback.
- Example: **Amazon’s AI-driven review summarization.**

Price Comparison & Competitive Analysis:

- NLP analyzes competitor pricing and market trends.
- Example: **Google Shopping AI, PriceRunner.**

🔹 8. Social Media & Content Moderation

Social Media Sentiment Analysis:

- Tracks brand reputation and public opinion.
- Example: **Brand monitoring tools like Brandwatch, Hootsuite.**

Toxicity Detection & Content Moderation:

- Detects hate speech, offensive language, and misinformation.
- Example: **Facebook's AI moderation, Twitter's automated filters.**

Trend Analysis & Hashtag Tracking:

- AI identifies viral trends and hashtags.
- Example: **Twitter Trending Topics AI.**

🔹 9. Education & Learning

AI-Powered Language Learning:

- NLP assists in learning new languages.
- Example: **Duolingo, Rosetta Stone, Babbel.**

Automated Essay Scoring & Grading:

- AI evaluates and grades student essays.
- Example: **ETS e-rater (used in TOEFL grading).**

Plagiarism Detection:

- NLP checks for duplicate content.
- Example: **Turnitin, Grammarly Plagiarism Checker.**

AI Tutoring Systems:

- NLP-powered tutors answer student queries.
- Example: **Socratic by Google, Khan Academy AI Tutor.**

📌 Summary Table: NLP Applications & Examples

Category	Application	Examples
Text Understanding	Search Engines, OCR, Text Summarization	Google Search, Google Lens
Conversational AI	Chatbots, Virtual Assistants	ChatGPT, Alexa, Siri
Text Understanding	Sentiment Analysis, Spam Detection	Twitter AI, Gmail Spam Filter
Language Translation	Machine Translation, Speech-to-Text	Google Translate, DeepL
Healthcare NLP	Medical Chatbots, Clinical Text Analysis	IBM Watson Health, Nuance Dragon AI
Finance & Legal	Contract Analysis, Fraud Detection	ROSS Intelligence, PayPal AI
E-Commerce	Recommendations, Review Analysis	Amazon, Netflix AI
Social Media	Toxicity Detection, Sentiment Analysis	Facebook AI, Twitter Moderation
Education	AI Tutors, Automated Essay Scoring	Duolingo, Turnitin AI

Is NLP divided into two parts NLP with ML and NLP with DL ?

Let's dive into the details of **NLP with Machine Learning (ML)** and **NLP with Deep Learning (DL)** to understand the differences, the methods used, and how they evolve from one another. NLP (Natural Language Processing) can broadly be categorized based on the type of algorithms and techniques (techniques and models) used to process and analyze natural language.

1. NLP with Machine Learning (ML)

This involves techniques that rely on manually designed rules or statistical models without requiring significant amounts of labeled data. Traditional NLP techniques primarily involve **feature engineering** and the use of machine learning algorithms.

Here's a breakdown of how it works:

Key Components: In these methods, feature engineering plays a significant role. Textual data is preprocessed and transformed into numerical representations (like **Bag of Words (BoW)** or **TF-IDF** vectors), which are then used by machine learning models.

1. **Feature Engineering:**
    - Feature Engineering is the first step in classical NLP that plays a significant role . It involves transforming raw text data into numerical features that machine learning algorithms can process. This process is called **feature engineering**.
    - Common techniques for feature extraction include:
        - **Bag of Words (BoW)**: Represents text as a collection of word counts or binary word occurrences, disregarding grammar and word order.
        - **Term Frequency-Inverse Document Frequency (TF-IDF)**: Weighs words based on their importance in a document relative to all documents in the corpus. This technique helps prioritize important words and reduces the weight of common words.
        - **Word n-grams**: Considers consecutive sequences of words (like bigrams or trigrams), capturing some level of word order.
2. **Algorithms Used:**
    - Once the features are extracted, traditional machine learning algorithms are applied to solve NLP tasks. These algorithms include:
        - **Logistic Regression**: A simple but powerful model for binary classification tasks (e.g., sentiment analysis: positive or negative).
        - **Support Vector Machines (SVM)**: Used for classification tasks, where the goal is to separate classes with the maximum margin.
        - **Decision Trees and Random Forests**: Used for classifying documents based on extracted features.
        - **Naive Bayes**: A probabilistic classifier based on Bayes' theorem, commonly used for text classification tasks (e.g., spam detection).
        - Traditional NLP approaches often use **machine learning** algorithms, such as **decision trees, random forests, and support vector machines (SVM) .**

- These algorithms typically require labeled data for supervised learning and work well for relatively simpler NLP tasks like **text classification**, **spam detection**, part-of-speech tagging , **named entity recognition**. **Classical NLP tasks** , sentiment analysis) were often solved using these techniques.

3. **Applications of ML in NLP:**

    - **Text Classification**: Classifying text into predefined categories (e.g., spam vs. non-spam).
    - **Named Entity Recognition (NER)**: Identifying entities such as names, dates, locations, etc.
    - **Part-of-Speech (POS) Tagging**: Identifying the grammatical components of sentences.
    - **Sentiment Analysis**: Determining the sentiment (positive, negative, neutral) of a piece of text.

    ✅Terminology: Often referred to as **Statistical NLP** or **Machine Learning-based NLP**.

Limitations of ML for NLP:

- **Manual Feature Engineering**: Creating good features requires domain knowledge and can be time-consuming. This is a key limitation of ML-based NLP methods.
- **Sparse Representations**: Methods like Bag of Words can lead to very sparse, high-dimensional feature vectors that might not capture the nuanced meaning of words or phrases.
- **Difficulty with Context**: Traditional ML approaches struggle to capture the complex context of language (e.g., sarcasm, word ambiguity).

---

2. NLP with Deep Learning (DL) Deep Learning-Based NLP :

Description:With the rise of deep learning in the last decade, NLP has undergone a transformation. Deep learning methods use neural networks to automatically learn features and patterns from large datasets, eliminating the need for extensive manual feature engineering.

This leverages neural networks, especially deep learning architectures, to automatically learn representations of text data. It eliminates the need for extensive manual feature engineering by learning features from raw data.With the rise of deep learning, NLP has evolved significantly. **Deep Learning** (DL) uses neural networks, particularly **recurrent neural networks (RNNs), long short-term memory networks (LSTMs), and transformers** like **BERT** and **GPT**.

- These models can automatically learn complex patterns from vast amounts of data without requiring heavy manual feature engineering. Deep learning-based models have shown superior performance in various NLP tasks such as **language generation, machine translation, text classification, and question answering**.
- **Pretrained models** such as BERT and GPT are examples of deep learning methods that have revolutionized NLP.

Let's break down the key components of NLP with deep learning:

Key Components:

1. **Neural Networks in NLP:**
    - **Neural networks** are a key part of deep learning. A neural network consists of layers of nodes (neurons) connected in such a way that each layer learns to transform the input into increasingly abstract representations.
    - In NLP, **recurrent neural networks (RNNs)** and **Long Short-Term Memory (LSTM)** networks are widely used because they can capture sequential information, making them ideal for tasks involving text, which has inherent sequential structure.
2. **Word Embeddings:**
    - One of the major advances with deep learning in NLP is the use of **word embeddings**.
    - **Word embeddings** represent words as dense vectors in a continuous vector space where semantically similar words are closer together.
        - **Word2Vec**: A neural network-based model that learns distributed representations of words.
        - **GloVe (Global Vectors for Word Representation)**: Another method to learn word embeddings using co-occurrence statistics.
    - These embeddings capture the meaning of words and relationships between them, unlike traditional methods (BoW, TF-IDF) that ignore context and relationships.
3. **Recurrent Neural Networks (RNNs):**
    - RNNs are designed for sequential data, making them well-suited for text. They process input one step at a time while maintaining an internal state that captures information about previous steps (words).
    - **LSTM (Long Short-Term Memory)** is a special kind of RNN that solves the problem of **vanishing gradients**, making it better at learning long-term dependencies in text.
4. **Transformers:**

    - The introduction of the **transformer** model (in the paper “Attention Is All You Need” by Vaswani et al.) revolutionized NLP. Unlike RNNs, transformers

do not process input sequentially, but rather consider the entire context of a sentence or document in parallel.
    - **Attention mechanisms** allow the model to focus on the most relevant words in a sentence when making predictions. This makes transformers much more efficient and capable of capturing long-range dependencies in text.
    - **BERT (Bidirectional Encoder Representations from Transformers)**, **GPT (Generative Pretrained Transformer)**, and **T5** are state-of-the-art transformer models that have been pre trained on large datasets and fine-tuned for specific NLP tasks.
5. **Pretrained Models (Transfer Learning):**
    - One of the game-changers in NLP with deep learning is the use of **pretrained models**. Models like **BERT**, **GPT**, and **RoBERTa** are first trained on large, generic corpora (such as books, articles, etc.) and then fine-tuned on specific NLP tasks.
    - Pretrained models have significantly improved the performance of NLP tasks and require much less labeled data for fine-tuning.

Applications of DL in NLP:

- **Machine Translation**: Translating text from one language to another (e.g., Google Translate uses transformer-based models).
- **Question Answering**: Models like BERT can answer questions based on a given passage of text.
- **Text Generation**: Models like GPT can generate human-like text, which is used for chatbots, content generation, and more.
- **Text Summarization**: Creating concise summaries from long articles or documents.
- **Sentiment Analysis**: Deep learning models can better capture the nuanced sentiments in text, even in complex cases like sarcasm.

Key Components:

- Word Embeddings (e.g., Word2Vec, GloVe, fastText)
- Sequence Models (e.g., RNNs, LSTMs, GRUs)
- Attention Mechanisms and Transformers (e.g., BERT, GPT)
- End-to-end models for tasks like sentiment analysis, text classification, and machine translation

Advantages of DL over ML:

- **Minimal Feature Engineering**: Deep learning models automatically learn features, saving significant effort compared to manual feature extraction in traditional machine learning.

- **Contextual Understanding**: DL models, particularly transformers, capture long-range dependencies and the context of words, which is essential for understanding meaning in text.
- **Scalability**: Deep learning models can scale to massive datasets and can improve as more data becomes available.

Challenges with DL:

- **Data Requirements**: Deep learning models generally require large amounts of labeled data to perform well, which can be a limitation for some applications.
- **Computational Resources**: Training deep learning models requires powerful hardware, especially GPUs or TPUs, making it resource-intensive.
- **Interpretability**: Deep learning models, particularly deep neural networks, are often seen as "black boxes" because it's difficult to interpret why a particular decision was made.

✅**Terminology**: Commonly referred to as **Neural NLP** or **Deep Learning-based NLP**.

Summary of Differences:

Aspect	NLP with Machine Learning (ML)	NLP with Deep Learning (DL)
Approach	Relies on handcrafted features and traditional ML algorithms	Uses deep neural networks to automatically learn features
Feature Engineering	Required (e.g., Bag of Words, TF-IDF)	Not required (models learn features automatically)
Complexity	Simpler models, easier to interpret	More complex models, harder to interpret
Context Understanding	Limited context understanding (sequential models like SVM)	Superior context understanding (especially with transformers)
Data Requirements	Works with smaller datasets	Requires large datasets for effective training
Scalability	Can be scaled, but may struggle with very large datasets	Highly scalable with access to large data and computational power
Performance	Adequate for simpler tasks	State-of-the-art performance in many NLP tasks

Conclusion:

- **NLP with Machine Learning:** refers to traditional machine learning-based approaches for language tasks.It is suitable for simpler tasks and smaller datasets, where feature engineering can still capture meaningful patterns
- **NLP with Deep Learning :** uses deep learning methods that can handle large-scale, complex data with minimal manual intervention and have become dominant in recent years. It has revolutionized the field, especially for complex tasks like machine translation, text generation, and contextual understanding. It excels in handling large datasets, capturing deeper nuances in language, and often achieves superior performance with minimal feature engineering.

NLP with MLNLP with DL Both methods are widely used, but the deep learning approach is now more prominent in cutting-edge NLP applications.

While deep learning is now the dominant force in NLP, traditional ML approaches still have their place, particularly for smaller-scale projects or in cases where computational resources are limited.

📌 Other Resources :

- Book: "Speech and Language Processing" by Daniel Jurafsky and James H. Martin [Speech and Language Processing – Jurafsky & Martin] (A must-read book)

🚀 **What's Next?**

✅ **Performances Metrics**

✅ **Text Preprocessing (Tokenization, Stopword Removal, Lemmatization)**
✅ **Deep Dive into Word Embeddings (Word2Vec, GloVe, BERT Embeddings)**