Introduction to Natural Language Processing

Natural Language Processing

- Understanding the <mark>correct meaning of the sentence</mark>,
- <mark>Correct Named-Entity Recognition</mark>(NER)
- <mark>Correct prediction of various parts of speech</mark>
- <mark>Solving a complex problem in Machine Learning means building a pipeline</mark>.
- <mark>In simple terms, it means breaking a complex problem into a number of small problems, making models for each of them and then integrating these models</mark>.
- A similar thing is done in NLP.

Natural Language Processing…

- It is a subfield of computer science and artificial intelligence that deals with the interaction between computers and human languages.

- The goal of NLP is to enable computers to understand, interpret, and generate natural language, the way humans do.

- NLP involves a variety of techniques, including computational linguistics, machine learning, and statistical modeling.

- These techniques are used to analyze, understand, and manipulate human language data, including text, speech, and other forms of communication.

Natural Language Processing…

- NLP has a wide range of applications, including sentiment analysis, machine translation, text summarization, chatbots, and more.

- Text Classification:
  - Classifying text into different categories based on their content, such as spam filtering, sentiment analysis, and topic modeling.

- Named Entity Recognition (NER): Identifying and categorizing named entities in text, such as people, organizations, and locations.

- Part-of-Speech (POS) Tagging: Assigning a part of speech to each word in a sentence, such as noun, verb, adjective, and adverb.

Natural Language Processing

NLP
Natural Language
Processing

NLG
Natural Language
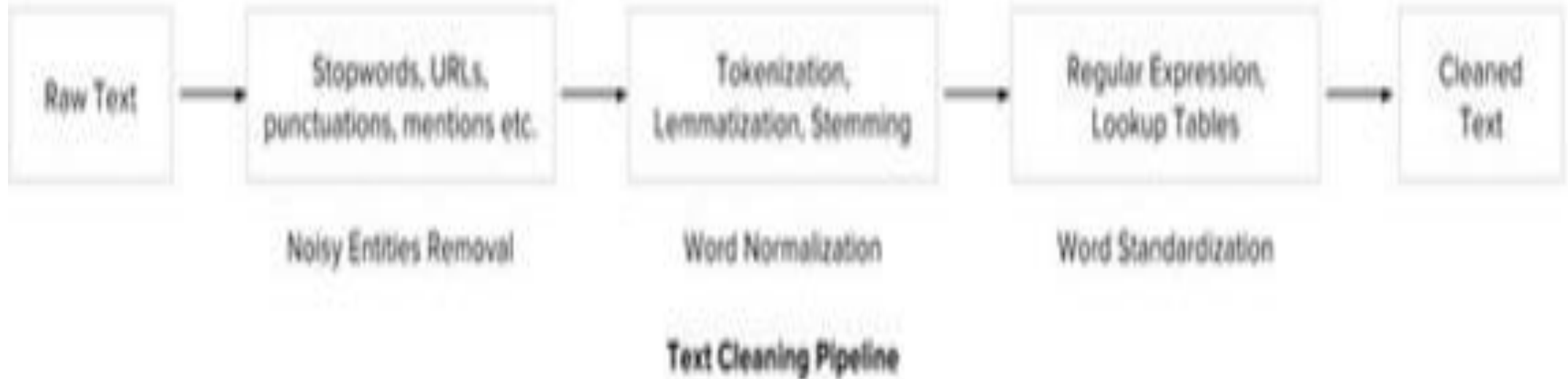Generation

NLU
Natural Language
Understanding

NLP Terminology

- Tokenization – process of converting a text into tokens

- Tokens – words or entities present in the text

- Text object – a sentence or a phrase or a word or an article

Text Preprocessing

- <mark>Text is the most unstructured form of all the available data</mark>.

- Various types of noise are present in it and the data is not readily analyzable without any pre-processing.

- The entire process of cleaning and standardization of text, making it noise-free and ready for analysis is known as text preprocessing.

- Noise Removal

- Lexicon Normalization

- Object Standardization

Text Preprocessing…

Raw Text	→	Stopwords, URLs, punctuations, mentions etc.	→	Tokenization, Lemmatization, Stemming	→	Regular Expression, Lookup Tables	→	Cleaned Text
		Noisy Entities Removal		Word Normalization		Word Standardization		

Text Cleaning Pipeline

Text Preprocessing…

- Noise Removal
  - Any piece of text which is not relevant to the context of the data and the end-output can be specified as the noise.
  - For example – language stop words (commonly used words of a language – is, am, the, of, in etc), URLs or links, social media entities (mentions, hashtags), punctuations and industry specific words.
  - This step deals with removal of all types of noisy entities present in the text.

- Lexicon Normalization:
  - Another type of textual noise is about the multiple representations exhibited by single word.
  - For example – "play", "player", "played", "plays" and "playing" are the different variations of the word – "play", Though they mean different but contextually all are similar.
  - The step converts all the disparities of a word into their normalized form (also known as lemma).

- Stemming: the process of reducing infected words to their stem ("ing", "ly", "es", "s" etc) from a word.

- Stemming is a process that stems or removes last few characters from a word, often leading to incorrect meanings and spelling.

- Examples: sentiment analysis, spam classification, restaurant reviews etc., getting base word is important to know whether the word is positive or negative.

- Stemming is used to get that base word.

History
Historical
> Histori

Finally
Final
> Fina

- Lemmatization:
  - Lemmatization considers the context and converts the word to its meaningful base form, which is called Lemma.
  - Lemmatization algorithms often rely on linguistic rules and patterns.
  - The purpose of lemmatization is same as that of stemming but overcomes the drawbacks of stemming.
  - Lemmatization takes more time as compared to stemming because it finds meaningful word/ representation.
  - Stemming has its application in Sentiment Analysis while Lemmatization has its application in Chatbots, human-answering.

  - stemming the word 'Caring' would return 'Car'.
  - lemmatizing the word '**Caring**' would return '**Care**'.

- Object Standardization:
  - Text data often contains words or phrases which are not present in any standard lexical dictionaries.
  - These pieces are not recognized by search engines and models.
  - examples are – acronyms, hashtags with attached words, and colloquial slangs.
  - With the help of regular expressions and manually prepared data dictionaries, this type of noise can be fixed, the code below uses a dictionary lookup method to replace social media slangs from a text.
  - other types of text preprocessing includes encoding-decoding noise, grammar checker, and spelling correction etc

Text to Features (Feature Engineering on text data)

- To analyze a preprocessed data, it needs to be converted into features.
- Depending upon the usage, text features can be constructed using assorted techniques – Syntactical Parsing, Entities / N-grams / word-based features, Statistical features, and word embeddings.
- Syntactic Parsing:
- It involves the analysis of words in the sentence for grammar and their arrangement in a manner that shows the relationships among the words.

- Dependency Trees:
- Sentences are composed of some word sewed together.
- The relationship among the words in a sentence is determined by the basic dependency grammar.
- Dependency grammar is a class of syntactic text analysis that deals with (labeled) asymmetrical binary relation between two lexical items (words).
- For example: consider the sentence – "Bills on ports and immigration were submitted by Senator Brownback, Republican of Kansas."
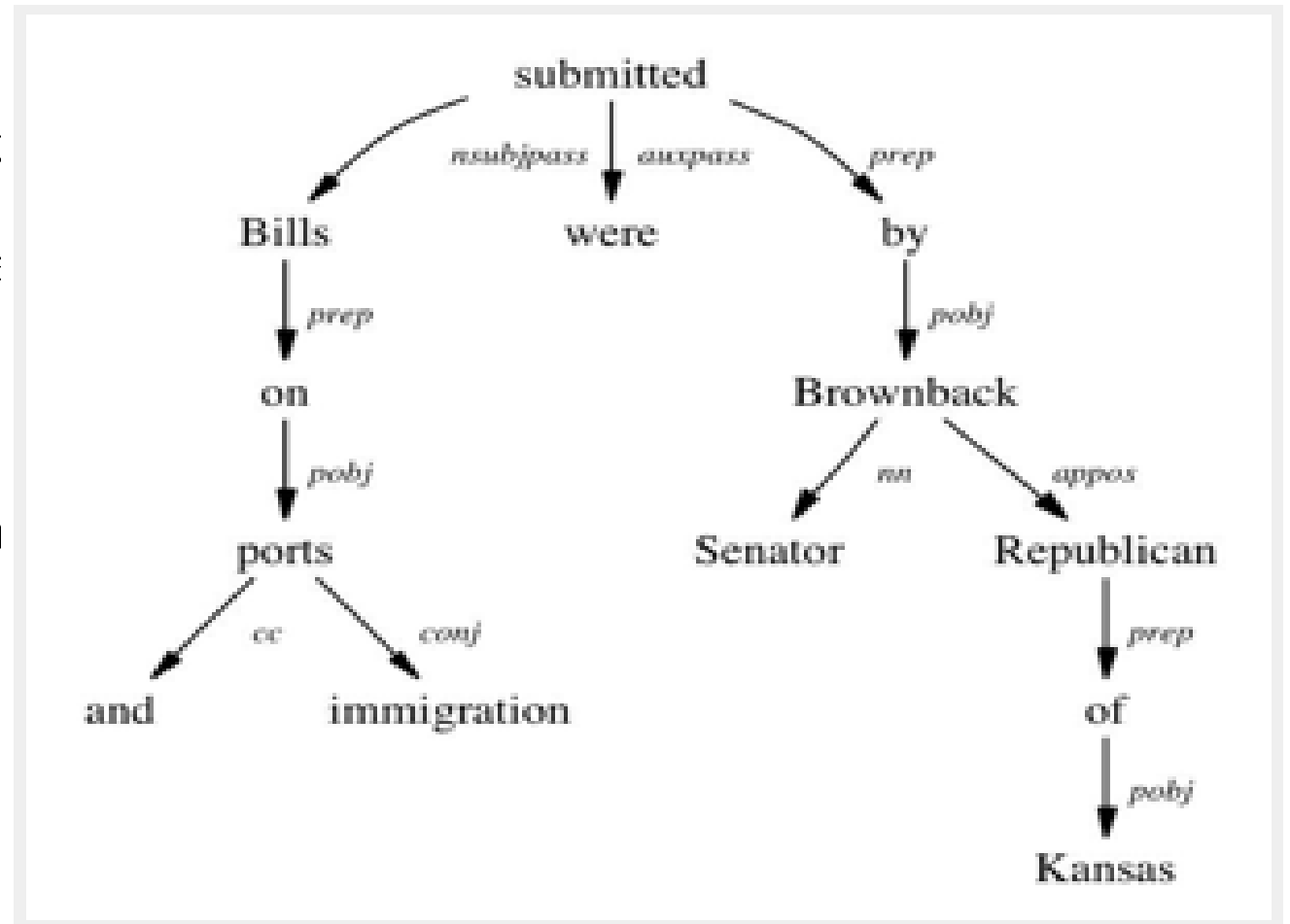
- Part of speech tagging :
  - Apart from the grammar relations, every word in a sentence is also associated with a part of speech (pos) tag (nouns, verbs, adjectives, adverbs etc).
  - The pos tags defines the usage and function of a word in the sentence.
  - text = "I am learning Natural Language Processing."
  - Pos: I, am, Learning, Natural, Langiuage, processing,
- Part of Speech tagging is used for many important purposes in NLP:
- Word sense disambiguation: Some language words have multiple meanings according to their usage.
- I. "Please book my flight for Delhi"
- II. "I am going to read this book in the flight"
- "Book" is used with different context, however the part of speech tag for both of the cases are different. In sentence I, the word "book" is used as verb, while in II it is used as noun.

- Improving word-based features:
- A learning model could learn different contexts of a word when used word as the features, however if the part of speech tag is linked with them, the context is preserved, thus making strong features.
- For example:
- Sentence -"book my flight, I will read this book"

- Tokens – ("book", 2), ("my", 1), ("flight", 1), ("I", 1), ("will", 1), ("read", 1), ("this", 1)

- Tokens with POS – ("book_VB", 1), ("my_PRP$", 1), ("flight_NN", 1), ("I_PRP", 1), ("will_MD", 1), ("read_VB", 1), ("this_DT", 1), ("book_NN", 1)

Named Entity Recognition(NER)

- NER systems look for how a word is placed in a sentence and make use of other statistical models to identify what kind of word actually it is.

- San Pedro is a town on the southern part of the island of Ambergris Caye in the Belize District of the nation of Belize, in Central America.

- Here, the NER maps the words with the real world places. The places that actually exist in the physical world. We can automatically extract the real world places present in the document using NLP.

- San Pedro - Geographic Entity

- Ambergris Caye - Geographic Entity

- Belize - Geographic Entity

- Central America - Geographic Entity

- For example – 'Washington' can be a geographical location as well as the last name of any person. A good NER system can identify this.
- Kinds of objects that a typical NER system can tag:
  - People's names.
  - Company names.
  - Geographical locations
  - Product names.
  - Date and time.
  - Amount of money.
  - Events.

Coreference resolution:

- Coreference resolution (CR) is the task of finding all linguistic expressions (called mentions) in a given text that refer to the same real-world entity.

- After finding and grouping these mentions we can resolve them by replacing, pronouns with noun phrases.

"I voted for Trump because he was most aligned with my values", John said.

The original sentence

"John voted for Trump because Trump was most aligned with John's values", John said.

The sentence with resolved coreferences

Advantages of Natural Language Processing:

- Improves human-computer interaction: NLP enables computers to understand and respond to human languages, which improves the overall user experience and makes it easier for people to interact with computers.

- Automates repetitive tasks: NLP techniques can be used to automate repetitive tasks, such as text summarization, sentiment analysis, and language translation, which can save time and increase efficiency.

- Enables new applications: NLP enables the development of new applications, such as virtual assistants, chatbots, and question answering systems, that can improve customer service, provide information, and more.

- Improves decision-making: NLP techniques can be used to extract insights from large amounts of unstructured data, such as social media posts and customer feedback, which can improve decision-making in various industries.

- Improves accessibility: NLP can be used to make technology more accessible, such as by providing text-to-speech and speech-to-text capabilities for people with disabilities.

- Facilitates multilingual communication: NLP techniques can be used to translate and analyze text in different languages, which can facilitate communication between people who speak different languages.

- Improves information retrieval: NLP can be used to extract information from large amounts of data, such as search engine results, to improve information retrieval and provide more relevant results.

- Enables sentiment analysis: NLP techniques can be used to analyze the sentiment of text, such as social media posts and customer reviews, which can help businesses understand how customers feel about their products and services.

- Improves content creation: NLP can be used to generate content, such as automated article writing, which can save time and resources for businesses and content creators.

- Supports data analytics: NLP can be used to extract insights from text data, which can support data analytics and improve decision-making in various industries.

- Enhances natural language understanding: NLP research and development can lead to improved natural language understanding, which can benefit various industries and applications.

Disadvantages of Natural Language Processing:

- Limited understanding of context: NLP systems have a limited understanding of context, which can lead to misinterpretations or errors in the output.
- Requires large amounts of data: NLP systems require large amounts of data to train and improve their performance, which can be expensive and time-consuming to collect.
- Limited ability to understand idioms and sarcasm: NLP systems have a limited ability to understand idioms, sarcasm, and other forms of figurative language, which can lead to misinterpretations or errors in the output.
- Limited ability to understand emotions: NLP systems have a limited ability to understand emotions and tone of voice, which can lead to misinterpretations or errors in the output.

- Difficulty with multi-lingual processing: NLP systems may struggle to accurately process multiple languages, especially if they are vastly different in grammar or structure.

- Dependency on language resources: NLP systems heavily rely on language resources, such as dictionaries and corpora, which may not always be available or accurate for certain languages or domains.

- Difficulty with rare or ambiguous words: NLP systems may struggle to accurately process rare or ambiguous words, which can lead to errors in the output.

- Lack of creativity: NLP systems are limited to processing and generating output based on patterns and rules, and may lack the creativity and spontaneity of human language use.

- Ethical considerations: NLP systems may perpetuate biases and stereotypes, and there are ethical concerns around the use of NLP in areas such as surveillance and automated decision-making.

NLP Steps

- Preprocessing: Before applying NLP techniques, it is essential to preprocess the text data by cleaning, tokenizing, and normalizing it.

- Feature Extraction: Feature extraction is the process of representing the text data as a set of features that can be used in machine learning models.

- Word Embeddings: Word embeddings are a type of feature representation that captures the semantic meaning of words in a high-dimensional space.

- Neural Networks: Deep learning models, such as neural networks, have shown promising results in NLP tasks, such as language modeling, sentiment analysis, and machine translation.

- Evaluation Metrics: It is important to use appropriate evaluation metrics for NLP tasks, such as accuracy, precision, recall, F1 score, and perplexity.

Python Libraries For Natural Language Processing

- Natural Language Toolkit (NLTK)
- Gensim
- SpaCy
- CoreNLP
- TextBlob
- AllenNLP
- Polyglot
- Scikit-Learn
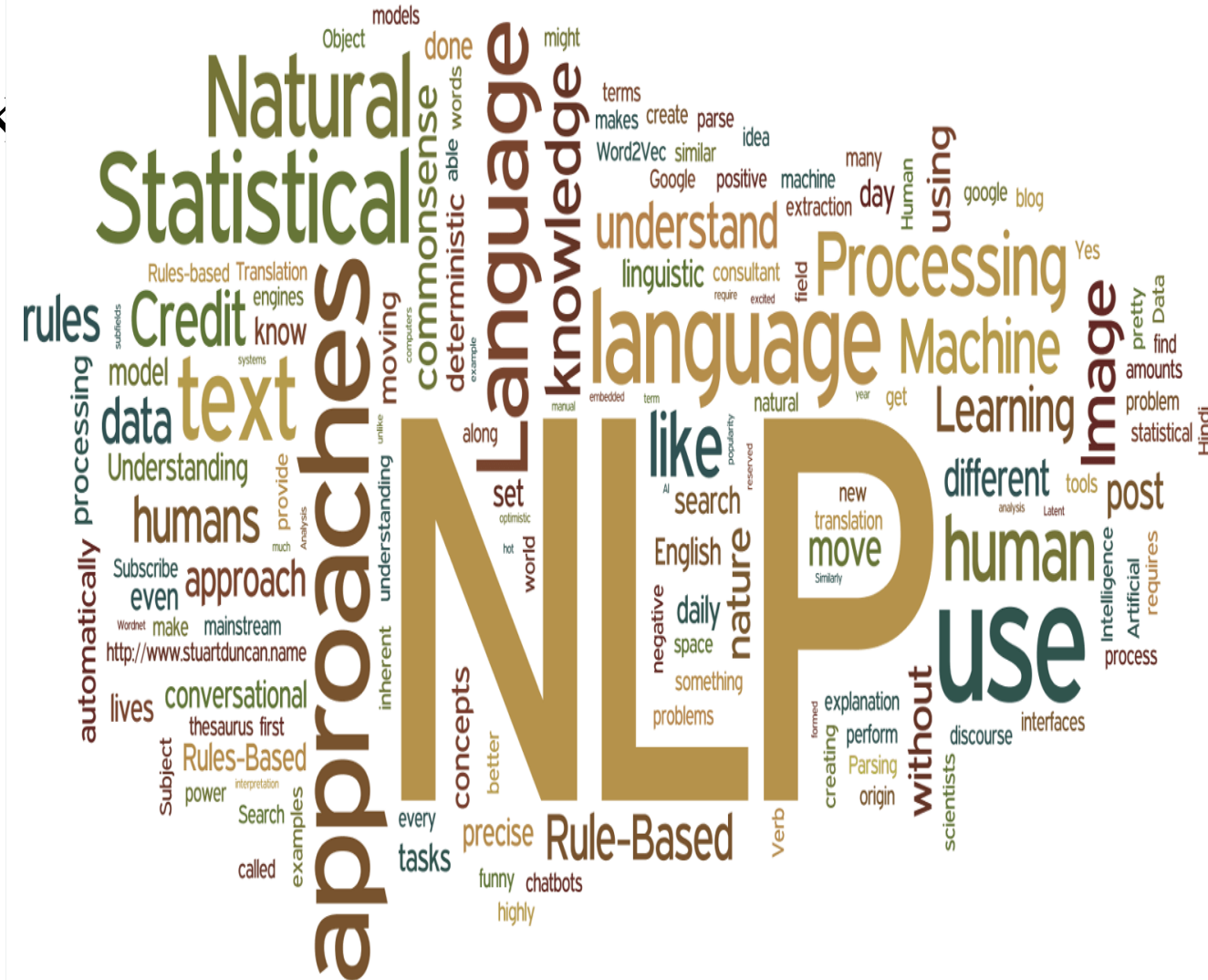
Natural Language Toolkit (NLTK)

- NLTK is the main library for building Python projects to work with human language data.

- NLTK is accessible for Windows, Mac OS, and Linux.

- The best part is that NLTK is a free, open-source library
  - Entity Extraction
  - Part-of-speech tagging
  - Tokenization
  - Parsing
  - Semantic reasoning
  - Stemming
  - Text classification

Gensim

- It is one of the best Python libraries for NLP tasks.
- It provides a special feature to identify semantic similarity between two documents using vector space modeling and the topic modeling toolkit.
- All algorithms in GenSim are memory-independent concerning corpus size, which means we can process input larger than RAM.
- It provides a set of algorithms that are very useful in natural language tasks such as the Hierarchical Dirichlet Process(HDP), Random Projections(RP), Latent Dirichlet Allocation(LDA), Latent Semantic Analysis(LSA/SVD/LSI) or word2vec deep learning.
- The most advanced feature of GenSim is its processing speed and fantastic memory usage optimization.

SpaCy

- SpaCy is one of the best open-source Python libraries for NLP. It is mainly designed for production usage- to build real-world projects and helps handle a large number of text data. Renowned for its rapidity and precision, SpaCy is a favored option for handling extensive datasets.

- It provides multi trained transformers like BERT.

- It is way faster than other libraries.

- Provides tokenization that is motivated linguistically In more than 49 languages.

- Provides functionalities such as text classification, sentence segmentation, lemmatization, part-of-speech tagging, named entity recognition, and many more.

- It has 55 trained pipelines in more than 17 languages.

CoreNLP

- Stanford CoreNLP contains a grouping of human language innovation instruments.

- It means to make the use of semantic analysis tools to a piece of text simple and proficient.

- With CoreNLP, you can extract a wide range of text properties (like part-of-speech tagging,named-entity recognition, and so forth) in a couple of lines of code.

TextBlob

- TextBlob is one of the famous Python libraries for NLP (Python 2 and Python 3) powered by NLTK.

- Constructed atop NLTK, it furnishes a streamlined API for typical Natural Language Processing (NLP) functions.

- It is the fastest NLP tool among all the libraries. It is beginners friendly.

- It provides an easy interface to help beginners and has all the basic NLP functionalities, such as sentiment analysis, phrase extraction, parsing, and many more. Some of the features of TextBlob are shown below:

- Sentiment analysis

- Parsing

- Word and phrase frequencies

- Part-of-speech tagging

- N-grams

- Spelling correction

- Tokenization

- Classification(Decision tree. Naïve Bayes)

- Noun phrase extraction

- WordNet integration

AllenNLP

- It is one of the most advanced Natural Language Processing Tools out there now.

- This is built on PyTorch tools and libraries.

- It is ideal for business and research applications. It has developed into an undeniable tool for various text-processing investigations.

- AllenNLP utilizes the SpaCy open-source library for data preprocessing while at the same time dealing with the lay cycles all alone.

- The fundamental component of AllenNLP is that it is easy to utilize. Unlike other NLP tools with numerous modules, AllenNLP simplifies the Natural Language Process.

Polyglot

- Following are the features of Polyglot:

  - Tokenization (165 Languages)
  - Language detection (196 Languages)
  - Named Entity Recognition (40 Languages)
  - Part of Speech Tagging (16 Languages)
  - Sentiment Analysis (136 Languages)
  - Word Embeddings (137 Languages)
  - Morphological analysis (135 Languages)
  - Transliteration (69 Languages)

Scikit-Learn

- It is one of the greater Python libraries for NLP and is most used among data scientists for NLP tasks.

- It provides a large number of algorithms to build machine learning models.

- It has excellent documentation that helps data scientists and makes learning easier.

- The main advantage of sci-kit learn is it has great intuitive class methods.

- It offers many functions for bag-of-words to convert text into numerical vectors. It has some disadvantages as well.

- It doesn't provide you with neural networks for text preprocessing.

- It is better to use other NLP libraries if you want to carry out more complex preprocessing, such as POS tagging for text corpora.

Application of NLP

- The goal is to recognize and classify items such as names of people, organizations, locations, and dates from a given text.
- Allowing important information to be extracted from unstructured data.
- Common datasets for NER include CoNLL-2003, OntoNotes, and Open Multilingual Wordnet.

Machine Translation

- NLP task that automatically translates text from one language to another, facilitating cross-lingual communication and accessibility.

- Popular datasets include WMT, IWSLT, and Multi30k. Data preprocessing involves tokenization, handling language-specific nuances, and generating the input-target pairs for training.

- Translate sentences or documents from the source language to the target language.

- Evaluate the translation quality using metrics like BLEU and METEOR.

Text Summarization

NLP Text Summarization — Extractive / Abstractive
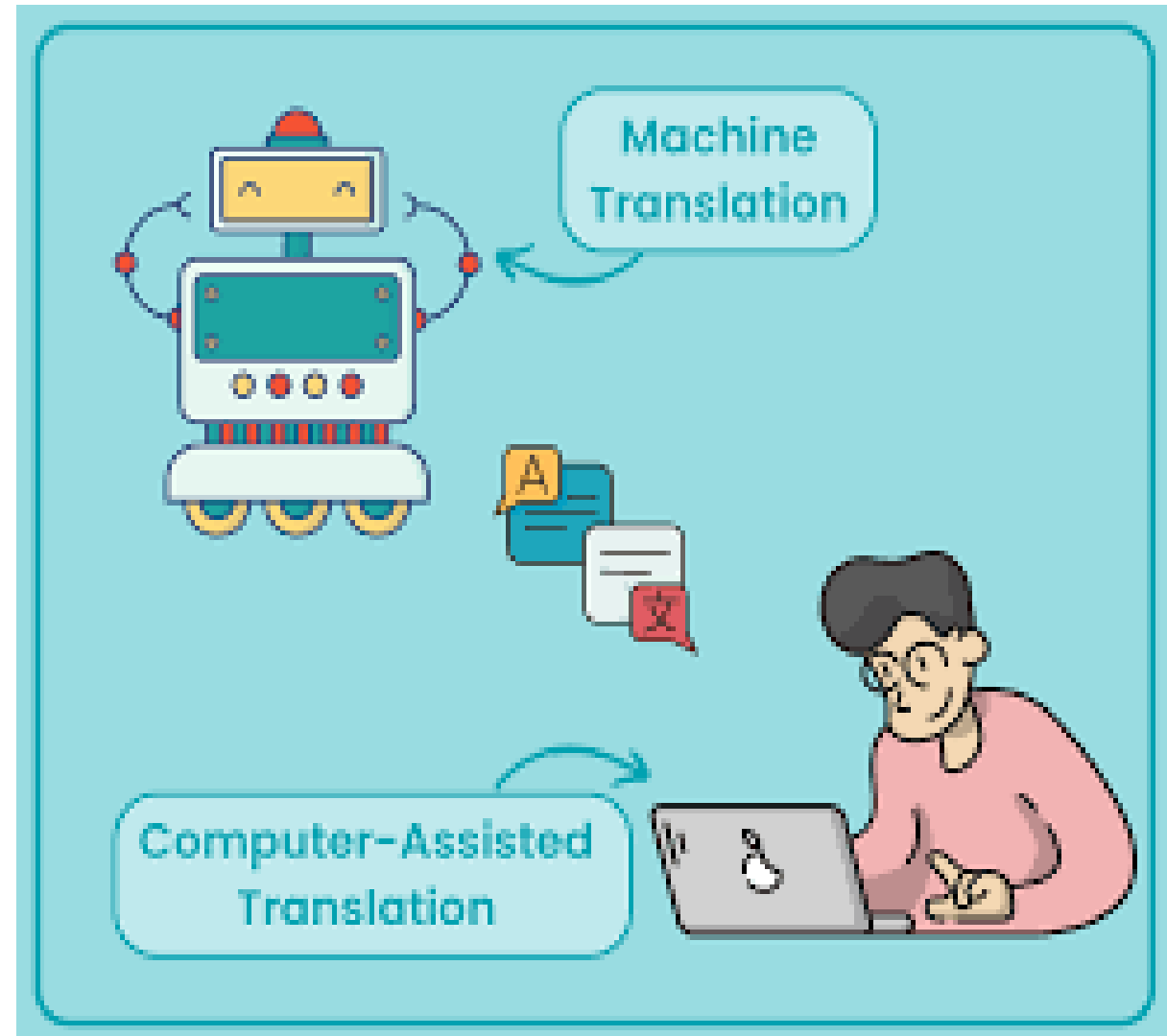
- involves generating concise and coherent summaries of longer pieces of text. It enables quick information retrieval and comprehension, making it invaluable for dealing with large volumes of textual data.

- Generate summaries for long articles or documents.

- Evaluate the quality of generated summaries using ROUGE and BLEU metrics.
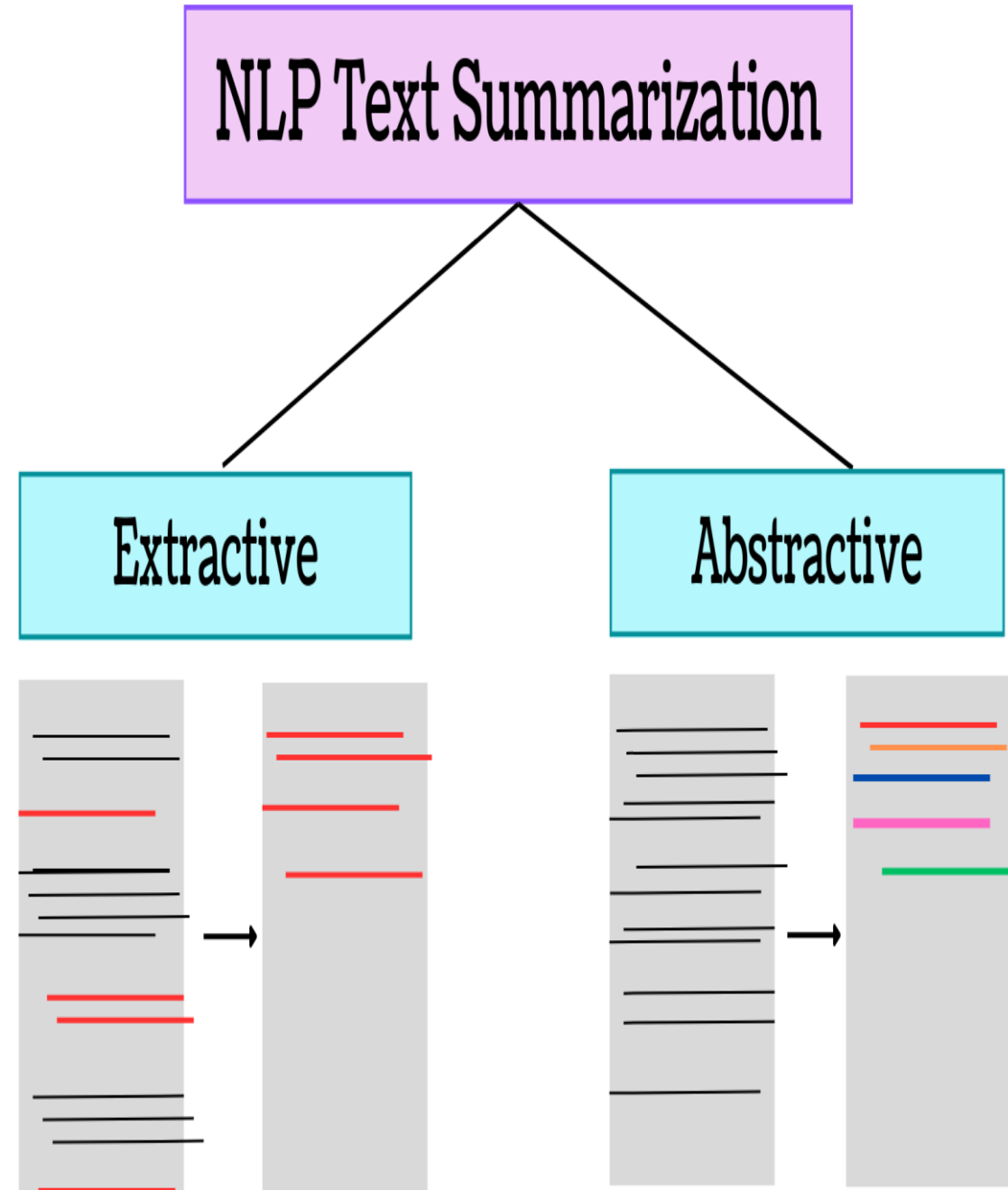
Text Correction and Spell Checking

- to develop algorithms that automatically correct spelling and grammatical errors in text data. It improves the accuracy and readability of written content.

- a dataset containing text with misspelled words and corresponding corrected versions. Data preprocessing involves handling capitalization, punctuation, and special characters.

- Detect and correct spelling errors in a given text.

- Suggest appropriate replacements for erroneous words based on context.

Phonetic spelling mistakes

Fizix is an interesting sudgekt.

Physics is an interesting subject.
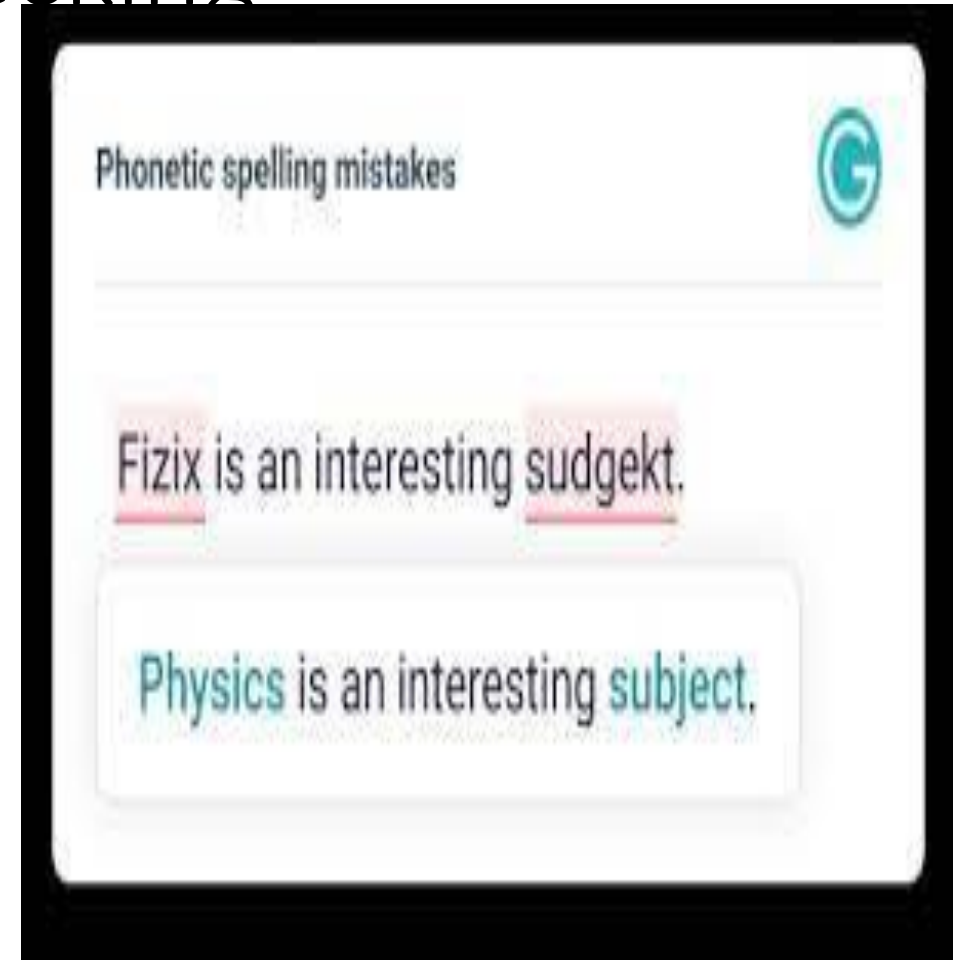
Sentiment Analysis

- It is NLP task that determines the sentiment expressed in a text, such as whether it is favorable, negative, or neutral. It is critical for analyzing client feedback, market attitudes, and social media monitoring.

- Analyze social media posts or product reviews to determine sentiment.

- Monitor changes in sentiment over time for specific products or topics.

SENTIMENT ANALYSIS

NEGATIVE   NEUTRAL   POSITIVE

Text Annotation and Data Labeling

- Text Annotation and Data Labeling are fundamental tasks in NLP projects, as they involve labeling text data for training supervised machine learning models. It is a crucial step to ensure the accuracy and quality of NLP models.

- Provide a platform for human annotators to label entities, sentiments, or other relevant information in the text.

- Ensure consistency and quality of annotations through validation and review mechanisms.

Deepfake Detection

- Deepfake technology has raised concerns regarding the authenticity and credibility of multimedia content, making Deepfake Detection a critical NLP task. Deepfakes are manipulated videos or audio that can deceive viewers into believing false information.

- Detects and classifies deepfake videos or audio.

- Evaluate the model's performance using precision, recall, and F1-score metrics.


ORIGINAL | DEEPFAKE

Voice Assistants for Smart Homes

- Voice Assistants have revolutionized smart home automation by enabling users to control various devices through natural language interactions. This technology enhances user experience and convenience.

- Create an intuitive voice assistant that understands and responds to voice commands.

- Integrate the voice assistant with smart home platforms for seamless device control.
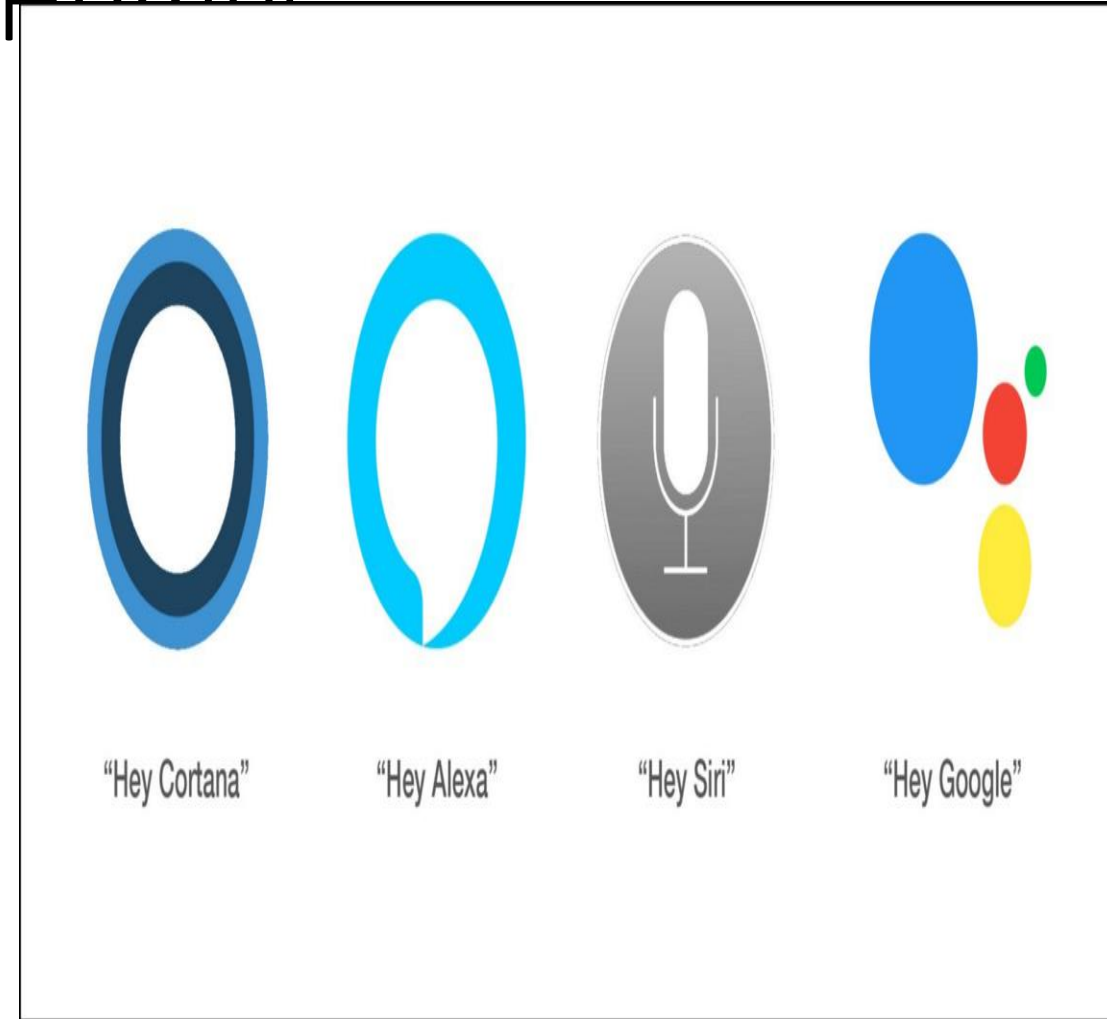
Creating Chatbots

- Creating Chatbots is a challenging NLP project that involves building highly sophisticated conversational agents capable of managing interactive and engaging user dialogues. Chatbots are exclusively used in customer service, virtual assistants, and various other applications.

- Develop a chatbot that understands user intents and provides contextually relevant responses.

- Evaluate the chatbot's performance through user satisfaction surveys and automated tests.
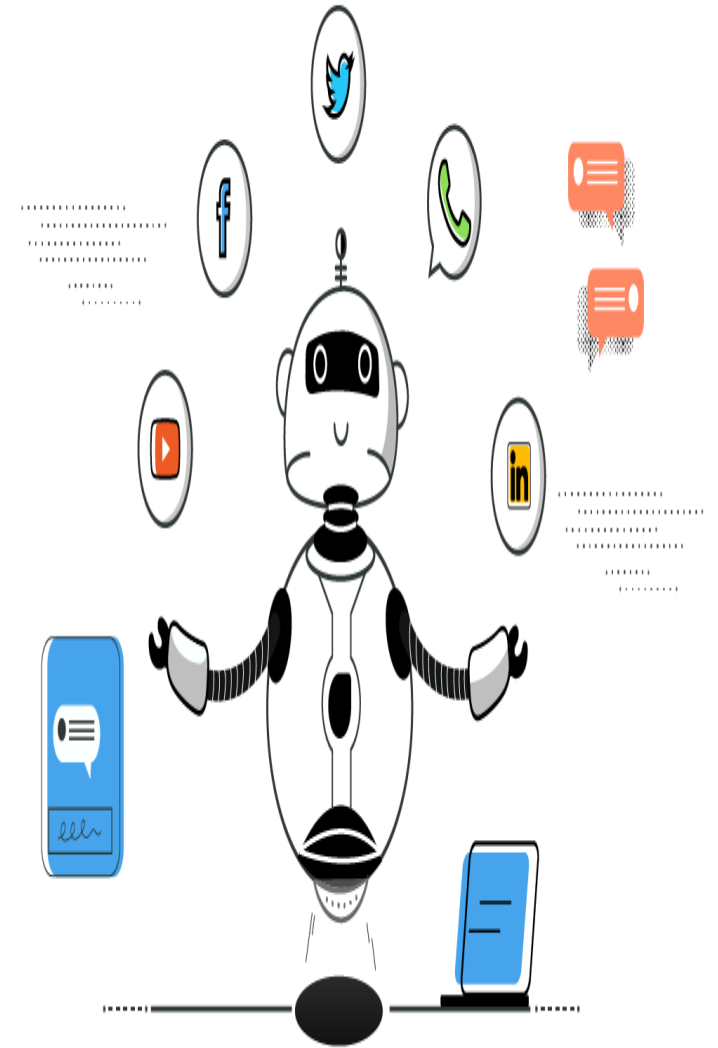
Text-to-Speech (TTS) and Speech-to-Text (STT)

- Text-to-Speech (TTS) and Speech-to-Text (STT) are significant components of Natural Language Processing, facilitating humans and machines to communicate effortlessly. The TTS generates written text in a human voice. In contrast, the STT converts spoken words into written text, creating a space to improve accessibility and seamless user interaction across various applications.

- Convert written text into human-like speech (TTS).

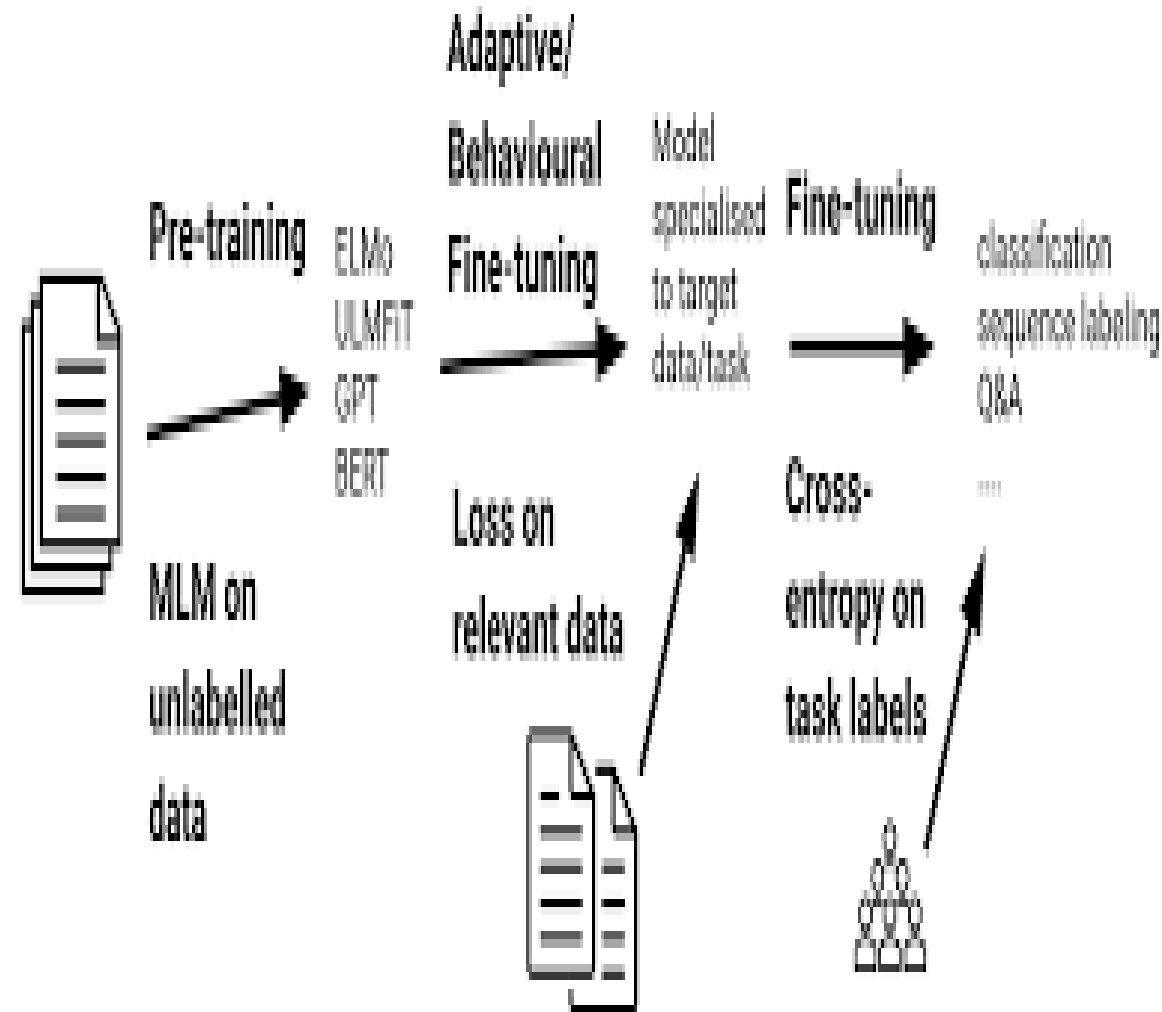- Transcribe spoken words into written text (STT) with high accuracy.
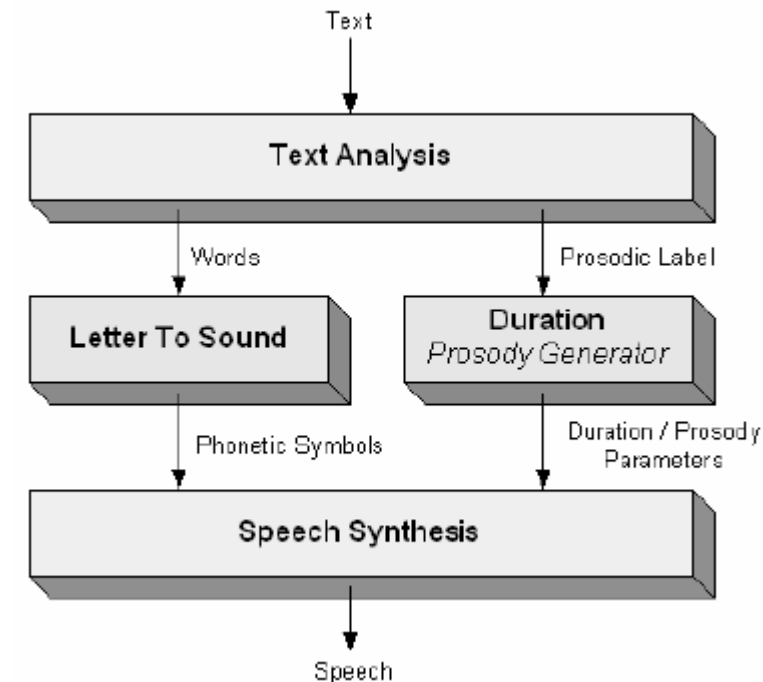
Emotion Detection

- Emotion Detection is a valuable NLP task that involves recognizing and understanding emotions conveyed through text. Its applications include sentiment analysis, customer service, and open human-computer interaction.

- Recognize emotions from spoken utterances.

- Evaluate the model's accuracy in emotion detection using metrics such as accuracy and confusion matrix.
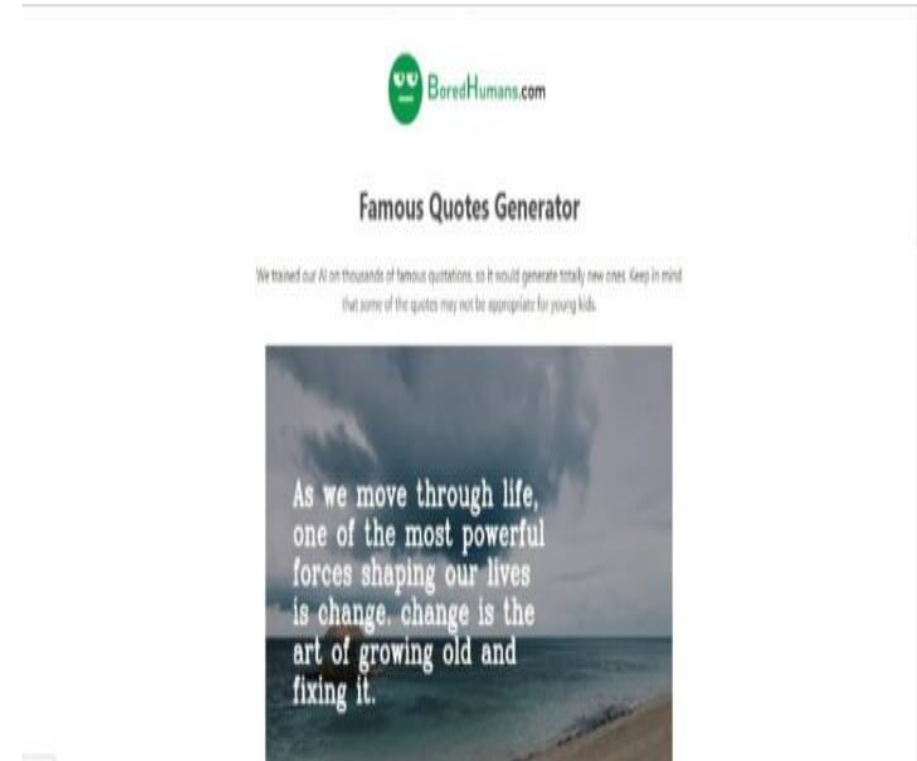

DEMO IMAGE
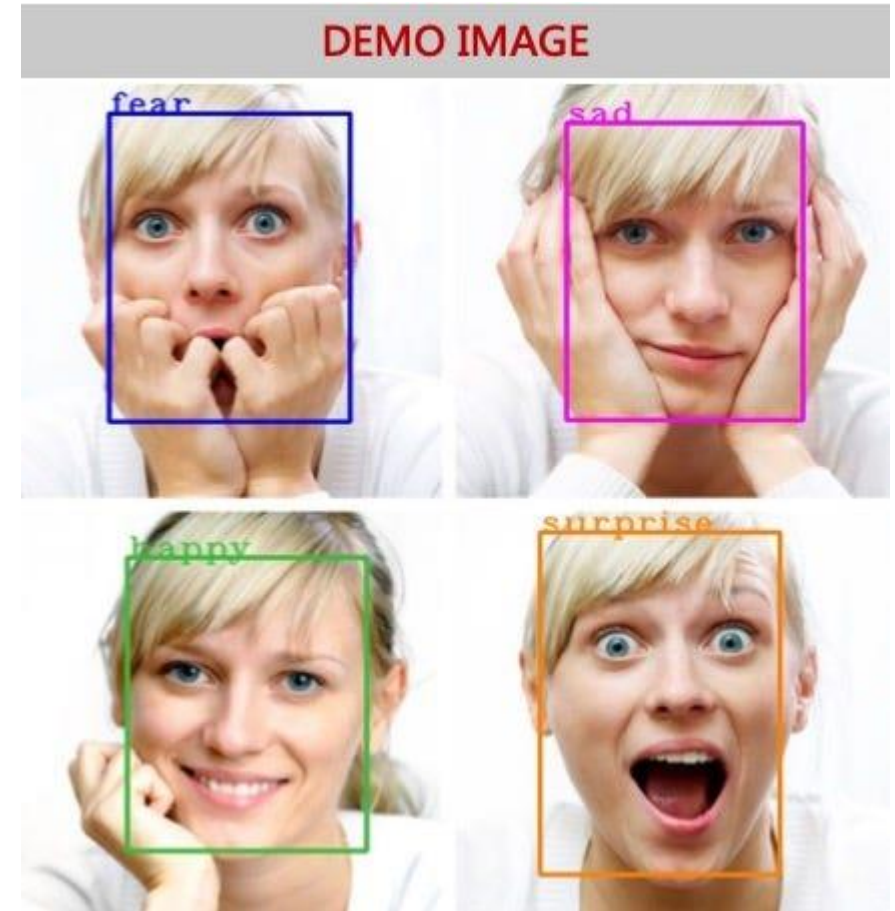
Language Model Fine-Tuning

- Language Model Fine-Tuning is a powerful technique in NLP that involves adapting pre-trained language models to perform specific tasks, enhancing model performance with limited labeled data.

- Fine-tune the pre-trained model on the target task.

- Evaluate the model's performance and compare it with the baseline model.

Pre-training   ELMo
                ULMFiT
                GPT
                BERT

MLM on
unlabelled
data

Adaptive/
Behavioural
Fine-tuning

Loss on
relevant data

Model
specialised
to target
data/task

Fine-tuning

Cross-
entropy on
task labels

classification
sequence labeling
Q&A
....

Inspiring Quote Generator

- The Inspiring Quote Generator is a creative NLP project that builds a model that generates motivational and uplifting quotes based on input keywords or themes.

- Generate inspiring quotes based on input keywords or themes.

- Evaluate the quality and coherence of generated quotes to ensure meaningful and motivational phrases.

- Thank you