NLP AND USE CASES

Advanced Natural Language Processing Models in Design, Development and NLP Use Cases

S. Tharabai

CIT, Chennai, India

Published November 2026

Abstract

Natural language processing, or NLP, is a field that focuses on studying the interaction between computers and human written language. The aim is to help machines understand and interpret human language helpfully. Many NLP use cases exist, such as sentiment analysis, text summarization, etc. However, there are so many things that NLP can do in a great way. They are Tokenization, Language Translation, Voice Assistance and Sentiment analysis. The PDF medical reports and treatment suggestions, Chat bot doctors pre-care can be integrated by the NLP. NLP models can parse structured and unstructured text from PDF medical reports lab results, prescriptions, discharge summaries. This involves PDF Extraction, Entity Recognition, Treatment suggestion mapping, Chatbot Integration.

Keywords: virtual culture chatbot; NLP applications; stemming; lemmatization; Natural Language Toolkit; sentiment analysis; tokenization; summarisation; types of NLP; NLP tasks; computation techniques of NLP; NLP challenges; medical NLP chatbots; NLP and NLU in medical analysis

1 Introduction

Recent advancements in neural networks, particularly transformer based architectures and large language models, have revolutionized tasks such as machine translation, sentiment analysis, question answering, text summarization, speech recognition, and conversational AI. These models can process vast amounts of linguistic data, learn contextual relationships between words, and generate human-like responses, making them valuable across various industries including healthcare, education, finance, customer service, and research.

1.1 Virtual culture chatbot

[1] A chatbot is an application that utilizes an NLP system to simulate an interaction via text to understand user input. Businesses usually use chatbots to automate redundant tasks such as standard customer support, appointment booking, and more. An unique application of an NLP powered chatbot has been used, and that is to become a Virtual Culture Chatbot that takes us on a walk into the museum. An example application for this is made by the Jewish Historical in Amsterdam, where they created a Culture Chatbot that explores the museum database, such as the collection and exhibition.

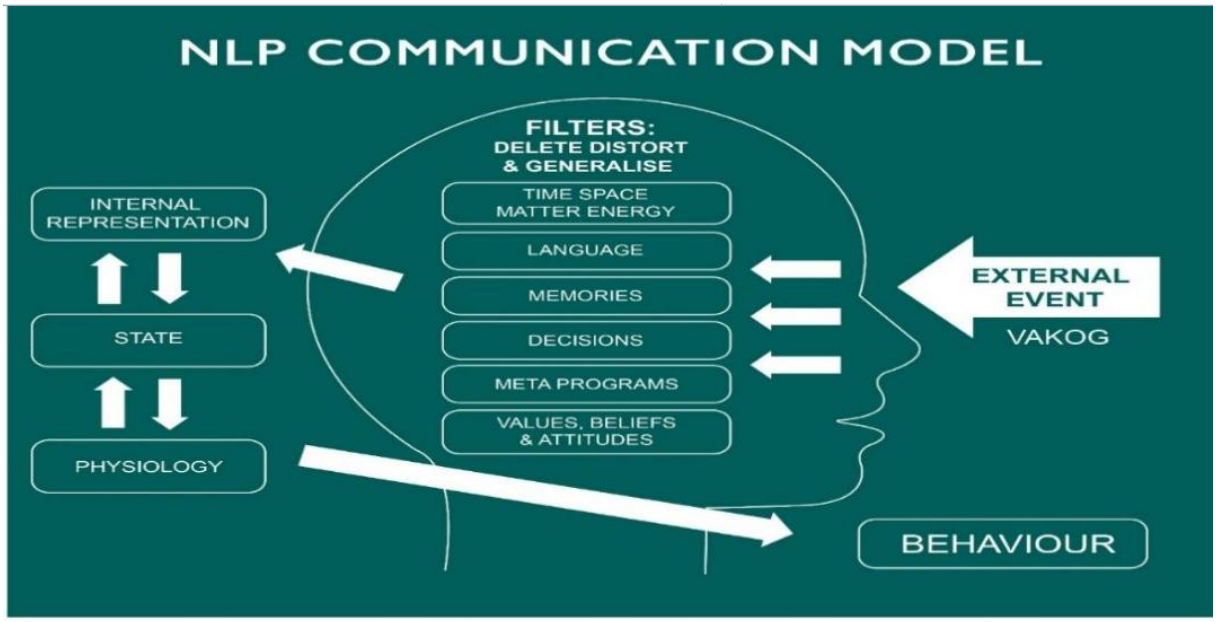


Figure 1. Natural language processing.

1.2 NLP applications

Natural Language Processing (NLP) has a wide range of applications, including:

- [1] **Voice Assistants**: Enabling devices like Alexa and Siri to interpret and respond to spoken commands.
- **Autocorrect And Spell Check**: Enhancing writing accuracy by correcting and suggesting spellings.
- **Chatbots**: Improving customer service by understanding and responding to human language.
- **Language Translation**: Converting text or speech from one language to another.
- **Sentiment Analysis**: Understanding the emotional tone behind textual data. The word "Love AI" earns positive score and it is a Good prompt. "Hate Coffee" earns negative score and it makes a Bad prompt.

1.3 Natural language processing (NLP) working principles

[1] NLP models work by finding relationships between the constituent parts of language for example, the letters, words, and sentences found in a text dataset. NLP architectures use various methods for data preprocessing, feature extraction and modelling. Some of these processes are:

- [1] **Data Preprocessing:** Before a model processes text for a specific task, the text often needs to be pre-processed to improve model performance or to turn words and characters into a format the model can understand. Data-centric AI is a growing movement that prioritizes data preprocessing. Various techniques may be used in this data preprocessing.
- [1] **Stemming And Lemmatization**: Stemming is an informal process of converting words to their base forms using heuristic rules. For example, "university," "universities," and "university's" might all be mapped to the base word **univers.** One limitation in this approach is that "universe" may also be mapped to **univers,** even though universe and university don't have a close semantic relationship. Lemmatization is a more formal way to find roots by analysing a word's morphology using vocabulary from a dictionary. Stemming and lemmatization are provided by libraries like **spaCy and NLTK.**

- [1] **Sentence Segmentation** breaks a large piece of text into linguistically meaningful sentence units. This is obvious in languages like English, where the end of a sentence is marked by a period, but it is still not trivial. A period can be used to mark an abbreviation as well as to terminate a sentence, and in this case, the period should be part of the abbreviation token itself. The process becomes even more complex in languages, such as ancient Chinese, that don't have a delimiter that marks the end of a sentence.

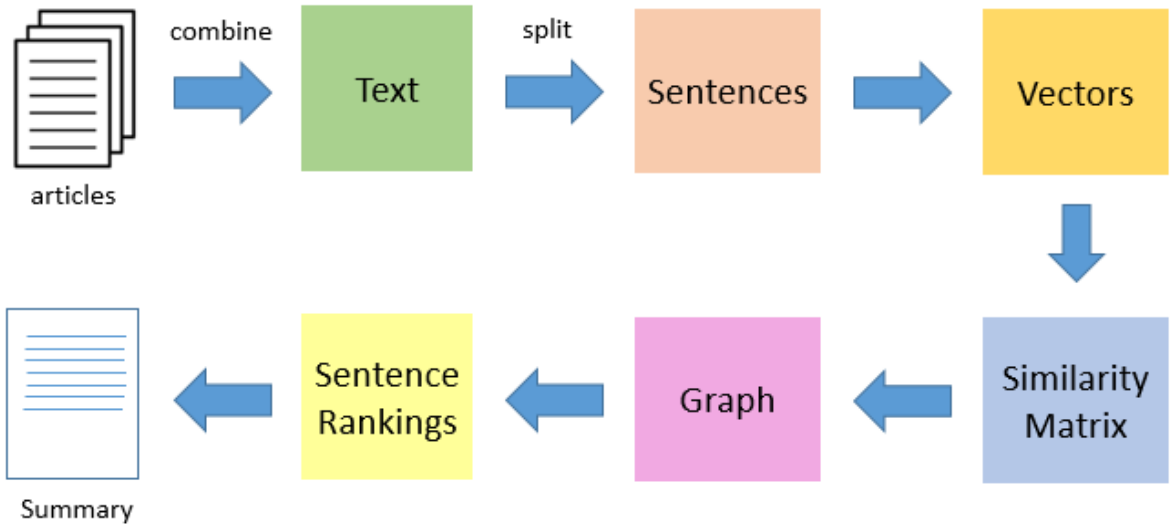


Figure 2. NLP summarization.

- **Stop Word Removal** aims to remove the most commonly occurring words that don't add much information to the text. For example, "the," "a," "an," and so on.
- [1] **Tokenization** splits text into individual words and word fragments. The result generally consists of a word index and tokenized text in which words may be represented as numerical tokens for use in various deep learning methods. A method that instructs language models to ignore unimportant tokens can improve efficiency.

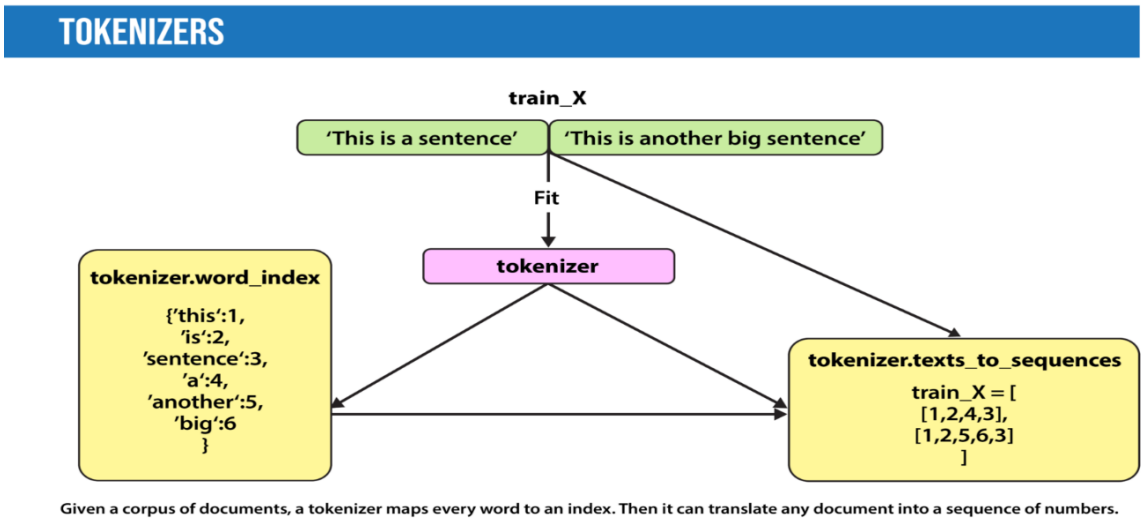


Figure 3. Tokenization.

- **Feature Extraction:** Most conventional machine learning techniques work on the features generally numbers that describe a document in relation to the corpus that contains. It created by either Bag-of-Words, TF-IDF, or generic feature engineering such as document length, word polarity, and metadata for instance, if the text has associated tags or scores. More recent techniques include Word2Vec, GLoVE, and learning the features during the training process of a neural network.
- **Bag-Of-Words:** Bag-of-Words counts the number of times each word or n-gram combination of n words appears in a document. For example, below, the Bag-of-Words model creates a numerical representation of the dataset based on how many of each word in the word_index occur in the document. Example Keywords count.

Many programming languages, libraries, and frameworks support natural language processing (NLP). Here are a few of the most useful languages,

- [1] **Python** is the most used programming language to tackle NLP tasks. Most libraries and frameworks for deep learning are written for Python. Here are a few that practitioners may find helpful.
- **Natural Language Toolkit (NLTK)** is one of the first NLP libraries written in Python. It provides easy-to-use interfaces to corpora and lexical resources such as WordNet. It also provides a suite of text processing libraries for classification, tagging, stemming, parsing, and semantic reasoning.
- **Deep Learning Libraries:** Popular deep learning libraries include TensorFlow and PyTorch, which make it easier to create models with features like automatic differentiation. These libraries are the most common tools for developing NLP models.

2 Approaches to NLP

2.1 Computational linguistics

[1] NLP combines the power of computational linguistics together with machine learning algorithms and deep learning. Computational linguistics uses data science to analyse language and speech. It includes two main types of analysis: syntactical analysis and semantical analysis. Syntactical analysis

determines the meaning of a word, phrase or sentence by parsing the syntax of the words and applying preprogrammed rules of grammar. Semantical analysis uses the syntactic output to draw meaning from the words and interpret their meaning within the sentence structure.

2.2 Speech recognition

[1] The parsing of words can take one of two forms. Dependency parsing looks at the relationships between words, such as identifying nouns and verbs, while constituency parsing then builds a parse tree or syntax tree. A rooted and ordered representation of the syntactic structure of the sentence or string of words. The resulting parse trees underlie the functions of language translators and speech recognition. Ideally, this analysis makes the output either text or speech understandable to both NLP models and people. The word "Hello", "Aullo" or "Hall--low" in every accent parsed and recognized as "Hello".

2.3 Self-supervised learning

[1] Self Supervised Learning (SSL) in particular is useful for supporting NLP because NLP requires large amounts of labeled data to train AI models. Because these labeled datasets require time-consuming annotation a process involving manual labeling by humans gathering sufficient data can be prohibitively difficult. Self supervised approaches can be more time effective and cost effective as they replace some or all manually labeled training data.

3 Types of NLP

There are three different approaches of NLP:

3.1 Rule-based NLP

[1] The earliest NLP applications were simple if-then decision trees, requiring preprogrammed rules. They are only able to provide answers in response to specific prompts, such as the original version of Moviefone, which had rudimentary natural language generation (NLG) capabilities. Because there is no machine learning or AI capability in rules based NLP, this function is highly limited and not scalable.

3.2 Statistical NLP

[1] Developed later, statistical NLP automatically extracts, classifies and labels elements of text and voice data and then assigns a statistical likelihood to each possible meaning of those elements. This relies on machine learning, enabling a sophisticated breakdown of linguistics such as part of speech tagging.

Statistical NLP introduced the essential technique of mapping language elements such as words and grammatical rules to a vector representation so that language can be modelled by using mathematical statistical methods, including regression or Markov models. This informed early NLP developments such as spellcheckers and T9 texting Text on 9 keys, to be used on Touch Tone telephones.

3.3 Deep learning NLP

[1] Recently, deep learning models have become the dominant mode of NLP, by using huge volumes of raw, unstructured data both text and voice to become ever more accurate. Deep learning can be

viewed as a further evolution of statistical NLP, with the difference that it uses neural network models. There are several subcategories of models:

- [1] **Sequence-To-Sequence (Seq2seq) Models**: Based on recurrent neural networks (RNN), they have mostly been used for machine translation by converting a phrase from one domain such as the German language into the phrase of another domain such as English.
- **Transformer Models:** They use tokenization of language the position of each token words or subwords and self attention capturing dependencies and relationships to calculate the relation of different language parts to one another. Transformer models can` be efficiently trained by using self supervised learning on massive text databases. A landmark in transformer models was Google's bidirectional encoder representations from transformers (BERT), which became and remains the basis of how Google's search engine works.
- **Autoregressive Models:** This type of transformer model is trained specifically to predict the next word in a sequence, which represents a huge leap forward in the ability to generate text. Examples of autoregressive LLMs include GPT, Llama, Claude and the open source Mistral.
- [1] **Foundation Models:** Prebuilt and curated foundation models can speed the launching of an NLP effort and boost trust in its operation. For example, the IBMGranite foundation models are widely applicable across industries. They support NLP tasks including content generation and insight extraction. Additionally, they facilitate Retrieval Augmented Generation (RAG), a framework for improving the quality of response by linking the model to external sources of knowledge. The models also perform named entity recognition which involves identifying and extracting key information in a text.

4 NLP tasks

Several NLP tasks typically help process human text and voice data in ways that help the computer make sense of what it's ingesting. Some of these tasks include:

- [2] Coreference resolution
- Prompt Engineering
- Word sense disambiguation

4.1 Coreference resolution

This is the task of identifying if and when two words refer to the same entity. The most common example is determining the person or object to which a certain pronoun refers, such as "she" = "Mary". But it can also identify a metaphor or an idiom in the text, such as an instance in which "bear" isn't an animal, but a large and hairy person.

4.2 Prompt engineering

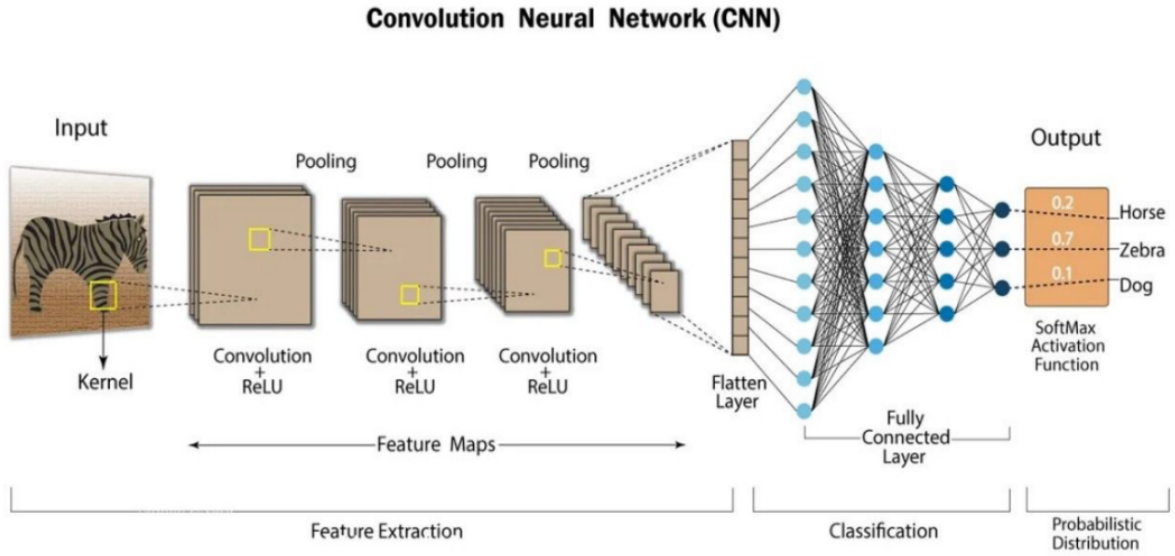


Figure 4. CNN for producing all possible outputs.

[2] Also called grammatical tagging, this is the process of determining which part of speech a word or piece of text is, based on its use and context. For example, part of speech identifies “make” as a verb in “I can make a paper plane,” and as a noun in “What make of car do you own”. Prompt engineering plays its role to do stemming work.Next word embedding is done by the LLM Multi layer Transformer. ANN, RNN , CNN and DNN are the Language Chain Models to empower the keywords.

4.3 Word sense disambiguation

[2] This is the selection of a word meaning for a word with multiple possible meanings. This uses a process of semantic analysis to examine the word in context. For example, word sense disambiguation helps distinguish the meaning of the verb “make” in “make the grade” to achieve versus “make a bet” to place.

5 Computational techniques of NLP

NLP works by combining various computational techniques to analyse, understand and generate human language in a way that machines can process.

5.1 Text preprocessing

[2] NLP text preprocessing prepares raw text for analysis by transforming it into a format that machines can more easily understand. It begins with tokenization, which involves splitting the text into smaller units like words, sentences or phrases. This helps break down complex text into manageable parts. Next, lowercasing is applied to standardize the text by converting all characters to lowercase, ensuring that words like "Apple" and "apple" are treated the same. Stop word removal is another common step, where frequently used words like "is" or "the" are filtered out because they don't add significant meaning to the text. Stemming or lemmatization reduces words to their root form Example: "running" becomes "run", making it easier to analyse language by grouping different forms of the same word.

Additionally, text cleaning removes unwanted elements such as punctuation, special characters and numbers that may clutter the analysis. After preprocessing, the text is clean, standardized and ready for machine learning models to interpret effectively.

5.2 Text generation

[3] Feature extraction is the process of converting raw text into numerical representations that machines can analyse and interpret. This involves transforming text into structured data by using NLP techniques like Bag of Words and TF-IDF, which quantify the presence and importance of words in a document.

5.3 Embedding the text

More advanced methods include word embeddings like Word2Vec or GloVe, which represent words as dense vectors in a continuous space, capturing semantic relationships between words. Contextual embeddings further enhance this by considering the context in which words appear, allowing for richer, more nuanced representations.

5.4 Text analysis

[3] Text analysis involves interpreting and extracting meaningful information from text data through various computational techniques. This process includes tasks such as part of speech (POS) tagging, which identifies grammatical roles of words and named entity recognition (NER), which detects specific entities like names, locations and dates.

[3] Dependency parsing analyses grammatical relationships between words to understand sentence structure, while sentiment analysis determines the emotional tone of the text, assessing whether it is positive, negative or neutral. Topic modelling identifies underlying themes or topics within a text or across a corpus of documents. Natural Language Understanding (NLU) is a subset of NLP that focuses on analysing the meaning behind sentences. NLU enables software to find similar meanings in different sentences or to process words that have different meanings. Through these techniques, NLP text analysis transforms unstructured text into insights.

5.5 Model training

[3] Processed data is then used to train machine learning models, which learn patterns and relationships within the data. During training, the model adjusts its parameters to minimize errors and improve its performance. Once trained, the model can be used to make predictions or generate outputs on new, unseen data. The effectiveness of NLP modelling is continually refined through evaluation, validation and fine tuning to enhance accuracy and relevance in real world applications.

[3] Different software environments are useful throughout the said processes. For example, the Natural Language Toolkit (NLTK) is a suite of libraries and programs for English that is written in the Python programming language. It supports text classification, tokenization, stemming, tagging, parsing and semantic reasoning functionalities. TensorFlow is a free and open source software library for machine learning and AI that can be used to train models for NLP applications. Tutorials and certifications abound for those interested in familiarizing themselves with such tools.

6 Challenges of NLP

[3] Even state of the art NLP models are not perfect, as human speech is prone to error. As with any AI technology, NLP comes with potential pitfalls. Human language is filled with ambiguities that make it difficult for programmers to write software that accurately determines the intended meaning of text or voice data. Human language might take years for humans to learn and many never stop learning. But then programmers must teach natural language powered applications to recognize and understand irregularities so their applications can be accurate and useful. Associated risks might include:

6.1 Misinterpretation

[3] As in programming, there is a risk of garbage in, garbage out (GIGO). Speech recognition, also known as speech to text, is the task of reliably converting voice data into text data. But NLP solutions can become confused if spoken input is in an obscure dialect, mumbled, too full of slang, homonyms, incorrect grammar, idioms, fragments, mispronunciations, contractions or recorded with too much background noise.

6.2 New vocabulary

[3] New words are continually being invented or imported. The conventions of grammar can evolve or be intentionally broken. In these cases, NLP can either make a best guess or admit it is unsure and either way, this creates a complication.

6.3 Tone of voice

[3] When people speak, their verbal delivery or even body language can give an entirely different meaning than the words alone. Exaggeration for effect, stressing words for importance or sarcasm can be confused by NLP, making the semantic analysis more difficult and less reliable. NLP use cases by industry. NLP applications can now be found across virtually every industry. Example: “Hello” or “Halloo”.

6.4 Finance

[3] In financial dealings, nanoseconds might make the difference between success and failure when accessing data, or making trades or deals. NLP can speed the mining of information from financial statements, annual and regulatory reports, news releases or even social media.

6.5 Healthcare

[4] New medical insights and breakthroughs can arrive faster than many healthcare professionals can keep up. NLP and AI based tools can help speed the analysis of health records and making better informed medical decisions possible, or assisting in the detection or even prevention of medical conditions.



Figure 5. Health care chatbot life cycle.

6.5.1 Medical NLP chatbots

6.5.1.1 Custom NLU for healthcare chatbots

- Developed a text-message-based chatbot for patient interaction.
- Used custom natural language understanding (NLU) tailored to medical terminology.
- Improved accuracy in identifying patient intent symptom reporting, medication queries.
- Result: Faster triage and reduced burden on healthcare staff.

6.5.1.2 LLM chatbot for care transitions

- [4] Conducted a randomized controlled trial with patients moving from primary to specialist care.
- Chatbot built on large language models (LLMs) guided patients through referrals, tests, and follow-ups.
- Outcome: Improved patient satisfaction and reduced delays in specialist consultations.
- Highlight: Demonstrated real-world safety and effectiveness of LLMs in healthcare.

6.5.1.3 Smart healthcare chatbot – CureAi

- [4] Uses sentence transformers + Llama-based LLM for contextual understanding.
- Classifies symptoms into domains diabetes, hypertension, skin, gastrointestinal.
- Integrates Gemini API for dynamic questioning and medical history recording.
- Provides root cause analysis with fishbone diagrams and ishikawa to visualize symptom causes.
- Outcome: Efficient preliminary medical advice and improved patient interviews.

6.5.2 Tools required

- [4] Accuracy boost: Custom NLU and LLMs outperform generic chatbots in medical contexts.
- Workflow integration: Chatbots are being tied into hospital systems for referrals and lab reports.
- Visualization tools: CureAi's fishbone diagrams show how lifestyle, genetics, and stress contribute to conditions.
- Patient empowerment: These systems provide pre-care guidance like diet, exercise, reminders before doctor visits.

6.5.3 Interpretation

- [4] Data privacy: Must comply with HIPAA/GDPR/Indian Data Protection Bill.
- Bias in models: LLMs may misinterpret rare conditions if training data is limited.
- Clinical validation: Chatbots must be tested in controlled trials before deployment.

6.6 Natural language processing and natural language understanding (NLP and NLU)

NLP (Natural Language Processing) can be a powerful bridge between PDF medical reports, treatment suggestions, and chatbot-based pre-care systems. This involves PDF Extraction, Entity Recognition, Treatment suggestion mapping, Chatbot Integration.

- PDF Extraction NLP models can parse structured and unstructured text from PDF medical reports like lab results, prescriptions, discharge summaries. This involves OCR for scanned docs and semantic parsing.
- Entity Recognition Key medical entities such as symptoms, diagnoses, medications, dosages, lab values) are identified and mapped to standardized vocabularies like SNOMED CT or ICD-10.
- Treatment Suggestion Mapping NLP can align doctor's recommendations with evidence-based guidelines, flagging contradictions or highlighting missing pre-care steps.
- Chatbot Integration

[4] A chatbot can use this structured data to provide pre-care guidance like:

- Reminders for medication schedules
- Lifestyle advice on diet, exercise, sleep etc.
- Alerts for abnormal lab values
- Prescribed drugs or procedures
- Predictive Analytics By analyzing historical patient data, NLP driven chatbots can suggest preventive care measures example: early diabetes management if blood sugar trends are rising.

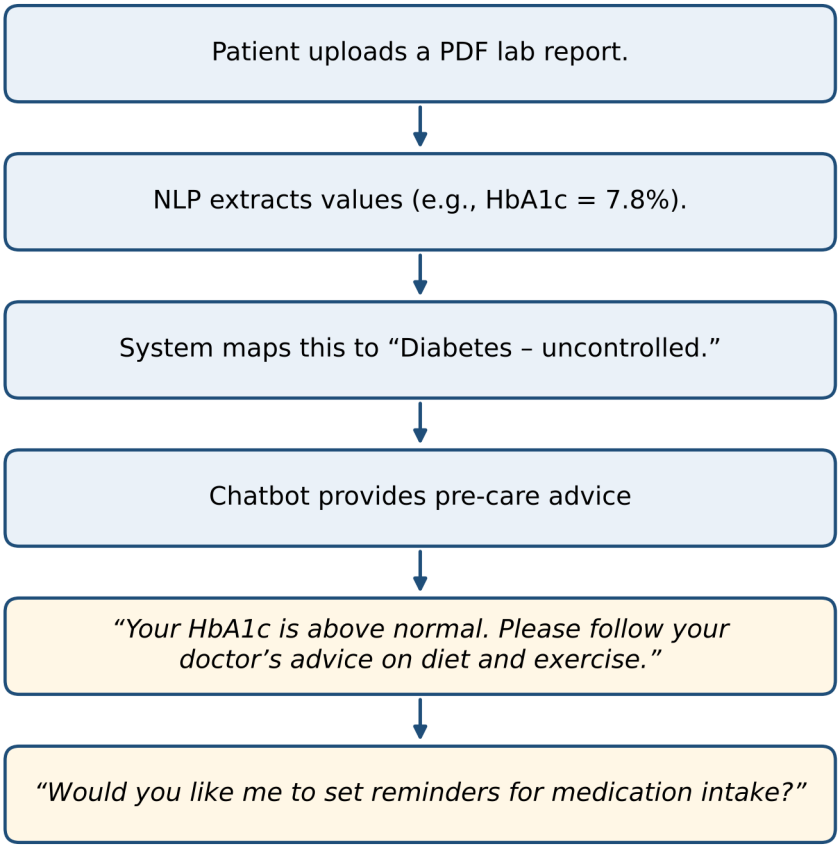


Figure 6. NLP and NLU workflow.

6.6.1 Advantages

- [4] Efficiency: Doctors save time; patients get instant guidance.
- Personalization: Chatbots tailor advice to patient history.
- Accessibility: Patients in remote areas receive continuous support.

6.7 Insurance

NLP can analyse claims to look for patterns that can identify areas of concern and find inefficiencies in claims processing, leading to greater optimization of processing and employee efforts.

7 Conclusion

NLP is solving a problem of every search and recommendation system faces. The goal is to retrieve, from a collection of documents that are indexed and numbered in millions. And Match the set of documents that are most relevant to the query. Document retrieval system extracts the matched contents. Thus NLP provides sophisticated retrieval mechanism. Even the PDF medical reports and treatment suggestions, Chat bot doctors can be integrated by the NLP to start and assist in patients care.

References

[1] Microsoft Bing. *Applications of NLP* [search results] [Internet]. https://www.bing.com/search?q=Applications+of+NLP

[2] DeepLearning.AI. *Natural language processing* [Internet]. https://www.deeplearning.ai/resources/natural-language-processing/

[3] IBM. *Natural language processing* [Internet]. https://www.ibm.com/think/topics/natural-language-processing

[4] Parseur. *AI data extraction* [Internet]. https://parseur.com/blog/ai-data-extraction