HAIMLC701 AI & ML in Healthcare

4.0		**Natural Language Processing in Healthcare**	**08**
	4.1	NLP tasks in Medicine, Low-level NLP components, High level NLP components, NLP Methods.	
	4.2	Clinical NLP resources and Tools, NLP Applications in Healthcare. Model Interpretability using Explainable AI for NLP applications.	

NLP

- **Natural Language Processing**-part of **Computer Science, Human language,** and **Artificial Intelligence**
- Technology that is used by machines to understand, analyze, manipulate, and interpret human's languages
- helps developers to organize knowledge for performing tasks such as
  - **translation**
  - **automatic summarization**
  - **Named Entity Recognition (NER)**
  - **speech recognition**
  - **relationship extraction**
  - **topic segmentation**

Components of NLP

- **Natural Language Understanding (NLU)**
  - helps the machine to understand and analyse human language by extracting the metadata from content such as concepts, entities, keywords, emotion, relations, and semantic roles
  - used in Business applications to understand the customer's problem in both spoken and written language
- NLU involves the following tasks -
  - used to map the given input into useful representation
  - used to analyze different aspects of the language

Natural Language Generation (NLG)

- Acts as a translator that converts the computerized data into natural language representation
- Involves Text planning, Sentence planning, and Text Realization

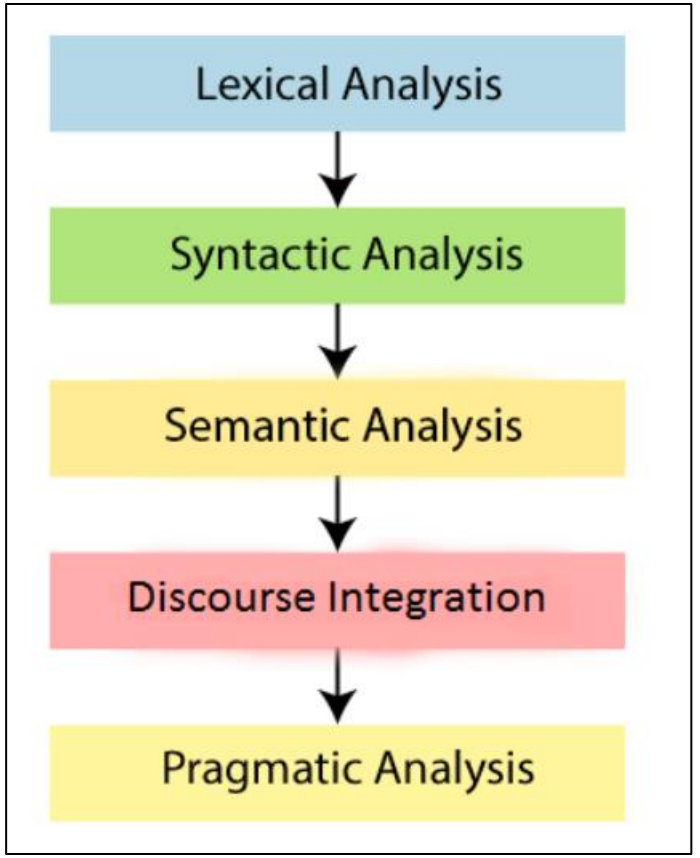
Phases of NLP

1. Lexical Analysis and Morphological

- This phase scans the source code as a stream of characters and converts it into meaningful lexemes
- A lexicon is defined as a collection of words and phrases in a given language
- It divides the whole text into paragraphs, sentences, and words

2. Syntactic Analysis (Parsing)

- used to check grammar, word arrangements, and shows the relationship among the words
- **Example:** Agra goes to the Poonam
- In the real world, Agra goes to the Poonam, does not make any sense, so this sentence is rejected by the Syntactic analyzer



Phases of NLP

3. Semantic Analysis

- concerned with the meaning representation.
- mainly focuses on the literal meaning of words, phrases, and sentences
- The guava ate an apple

4. Discourse Integration

- Depends upon the sentences that preceeds it and also invokes the meaning of the sentences that follow it
- Billy bought it.

5. Pragmatic Analysis

- helps to discover the intended effect by applying a set of rules that characterize cooperative dialogues
- It comprehends how people communicate with one another, the context in which they converse, and a variety of other factors. It refers to the process of abstracting or extracting the meaning of a situation's use of language.
- **For Example:** "Open the door" is interpreted as a request instead of an order

How to build an NLP pipeline

Step1: Sentence Segmentation

- It breaks the paragraph into separate sentences

Example: Consider the following paragraph -

- **Independence Day is one of the important festivals for every Indian citizen. It is celebrated on the 15th of August each year ever since India got independence from the British rule. The day celebrates independence in the true sense.**
- **Sentence Segment produces the following result:**

1. "Independence Day is one of the important festivals for every Indian citizen."
2. "It is celebrated on the 15th of August each year ever since India got independence from the British rule."
3. "This day celebrates independence in the true sense."

How to build an NLP pipeline

Step2: Word Tokenization

- used to break the sentence into separate words or tokens

Example:

- XYZ offers Corporate Training, Summer Training, Online Training, and Winter Training
- Word Tokenizer generates the following result:
  - “XYZ", "offers", "Corporate", "Training", "Summer", "Training", "Online", "Training", "and", "Winter", "Training", "."

How to build an NLP pipeline

Step3: Stemming

- used to normalize words into its base form or root form
  - example, celebrates, celebrated and celebrating, all these words are originated with a single root word "celebrate.“
- sometimes it produces the root word which may not have any meaning
  - **For Example,** intelligence, intelligent, and intelligently, all these words are originated with a single root word "intelligen."
  - In English, the word "intelligen" do not have any meaning

How to build an NLP pipeline

Step 4: Lemmatization

- Quite similar to Stemming used to group different inflected forms of the word, called Lemma
- The main difference between Stemming and lemmatization is that it produces the root word, which has a meaning
  - **For example:** In lemmatization, the words intelligence, intelligent, and intelligently has a root word intelligent, which has a meaning

How to build an NLP pipeline

Step 5: Identifying Stop Words

- In English, there are a lot of words that appear very frequently like "is", "and", "the", and "a".
- NLP pipelines will flag these words as stop words.
- **Stop words** might be filtered out before doing any statistical analysis.
  - **Example:** He **is a** good boy.

How to build an NLP pipeline

Step 6: Dependency Parsing

- used to find how all the words in the sentence are related to each other

Step 7: POS tags

- stands for parts of speech, which includes Noun, verb, adverb, and Adjective
- indicates how a word functions with its meaning as well as grammatically within the sentences
- A word has one or more parts of speech based on the context in which it is used
- **Example: "Google"** something on the Internet.
  - In the above example, Google is used as a verb, although it is a proper noun

How to build an NLP pipeline

Step 8: Named Entity Recognition (NER)

- process of detecting the named entity such as person name, movie name, organization name, or location
- **Example: Steve Jobs** introduced iPhone at the Macworld Conference in San Francisco, California

Step 9: Chunking

- used to collect the individual piece of information and grouping them into bigger pieces of sentences

Challenges in NLP

- **Ambiguity**
  - **Lexical Ambiguity**
    - Lexical Ambiguity exists in the presence of two or more possible meanings of the sentence within a single word.
  - **Example:**
    - Manya is looking for a **match**.
  - In the above example, the word match refers to that either Manya is looking for a partner or Manya is looking for a match. (Cricket or other match)

Challenges in NLP

- **Syntactic Ambiguity**
  - exists in the presence of two or more possible meanings within the sentence
  - **Example:**
    - I saw the girl with the binocular.
    - In the above example, did I have the binoculars? Or did the girl have the binoculars?
- **Referential Ambiguity**
  - exists when we are referring to something using the pronoun
  - **Example:** Kiran went to Sunita. She said, "I am hungry."
    - In the above sentence, we do not know that who is hungry, either Kiran or Sunita

NLP Libraries

- **Scikit-learn:**
  - provides a wide range of algorithms for building machine learning models in Python
- **Natural language Toolkit (NLTK):**
  - NLTK is a complete toolkit for all NLP techniques
- **Pattern:**
  - a web mining module for NLP and machine learning
- **TextBlob:**
  - provides an easy interface to learn basic NLP tasks like sentiment analysis, noun phrase extraction, or pos-tagging
- **Quepy:**
  - used to transform natural language questions into queries in a database query language
- **SpaCy:**
  - an open-source NLP library which is used for Data Extraction, Data Analysis, Sentiment Analysis, and Text Summarization
- **Gensim:**
  - Gensim works with large datasets and processes data streams

NLP Tasks

- Three broad categories in order of increasing difficulty and complexity:
- pre-processing, low-level, and high-level tasks

Pre-Processing Tasks

Books are on the table.

1. **Tokenization**: splits the text into smaller pieces, such as sentences

   ['Books', 'are', 'on', 'the', 'table']

2. **Normalization**: converts the text to the same format, such as converting uppercase to lowercase

   books are on the table

3. **Noise Removal**: removes invalid or non-text characters, such as punctuation
4. **Stop word Removal**: removes very common words with little meaning, such as “the”, “over”, or “for” (note that this is sometimes not desirable for NLP problems where such words are meaningful)
5. **Lemmatization**: converts words to their base forms, such as “jumps” to “jump”

Low-Level Tasks – Keyword Extraction

- After pre-processing, simple rule-based methods (or even statistical methods sometimes) can be applied to the extracted keywords from the text

1. **Part-Of-Speech (POS) tagging**: labels each word by its grammatical role (i.e. part-of-speech)
    1. e.g. “And now for something completely different.” → {“And”: Coordinating conjunction, “now”: Adverb, “for”: Preposition, “something”: Singular noun, “completely”: Adverb, “different”: Adjective}
    2. **Applications**: usually a stepping stone for other, more advanced tasks (unless your goal is analyzing grammatical functions in a text)

Low-Level Tasks – Keyword Extraction

1. **Named Entity Recognition (NER)**: locates and classifies named entities (e.g. persons, locations, organizations) into pre-defined categories
    1. e.g. “Jim bought 300 shares of Acme Corp. in 2006.” → {“Jim”: Person, “Acme Corp.”: Corporation, “2006”: Time}
    2. **Applications**: extracting events from a text, building a search engine, building a database of entities in a large collection of texts

Low-Level Tasks – Keyword Extraction

1. **Relationship Extraction / Event Extraction: extract** subject-verb-object triplets, actor-action pairs, or actor-object pairs from a text
    1. e.g. “Paris is the capital of France” → {Country: “France”, Capital city: “Paris”}
    2. e.g. “The United States won the most gold medals in the 2016 Olympics.” → {Subject: “United States”, Verb: “won”, Object: “gold medals”}
    3. **Applications**: populating a database (relational or non-relational) with a fixed schema (e.g. subject-verb-object, actor-action, actor-object) of company acquisitions, business news

Low-Level Tasks – Keyword Extraction

1. **Document Classification / Clustering**: label each document in a collection with a pre-defined class (classification) or cluster similar documents together without pre-defined classes (clustering)

Applications: automatically tagging or organizing a large collection of texts for routing to the correct employee, facilitating document search (e.g. in a library)

High-Level Tasks – Content Analysis

1. **Sentiment Analysis**: quantifies the subjective sentiment of a text on a scale from 0 (most negative) to 1 (most positive)
    1. e.g. “I love Alphabyte Solutions.” → 0.92 sentiment (very positive)
    2. **Applications**: quantifying user responses to a given product where there are many text reviews and few or unreliable numerical ratings, analyzing a social media feed

High-Level Tasks – Content Analysis

1. **Recommender Systems**: analyzes the content of an unstructured text to provide a content-based recommendation
    1. e.g. A user reads a book on the political system in Canada → the system recommends a book on the political system in the United States
    2. **Applications**: improving customer engagement of text-based products by recommending relevant products similar to query texts

High-Level Tasks – Content Analysis

1. **Natural Language Understanding**: extracts the semantic meaning from a text (e.g. query) beyond keyword matching and (optionally) converts it into a database query
    1. e.g. “Which pandemics in the last 100 years had a greater death toll than COVID-19?” → “”SELECT Pandemics by death toll WHERE start date is between 1920 and 2020 AND death toll > COVID-19 GROUP BY Pandemic name SORT BY death toll DESC”
    2. **Applications**: building search engines that can take natural language queries, enabling non-technical users to easily query a database (e.g. Advanced Search on Google)

High-Level Tasks – Content Analysis

1. **Information Retrieval / Semantic Search / Question Answering**: given a text query, retrieves content-relevant documents or returns a natural language answer
    1. e.g. “Which pandemics in the last 100 years had a greater death toll than COVID-19?” → “Spanish flu, Russian typhus, Asian flu, etc. all had the same death count as COVID-19 in the last 100 years.”
    2. **Applications**: building systems that can answer user questions in natural language (e.g. chatbots) or create a user-interactive FAQ tool

NLP tasks in Medicine

Clinical Documentation

- helps free clinicians from the laborious physical systems of EHRs and permits them to invest more time in the patient
- Both speech-to-text dictation and formulated data entry have been a blessing

NLP tasks in Medicine

Speech Recognition

- Important use case in speech recognition over the years by allowing clinicians to transcribe notes for useful EHR data entry
- Front-end speech recognition eliminates the task of physicians to dictate notes instead of having to sit at a point of care, while back-end technology works to detect and correct any errors in the transcription before passing it on for human proofing

Computer-Assisted Coding (CAC)

- CAC captures data of procedures and treatments to grasp each possible code to maximise claims
- Most popular uses of NLP, but unfortunately, its adoption rate is just 30%
- It has enriched the speed of coding but fell short at accuracy

NLP tasks in Medicine

Data Mining Research

- The integration of data mining in healthcare systems allows organizations to reduce the levels of subjectivity in decision-making and provide useful medical know-how
- Once started, data mining can become a cyclic technology for knowledge discovery, which can help any HCO create a good business strategy to deliver better care to patients

NLP tasks in Medicine

Automated Registry Reporting

- An NLP use case is to extract values as needed by each use case
- Many health IT systems are burdened by regulatory reporting when measures such as ejection fraction (the amount of blood that your heart pumps each time it beats) are not stored as discrete values
- For automated reporting, health systems will have to identify when an ejection fraction is documented as part of a note, and also save each value in a form that can be utilized by the organization's analytics platform for automated registry reporting

NLP tasks in Medicine

- **Clinical Decision Support**
- The presence of NLP in Healthcare will strengthen clinical decision support. Nonetheless, solutions are formulated to bolster clinical decisions more acutely
- There are some areas of processes, which require better strategies of supervision, e.g., medical errors
- In addition, with the help of Isabel Healthcare, NLP is aiding clinicians in diagnosis and symptom checking

NLP tasks in Medicine

- **Clinical Trial Matching**
- Using NLP and machines in healthcare for recognising patients for a clinical trial is a significant use case
- Some companies are striving to answer the challenges in this area using Natural Language Processing in Healthcare engines for trial matching
- With the latest growth, NLP can automate trial matching and make it a seamless procedure
- One of the use cases of clinical trial matching is IBM Watson Health and Inspirata, which have devoted enormous resources to utilise NLP while supporting oncology trials

NLP tasks in Medicine

- **AI Chatbots and Virtual Scribe**
- Although no such solution exists presently, the chances are high that speech recognition apps would help humans modify clinical documentation
- Chatbots or Virtual Private assistants exist in a wide range in the current digital world, and the healthcare industry is not out of this
- Presently, these assistants can capture symptoms and triage patients to the most suitable provider
- New startups formulating chatbots comprise BRIGHT.MD, which has generated Smart Exam, “a virtual physician assistant” that utilises conversational NLP to gather personal health data and compare the information to evidence-based guidelines along with diagnostic suggestions for the provider

NLP tasks in Medicine

- **Risk Adjustment and Hierarchical Condition Categories**
- Hierarchical Condition Category coding, a risk adjustment model, was initially designed to predict the future care costs for patients
- In value-based payment models, HCC coding will become increasingly prevalent
- HCC relies on ICD-10 coding to assign risk scores to each patient
- Natural language processing can help assign patients a risk factor and use their score to predict the costs of healthcare

NLP tasks in Medicine

- **Computational Phenotyping**
- NLP is altering clinical trial matching; it even had the possible chances to help clinicians with the complicatedness of phenotyping patients for examination
- Example, NLP will permit phenotypes to be defined by the patients' current conditions instead of the knowledge of professionals
- To assess speech patterns, it may use NLP that could validate to have diagnostic potential when it comes to neurocognitive damages, for example, Alzheimer's, dementia, or other cardiovascular or psychological disorders
- In addition, Winterlight Labs is discovering unique linguistic patterns in the language of Alzheimer's patients

NLP tasks in Medicine

- **Review Management & Sentiment Analysis**
- NLP can also help healthcare organisations manage online reviews
- It can gather and evaluate thousands of reviews on healthcare each day
- In addition, NLP finds out PHI or Protected Health Information, further data related to HIPPA compliance
- Some systems can even monitor the voice of the customer in reviews; this helps the physician get a knowledge of how patients speak about their care and can better articulate with the use of shared vocabulary
- Similarly, NLP can track customers' attitudes by understanding positive and negative terms within the review

NLP tasks in Medicine

- **Dictation and EMR Implications**
- On average, EMR lists between 50 and 150 MB per million records, whereas the average clinical note record is almost 150 times extensive
- For this, many physicians are shifting from handwritten notes to voice notes that NLP systems can quickly analyse and add to EMR systems
- By doing this, the physicians can commit more time to the quality of care
- Much of the clinical notes are in amorphous form, but NLP can automatically examine those
- In addition, it can extract details from diagnostic reports and physicians' letters, ensuring that each critical information has been uploaded to the patient's health profile

NLP tasks in Medicine

- **Root Cause Analysis**
- Another exciting benefit of NLP is how predictive analysis can give the solution to prevalent health problems
- Applied to NLP, vast caches of digital medical records can assist in recognising subsets of geographic regions, racial groups, or other various population sectors which confront different types of health discrepancies
- The current administrative database cannot analyse socio-cultural impacts on health at such a large scale, but NLP has given way to additional exploration
- In the same way, NLP systems are used to assess unstructured response and know the root cause of patients' difficulties or poor outcomes

Techniques Used Within NLP

- Once data preprocessing, lexical, syntactical, and semantic analysis of corpus has taken place, we are required to transform text into mathematical representations for evaluation, comparison, and retrieval
- This is achieved through transforming documents into the vector space model with scoring and term weighting essential for query ranking and search retrieval
- Documents can be in the form of patient records, web pages, digitalized books, and so forth

N-grams

- N-grams are used in many NLP problems
- If X = number of words in a given sentence K, the number of n-grams for sentence K would be

$$N\ grams_K = X - (N - 1)$$

- For example, if N = 2 (bigrams), the sentence “David reversed his metabolic syndrome” would result in n-grams of the following:
- David reversed
- Reversed his
- His metabolic
- Metabolic syndrome
- N-grams preserve the sequences of N items from the text input
- N-grams can have a different N value:
- unigrams for when N = 1,
- bigrams when N = 2, and
- trigrams when N = 3
- N-grams are used extensively for spelling correction, word decomposition, and summating texts

TF IDF- Term Frequency Inverse Document Frequency

- a statistical technique that tells how important a word is to a document in a collection of documents
- statistical measure is calculated by multiplying 2 distinct values- term frequency and inverse document frequency
- **Term Frequency**
- used to calculate how frequently a word appears in a document
  - **TF (t, d) = count of t in d/ number of words in d**
- The words that generally occur in documents like stop words- “the”, “is”, “will” are going to have a high term frequency

TF IDF

- **Inverse Document Frequency**
- Before getting to Inverse Document Frequency, let's understand Document Frequency first
- In a corpus of multiple documents, Document Frequency measures the occurrence of a word in the whole corpus of documents(N).
    - **DF(t)= occurrences of t in N documents**
- This will be high for commonly used words in English Inverse Document Frequency is just the opposite of Document Frequency
    - **IDF(t)= N / occurrences of t in N documents**
- This basically measures the usefulness of a term in our corpus
- Terms very specific to a particular document will have high IDF.
- Terms like- biomedical, genomic, etc. will only be present in documents related to biology and will have a high IDF.

TF IDF

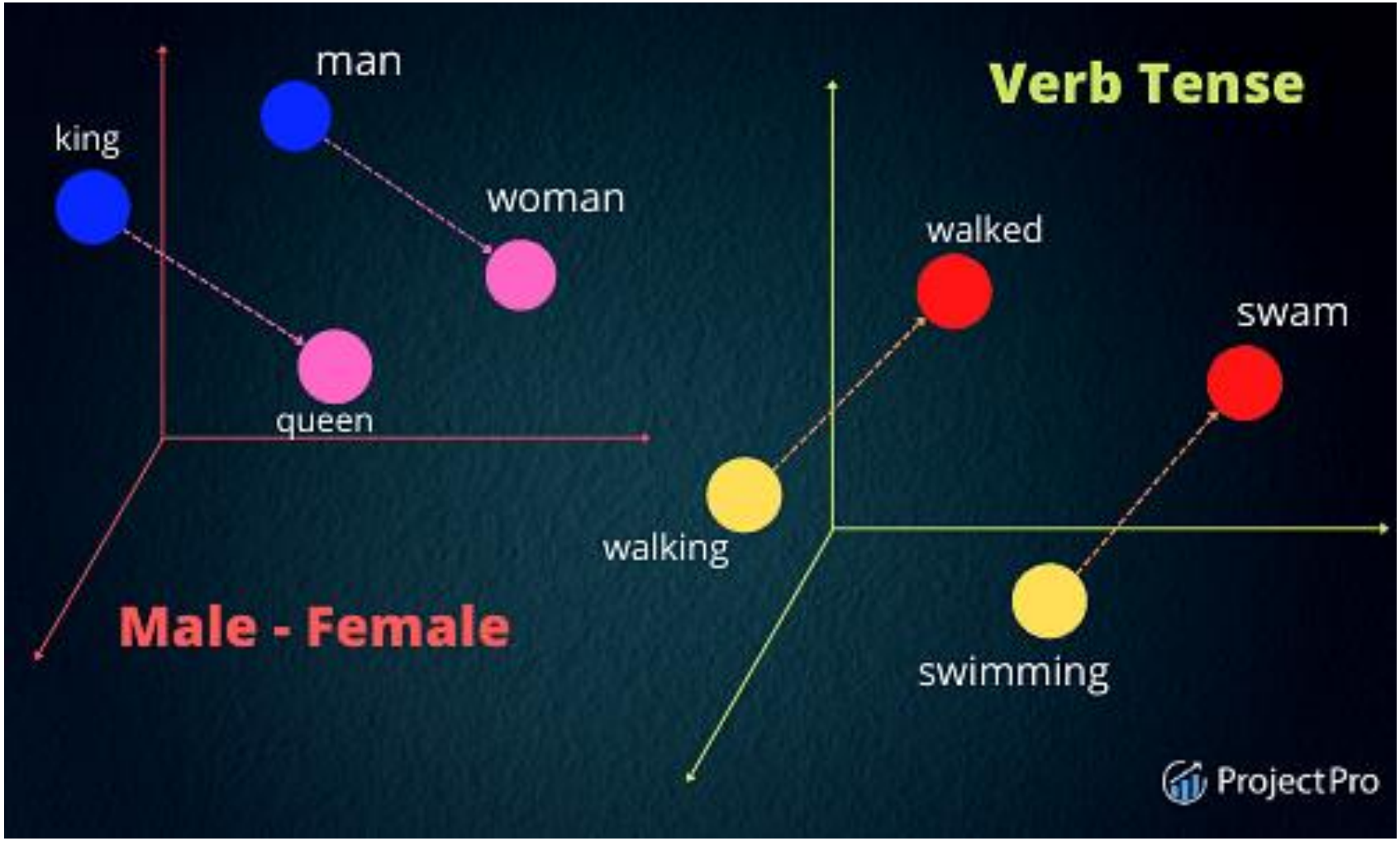
- **TF-IDF = Term Frequency * Inverse Document Frequency**
- The whole idea behind TF-IDF is to find important words in a document by finding those words that have a high frequency in that document but not anywhere else in the corpus
- For a document related to Computer Science, these words could be – Computational, data, processor, etc. but for an astronomical document, it would be- extraterrestrial, galactic, black hole, etc

Word Embeddings

- As we know that machine learning and deep learning algorithms only take numerical input, so how can we convert a block of text to numbers that can be fed to these models
- When training any kind of model on text data be it classification or regression- it is a necessary condition to transform it into a numerical representation
- The answer is simple, follow the word embedding approach for representing text data
- This NLP technique lets you represent words with similar meanings to have a similar representation

Word Embeddings

- Word Embeddings also known as vectors are the numerical representations for words in a language
- These representations are learned such that words with similar meaning would have vectors very close to each other
- Individual words are represented as real-valued vectors or coordinates in a predefined vector space of n-dimensions


man
king
woman
queen
Male - Female
Verb Tense
walked
swam
walking
swimming
ProjectPro

Word Embeddings

- Consider a 3- 3-dimensional space as represented above in a 3D plane
- Each word is represented by a coordinate(x,y,z) in this space
- Words that are similar in meaning would be close to each other in this 3-dimensional space
- The distance between walked and the king would be greater than the distance between walked and walking since they have the same root word-walk
- Word embeddings are also useful in understanding the relationship between words- what king is to the queen, a man is to woman
- Hence, in the vector space, the distance between king and queen would approximately be equal to the distance between man and woman

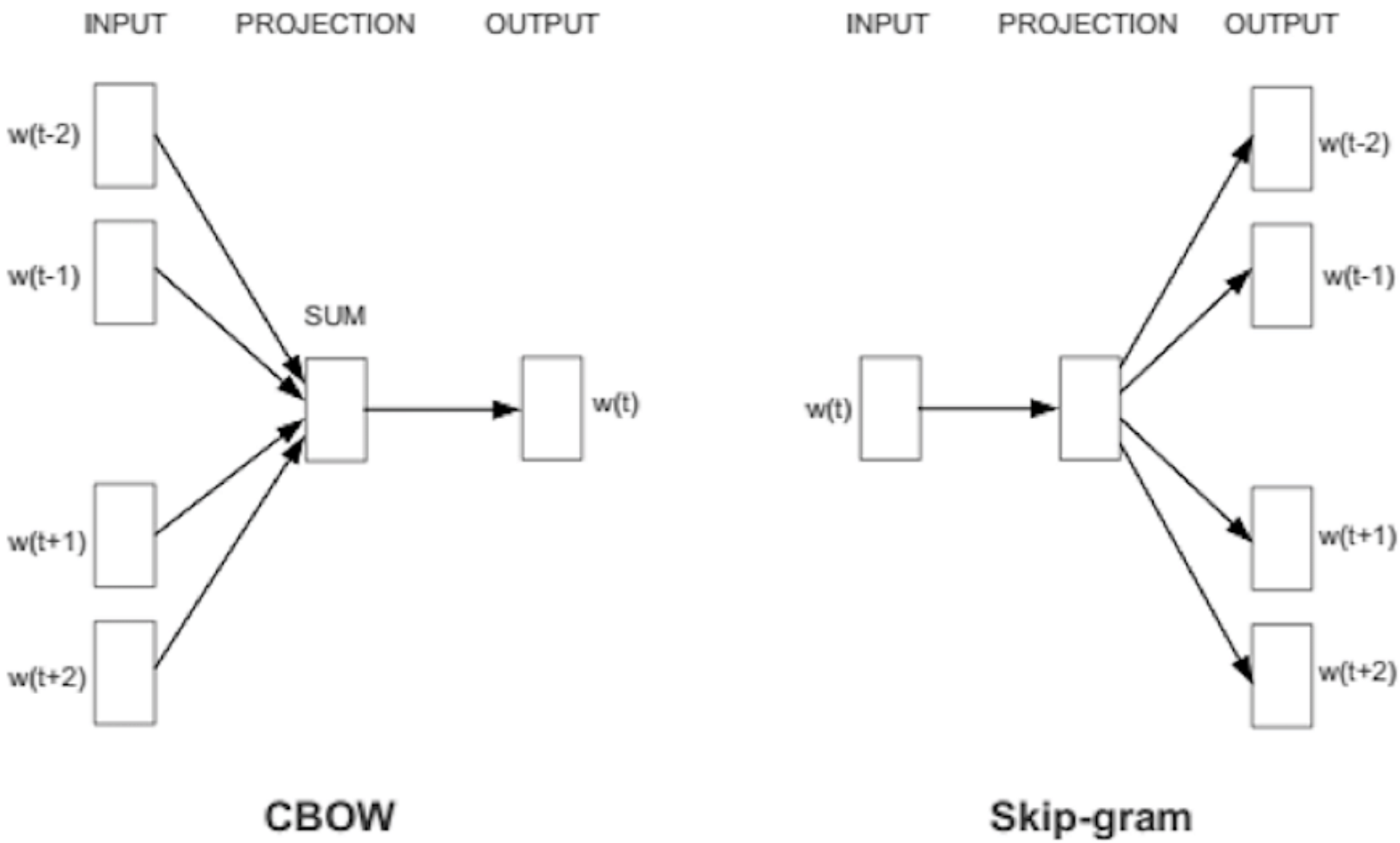
Word Embeddings

- One can either use predefined Word Embeddings (trained on a huge corpus such as Wikipedia) or learn word embeddings from scratch for a custom dataset
- There are many different kinds of Word Embeddings out there like GloVe, Word2Vec, TF-IDF, CountVectorizer, BERT, ELMO etc

Word2Vec

- neural network model that learns word associations from a huge corpus of text
- trained in two ways, either by using the Common Bag of Words Model (CBOW) or the Skip Gram Model

Word2Vec


INPUT
PROJECTION
OUTPUT
w(t-2)
w(t-1)
SUM
w(t)
w(t+1)
w(t+2)
CBOW
INPUT
PROJECTION
OUTPUT
w(t)
w(t-2)
w(t-1)
w(t+1)
w(t+2)
Skip-gram

Word2Vec

- In the CBOW model, the context of each word is taken as the input and the word corresponding to the context is to be predicted as the output
  - Consider an example sentence- “The day is bright and sunny.”
  - In the above sentence, the word we are trying to predict is sunny, using the input as the average of one-hot encoded vectors of the words- “The day is bright”.
  - This input after passing through the neural network is compared to the one-hot encoded vector of the target word, “sunny”
  - The loss is calculated, and this is how the context of the word “sunny” is learned in CBOW

Word2Vec

- The Skip Gram model works just the opposite of the above approach, we send input as a one-hot encoded vector of our target word “sunny” and it tries to output the context of the target word
- For each context vector, we get a probability distribution of V probabilities where V is the vocab size and also the size of the one-hot encoded vector in the above technique

Advantages of NLP

- helps users to ask questions about any subject and get a direct response within seconds
- offers exact answers to the question means it does not offer unnecessary and unwanted information
- helps computers to communicate with humans in their languages
- very time efficient
- Most of the companies use NLP to improve the efficiency of documentation processes, accuracy of documentation, and identify the information from large databases

Disadvantages of NLP

- NLP may not show context
- NLP is unpredictable
- NLP may require more keystrokes
- NLP is unable to adapt to the new domain, and it has a limited function that's why NLP is built for a single and specific task only

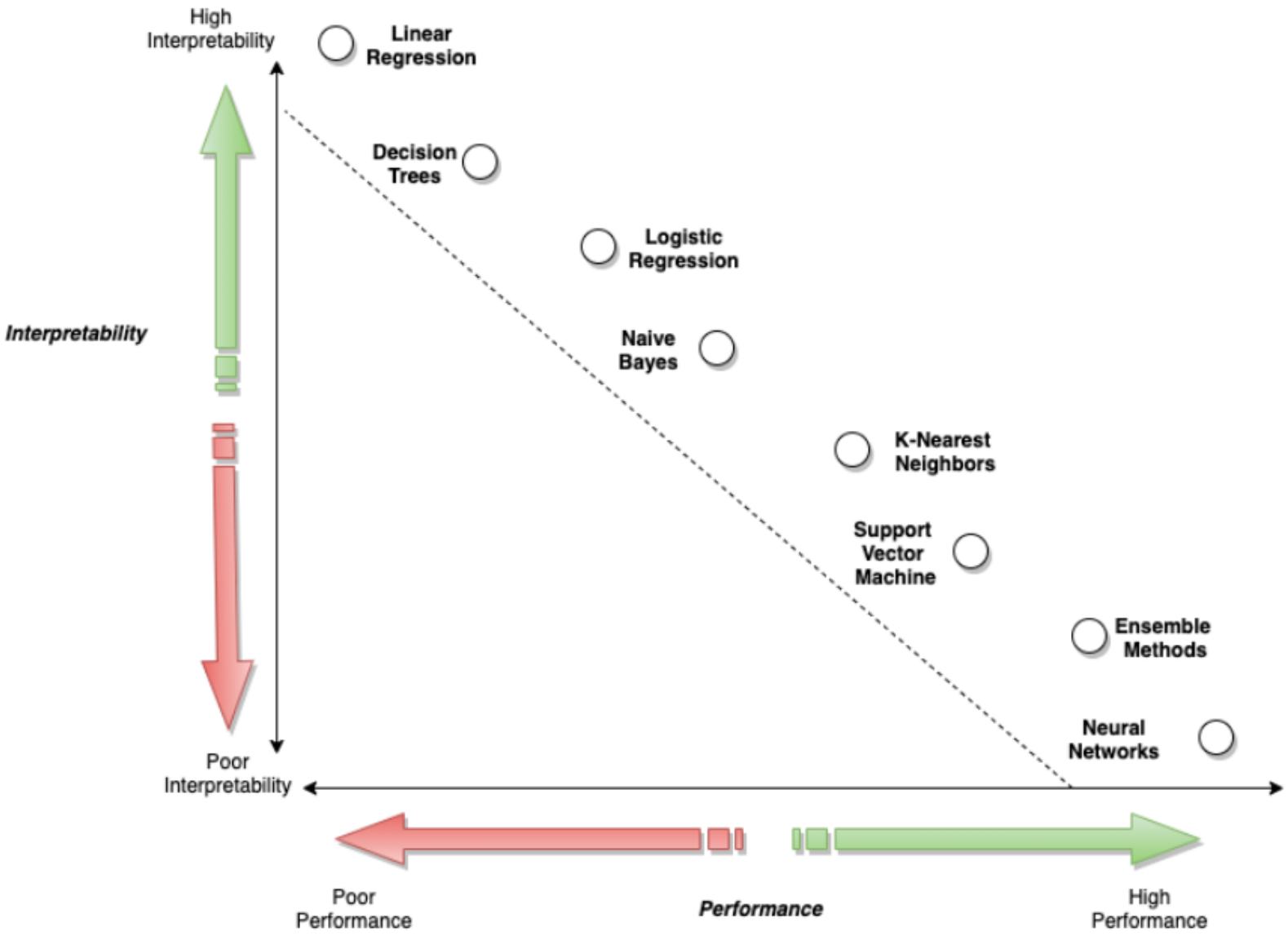
Applications of NLP

- Question Answering
- Spam Detection
- Sentiment Analysis
- Machine Translation
- Spelling correction
- Speech Recognition
- Chatbot
- Information extraction
- Natural Language Understanding (NLU)

Model Interpretability using Explainable AI in NLP applications

Interpretability

- Need: If a business wants to understand exactly why and how the model is generating predictions
- Interpreting the model's weights and features to determine the given output called interpretability
- Eg: An economist builds a multi-variate regression model to predict an inflation rate, they can view the estimated parameters to measure the expected output
- However, high interpretability typically comes at the cost of performance


High Interpretability
Linear Regression
Decision Trees
Logistic Regression
Naive Bayes
K-Nearest Neighbors
Support Vector Machine
Ensemble Methods
Neural Networks
Interpretability
Poor Interpretability
Poor Performance
Performance
High Performance

Explainability /Explainable AI(XAI)

- Set of processes and methods that allows human users to comprehend and trust the results by ML algorithms
- Example:
  - A news media outlet uses a neural network to assign categories to different articles. The news outlet cannot interpret the model in depth; however, they can use a model agnostic approach to evaluate the input article data versus the model predictions
  - With this approach, they find that the model is assigning the *Sports* category to business articles
  - Although the news outlet did not use model interpretability, they were still able to derive an explainable answer to reveal the model's behavior

When datasets are large and the data is images or text, neural networks can meet the customer's AI/ML objective with high performance. In such cases, where complex methods are required to maximize performance, data scientists focus on model explainability instead of interpretability

Explainable Models

- **Intrinsically explainable- directly interpretable**
  - Designed to be simple and transparent enough that we can get a sense of how it works by looking at its structure
  - e.g. simple regression models and small decision trees
- **Post-hoc explainable**
  - For more complicated, already trained models, we can use explainability tools to obtain post-hoc explanations
  - Explanations of sufficiently complex models such as deep neural networks are always post-hoc explanations as they are not directly interpretable

Types of Explanations

- **Global explanations**
  - Refers to details of what features are important to the model overall
  - Measured by looking at effect sizes or determining which features have the biggest impact on model accuracy
  - Helpful for guiding policy or finding evidence for, or rejecting a hypothesis that a particular feature is important

Types of Explanations

- **Local explanations** - A local explanation details how a ML model arrived at a specific prediction
- For tabular data, it could be a list of features with their impact on the prediction
- For a computer vision task, it might be a subset of pixels that had the biggest impact on the classification
- Local explanations are useful for deep-dive insights or diagnosing issues and can provide answers to questions like:
  - Why did the model return this output for this input?
  - What if this feature had a different value?

Techniques for Explainable AI

Counterfactual explanations:

- Technique used to explain the output of a model by showing how the output would have changed if certain inputs had been different
- useful for understanding how the model arrived at its decision and identifying which inputs were most important in generating the output

Explainable NLP

- NLP models have become more sophisticated and capable of handling a wide range of tasks, such as machine translation, sentiment analysis, and question-answering
- However, these models have one major limitation – they are often considered black boxes, meaning that it's challenging to understand how they reach their conclusions or what information they are relying on to make decisions
- This limitation has led to the development of Explainable NLP, which aims to make these models more transparent and interpretable

Need for Explainable NLP

- Increasingly being used in critical applications such as healthcare, finance, and legal systems, where the consequences of errors can be severe
  - Example: In disease diagnosis, it's crucial to know how the model arrived at its decision to determine if it's accurate & reliable
- XAI can help us understand the limitations of current models and identify areas for improvement
  - Example, if an NLP model is biased against certain demographics, we can take steps to mitigate those biases and improve the model's performance
- Explainable NLP can help build trust in the model's output, which is essential for widespread adoption
  - If people don't understand how a model is making decisions, they may be hesitant to trust its output, which can limit its usefulness

Techniques for Explainable AI

Layer-wise relevance propagation (LRP):

- LRP is a technique used to explain the predictions of deep neural networks
- It works by propagating the relevance of each output back through the layers of the network to identify which input features were most important in making the prediction
- LRP is useful for interpreting the output of NLP models because it can tell us which words or phrases were most important in generating the output

Techniques for Explainable AI

Model distillation:

- Technique used to simplify complex models into simpler ones that are easier to interpret
- Involves training a simpler model to mimic the output of a more complex model
  - Simpler model can then be used to generate predictions and is often easier to interpret than the original model

Techniques for Explainable AI

Attention mechanisms:

- Technique used in NLP to help models focus on specific parts of the input sequence
- Useful for interpreting the model's output because they can tell us which parts of the input the model is paying attention to
  - Example, in a machine translation task, we can use attention mechanisms to see which parts of the input sentence are being used to generate each output word

Techniques for Explainable AI

Rule-based models:

- Models that rely on explicit rules or heuristics to generate predictions
- Useful for NLP tasks that require a high level of interpretability, such as legal or regulatory compliance tasks
- Can also be used to identify biases in more complex models

Techniques for Explainable AI

Interpretable neural networks:

- type of neural network that is designed to be more transparent and interpretable
- They use techniques such as attention mechanisms, sparse activations, and structured layers to make them more interpretable
- Example, the Structured Self-Attention Network (SSAN) is an interpretable neural network that uses a structured attention mechanism to identify the most relevant parts of the input

Techniques for Explainable AI

Human-in-the-loop approaches:

- Involve incorporating human feedback into the model to improve its interpretability
  - Example, we can ask human annotators to label certain input features as important or unimportant and use this feedback to improve the model's interpretability

Explainable AI in different Industries

- **Healthcare:**
  - essential for ensuring patient safety and privacy.
  - Medical professionals rely on NLP models to analyze patient records and make treatment decisions.
  - Explainable NLP helps to ensure that these decisions are based on accurate, reliable, and transparent information.
- **Finance:**
  - relies heavily on NLP models to analyze financial data and make investment decisions
  - crucial for ensuring these decisions are based on accurate and transparent data and that there is no bias or unfairness in the analysis
- **Legal:**
  - NLP models are used for a variety of purposes, including contract analysis and legal research
  - Legal decisions are based on accurate and reliable information and that there is no bias or unfairness in the analysis

Explainable AI in different Industries

- **Customer Service:**
  - In industries such as retail and hospitality, NLP models are used to provide customer service and support
  - Ensuring that customers receive accurate and relevant information and that there is no bias or unfairness in the responses
- **Government:**
  - Used for a variety of purposes, including analyzing public opinion and sentiment analysis
  - Important for ensuring that decisions made based on NLP analysis are transparent and accountable.