Journal of **N**onlinear **A**nalysis and **O**ptimization
Vol. 15, Issue. 1, No.8 : 2024
ISSN : **1906-9685**

A STUDY ON NATURAL LANGUAGE PROCESSING "EXPLORING THE DEPTHS OF NATURAL LANGUAGE PROCESSING WITH SEMANTIC SENSE"

Dr. M. Santhanalaskshmi, Assistant Professor, SDNB Vaishnav College for Women, Chromepet.
Esther Rani. A, SDNB Vaishnav College for Women, Chromepet
Nithiyashree. K. M, SDNB Vaishnav College for Women, Chromepet

Abstract
Natural language processing(NLP) is an extend of Artificial Intelligence and Linguistics, dedicated to make computers can learn to think more like a person, closing the gap between them. Enabling computers to process written and spoken language similarly to humans is the aim of research into natural language processing. Natural language processing(NLP), is thoroughly explored in this study, which connects on all the major topics from its historical origins to its current relevance. The introduction explores the significance and exigency of natural language processing. Word embedding and word conversion into numerical representations are two examples of the methods used by robots to interpret language. The paper explains the complicated workings of natural language processing (NLP). In addition, it looks into how statistics fit into NLP and examines some essential statistical machine learning methods. The many methods to NLP are examined, and the paper outlines the duties required, illuminating the difficulties encountered in this dynamic domain. A case study illustrating the real-world application of NLP concepts is provided, featuring a practical application of sentiment analysis. For scholars, practitioners, and hobbyists interested in comprehending the complex field of natural language processing, this paper provides an extensive resource.

Keywords: natural language processing, sentiment analysis, machine learning

Introduction:
Natural language processing (NLP) has recently obtained much attention for representing and analyzing human language computationally[1].It has spread its applications in various fields such as machine translation, email spam detection, information extraction, summarization, medical, and question answering[2].NLP is a extend of Artificial Intelligence and Linguistics, dedicated to make computers understand the statements or words written in human languages A subfield of artificial intelligence known as "natural language processing" enables machines to comprehend, analyze, and interpret human languages. This discipline blends computer science, languages, and machine learning.Numerous technologies are powered by it, such as automated text summarization, virtual assistants, speech recognition, machine translation, and sentiment analysis, among many more.

1.1.The field of NLP is divided into three parts

1.1.1. Speech recognition
Often referred to as speech-to-text, speech recognition is the process of accurately translating voice data into text data.A key component of applications that employ voice commands or provide spoken question replies is speech recognition. The way individuals speak quickly, slurring words together, with varied emphasis and intonation, in diverse dialects, and frequently using improper grammar makes speech detection particularly difficult.

1.1.2. Natural language understanding

Computer's ability to understand language.With the development of a digital system that can comprehend and react correctly to human speech, natural language understanding seeks to enable human-computer communication.

1.1.3. Natural language generation

Generation of natural language by a computer.It is the most crucial component of natural language processing (NLP) since it enables computers to respond to users in natural language. Develops genuine written or spoken words from organized and unstructured data using a software process powered by artificial intelligence. Moreover, it can be utilized to convert complex data, such as numerical input, into reports that are simple to comprehend. Reports on finances or weather conditions might be automatically generated using NLG.

This field integrates computer science, linguistics, and machine learning. In this paper, we'll talk about this in brief.

2. History

The pioneer of Natural Language processing is **Alan Turning.** Following World War II, the field of natural language processing was founded in the 1940's. People at the time realized how important it was to translate text between languages and planned to build a system that could perform this kind of translation automatically. But it was clear that the work was more difficult than many had initially thought. By 1958, a few scholars had noted important problems with the advancement of NLP. Among these scholars was Noam Chomsky, who considered it concerning that grammatical model of language treated nonsensical but grammatically accurate sentences as equally irrelevant as nonsense but non-grammatically correct ones.

Between 1957 and 1970, researchers in the field of natural language processing divided into two groups: symbolic and stochastic. Many computer scientists and linguists who studied formal languages and syntax were part of the symbolic, or rule-based, research community, which saw this area as the start of artificial intelligence study. Statistical and probabilistic approaches to natural language processing, particularly in the areas of optical character identification and textual pattern recognition, piqued the interest of stochastic researchers. This is how natural language processing comes into picture.

3. Is Natural language processing important?

Machine translation is the most popular use of natural language processing (NLP) that helps overcome language barriers. The need to access and evaluate the growing amount of data that is available online is becoming more and more important. Machine translation can be used to translate data from one language to another. The efficiency of machine translation is increased when the machine is able to understand the meaning of sentences thanks to the use of NLP techniques. Sentiment analysis benefits greatly from the application of NLP techniques. It helps identify the sentiment present in a number of internet postings and comments. Businesses use natural language processing (NLP) techniques to mine internet reviews for information about what customers think about their goods and services. Automatic summarization can be completed more quickly and effectively by using NLP. For example, while obtaining information from social media, automatic summarization is crucial not only for acingulate the relevance of documents and data but also for comprehending the emotional connotations of the data. When utilized to provide an overview of a news article or blog post while keeping a strategic distance from several sources and optimizing the diversity of content acquired, automatic summarizing is very important. By doing this, the difficulty of finding a crucial piece of information from a massive database can be decreased. The sophisticated natural language processing (NLP) methods enable non-developers to interact with computer systems and extract useful information from them. The fundamental counterparts for the input phrases can be found using natural language processing (NLP), and this assists users who are unfamiliar with the technical terms used in computers by coordinating them with the appropriate responses. The other common uses of NLP are speech-to-text, voice recognition, language comprehension, text classification, information extraction, question answering, social media feeds, and spam filtering.

4. Approaches to Natural language processing

The field of language processing, or NLP, includes a range of methods and strategies for interpreting, analyzing, and producing data related to human language. These are a few of the major NLP methodologies.

4.1.Rule-based Systems

Rule-based NLP systems use human-crafted linguistic patterns and rules to process and analyze text. Usually created by linguists or subject matter experts, these rules encode particular linguistic phenomena including syntactic structures, semantic links, and grammatical rules. Although rule-based systems are highly accurate in some situations, they are not always scalable or able to handle complicated linguistic phenomena.

4.2.Statistical Methods

Through the use of probabilistic models and machine learning algorithms, statistical approaches to natural language processing (NLP) can extract patterns and correlations from massive volumes of annotated text data. n-gram language models, Conditional Random Fields (CRFs), and Hidden Markov Models (HMMs) are a few of these strategies. For applications such as named entity recognition, machine translation, and part-of-speech tagging, statistical approaches work well.

4.3.Machine Learning

Especially in tasks like text classification, sentiment analysis, and document clustering, machine learning techniques are important to natural language processing (NLP). Many supervised learning algorithms and deep learning architectures, including Recurrent Neural Networks (RNNs) and Transformer models, have been widely used in NLP applications. These include Support Vector Machines (SVM), Naïve Bayes, and others.

4.4.Deep Learning

Because deep learning allows models to learn hierarchical representations of linguistic data, it has transformed several areas of natural language processing (NLP). State-of-the-art performance has been attained in tasks like language modelling, question answering, and text generation by deep learning architectures like Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks, and more recently, Transformer-based models like BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer).

4.5.Semantic Analysis

Finding the meaning in text data and comprehending its context are the main goals of semantic analysis. To represent and study the semantics of words, phrases, and sentences, methods including semantic role labelling, word embedding’s (e.g., Word2Vec, GloVe), and semantic parsing are utilized. For tasks like question answering, sentiment analysis, and information retrieval, semantic analysis is essential.

4.6.Neural Language Models

Neural network designs are utilized by neural language models to acquire continuous word and sentence representations. These models can recognize intricate language links and patterns since they have been trained on vast corpora of text data. Language generation, text summarization, and dialogue generation are made possible by neural language models.

4.7.Hybrid Approaches

Numerous NLP applications integrate various methodologies to capitalize on the advantages of distinct approaches. For text normalization, for instance, a system might employ rule-based preprocessing. For additional analysis and interpretation, it might next employ a statistical or deep learning model. In complex natural language processing (NLP) problems, hybrid techniques are frequently used when one methodology is insufficient to deliver the desired results.

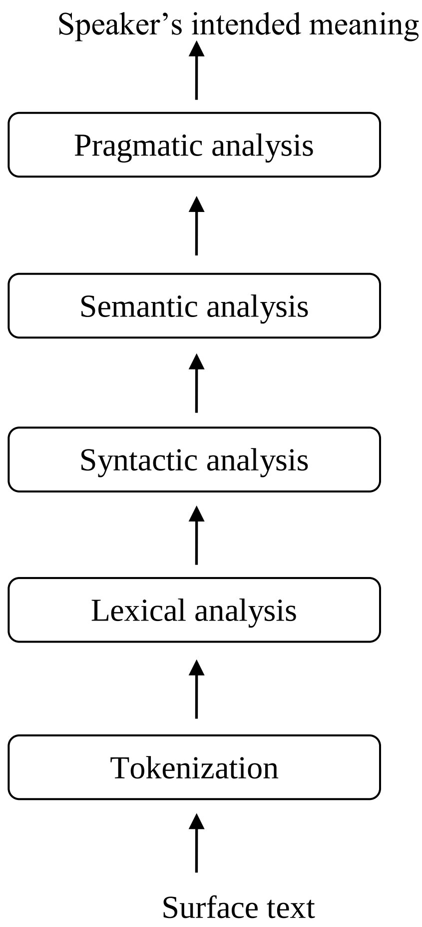
Because of developments in computational linguistics, deep learning, and machine learning, these methods to NLP are always changing. In order to effectively handle the difficulties associated with comprehending and processing human language, researchers and practitioners are still investigating novel approaches.

5. How natural language processing works?

NLP employs a wide range of methods to give computers the same level of natural language comprehension as people. Natural language processing uses artificial intelligence (AI) to analyze and interpret spoken or written language in a way that is understandable to a computer. Computers have programs to read and microphones to gather audio, just as people have various s0ensors, such as ears to hear and eyes to see. Computers have programs to process the inputs that they receive, much as humans have brains to process information. The input is eventually changed into computer-understandable code during processing.

Linguistics is one of the core disciplines of NLP. Linguistics isthe scientific study of language and its structure. It includes:

- Phonetics and phonology
- Morphology
- Syntax
- Semantics
- Pragmatics



All these are analyzed by NLP to determine the structure and meaning of the language. This linguistic information is then transformed by computer science into rule-based machine learning algorithm and carries out desired activities and address particular difficulties.

6. How machine understand human language

NLP is the application of statistical models and a set of algorithms that let computers interpret and analyze human language. This covers jobs like sentiment analysis, text summarization, speech recognition, and language translation.NLP has many difficulties, one of which is interpreting the ambiguity and complexity of human language. Human language, in contrast to computer languages, is frequently imprecise and subject to many interpretations.

Let’s consider an example:

Sentence 1: Hang him not, leave him.

Sentence 2: Hang him, not leave him.

Both these sentences look similar but differ in their meaning. So how do machine understand the dissimilarity in them since it is hard for human to understand it quickly.In order to get around this problem, NLP algorithms employ a variety of strategies to examine sentence structure and interpret meaning, including named entity recognition, syntactic parsing, and part-of-speech tagging. To further increase a model's accuracy over time, machine learning algorithms are employed to train it using big linguistic data sets.

7. What mechanism is used to make machine understand them?

A development in natural language processing called Word embedding or Word vectorizationhas enhanced computers' comprehension of textual material. Since computer can only understand binary language that is 0 and 1 so word embedding is the best way.

7.1.Word Embedding

Comparable words can have comparable vector representations thanks to this method, which uses numerical vectors to represent words and texts.Word embeddings are numerical depictions of words in a reduced dimension that encompass syntactic and semantic data.Once the words are turned to numbers, we use these numbers to perform a variety of operations to detect patterns in the words, sentences, paragraphs, books, and other text.Following the conversion of the words into vectors, we must utilize methods like Euclidean distance and Cosine Similarity to find words that are comparable.

The general method for matching comparable documents which is based on calculating the amount of common words between the documents is called Euclidean distance.[i] Even if there are more common words in the paper but it covers various themes, this strategy won't work. The “Cosine Similarity” method is used to determine the degree of similarity between the papers in order to get around this problem.

Let us see an example for Word embedding:

Sentence 1: Alwin likes coffee more than Aaron likes tea

Sentence 2: Aryn likes coffee more than Alwin likes tea

Now the similarity between these two sentences is:

Coffee	1	1
Aryn	0	1
Alwin	1	1
Aaron	1	0
Likes	2	2
More	1	1
Than	1	1
Tea	1	1

Now the vector format of the above 2 sentence is:

Item 1: [1, 0, 1, 1, 2, 1, 1, 1]

Item 2: [1, 1, 1, 0, 2, 1, 1, 1]

Once the ve

ctor is found then machine learning algorithm is applied to find out the expected output.

8. How statistics comes into role in NLP?

Basically gathering, analyzing, interpreting, presenting, and organizing of data is the subject of statistics.The foundation of machine learning is statistics, which offers the methods and instruments for data analysis and interpretation. Essentially, machine learning algorithms are constructed using the theoretical foundation that statistics gives.We can make sense of complicated phenomena by using statistics to analyze and summarize the data.In contrast, machine learning is a potent instrument that enables computers to learn from and anticipate future events based on data.This is where the elegance of the combination of machine learning and statistics is revealed.The foundations upon which machine learning is built are the laws of statistics which include:

- Building models for machine learning
- Analyzing the outcomes
- proving the models
- supporting sophisticated methods

Not only can a strong grasp of statistics help us build and evaluate machine learning models more effectively, but it also makes it possible for us to meaningfully and practically interpret the models' results.

9. Statistical Machine Learning Techniques

- **Machine learning** is an artificial intelligence application that makes use of statistical methods to teach computers to learn and make judgments without explicit programming.
- **Supervised learning** isone kind of machine learning algorithm that gains knowledge from labelled data is called supervised learning. Data that has been assigned a correct classification or answer is called labelled data.When we use well-labelled data to instruct or train a machine, this is known as supervised learning.

9.1.Un-supervised learning

Unsupervised learning is a machine learning approach that utilizes unlabelled data to learn. This indicates that there are no labels or categories present in the data. Unsupervised learning aims to find links and patterns in the data without providing explicit instructions.

9.2.Statistical machine learning

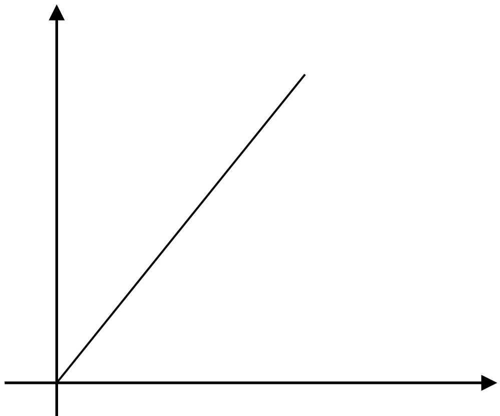
It is creating models that can learn from data and make predictions or judgments by applying statistical approaches.

The following are the statistical machine learning algorithms used for developing models:

- Linear Regression.
- Logistic Regression.
- Decision Trees.
- Random Forest.
- Navis Bais

9.3.Linear Regression

One technique for predicting the future course of events is linear regression, which establishes a linear relationship between an independent and dependent variable. This statistical approach is utilized for predictive analysis in data science and machine learning. When other factors change, the independent variable which also serves as a predictor or explanatory variable stays constant. On the other hand, variations in the independent variable affect the dependent variable. Predicting continuous or numerical variables, linear regression is a supervised learning approach that models a mathematical relationship between variables.



9.4.Types of linear regression

9.4.1. Simple Linear Regression

There is just one independent variable and one dependent variable in this kind of linear regression, which is the most basic kind. For basic linear regression, use the following equation:

$Y=\beta_0+\beta_1X$

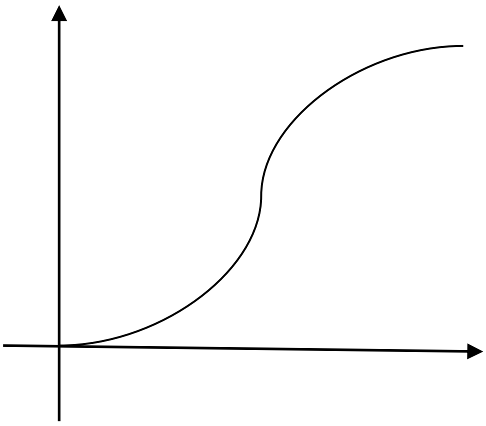
9.4.2. Multiple Linear Regression

This approach uses one dependent variable and multiple independent variables. For multiple linear regression, the equation is:

$Y=\beta_0+\beta_1X+\beta2X+...........\beta nX$

9.5.Logistic Regression

Supervised learning involves the use of algorithms like logistic regression, which is one of the most widely used machine learning techniques. It makes use of a predetermined collection of independent factors to predict the category dependent variable. The result needs to be a discrete or category value. It could be True or False, 0 or 1, Yes or No, etc. Except for their respective applications, logistic regression and linear regression are very similar. Those involving regression are solved using linear regression, while those involving classification are solved using logistic regression. Logistic regression uses a function called **Sigmoid function**. A mathematical tool called the Sigmoid function is used to convert expected values into probabilities. It converts any real number between 0 and 1 into another value. The logistic regression's result must lie between 0 and 1, and as it cannot be greater than this, it takes the shape of an "S" curve. The logistic or sigmoid function is another name for the S-form curve.



9.5.1. Logistic Regression Equation:

From the linear regression equation, one can get the logistic regression equation. The following are the mathematical procedures to obtain equations for logistic regression:

$$Y=b_0+b_1x_1+b_2x_2+b_3x_3+...........b_nx_n$$

9.6.Decision Trees

While decision trees are a supervised learning technique, they are mostly employed to solve classification problems. However, they can also be used to solve regression problems. They are simple to comprehend, interpret, and put into practice.[ii] It is organized like a hierarchical tree with a root node, branches, internal, and leaf nodes. The term itself implies that it displays the predictions that arise from a sequence of feature-based splits using a flowchart akin to a tree structure. A decision made by the leaves marks the end of it, which begins with a root node.

9.6.1. Decision Tree Terminologies

- Root Node

Since all other nodes in the tree are descended from the root node, it is the highest node. When the dataset is divided into subsets, it indicates the feature or characteristic that is taken into account at the initial decision point. In order to determine which feature best divides the data into subsets that are more homogeneous with regard to the target variable (in the case of classification) or produce better predictions (in the case of regression), the process of picking the root node requires analyzing many characteristics.

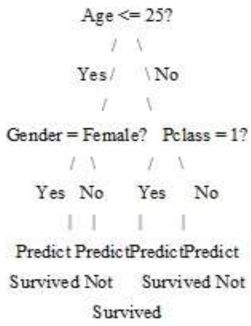
- Decision Nodes
  Nodes whose decisions are based on certain attribute values in the tree. These nodes have branches leading to different nodes.
- Leaf Nodes
  The terminal points of the branches, where decisions or projections are made. On leaf nodes, there are no longer any branches.
- Branch / Sub-Tree
  Split tree that produced a new tree.
- Parent Node
  A node that divides into smaller nodes.The initial node that gives rise to a split.
- Child Node
  Nodes created as a result of a split from a parent node.

Let us examine an alternative scenario in which our goal is to forecast a cruise passenger's likelihood of survival by taking into account their age, gender, and social status. This is a tiny dataset:

Age	**Gender**	**Class**	**Survived**
25	Male	Upper	No
40	Female	Lower	Yes
35	Female	Upper	Yes
15	Male	Lower	Yes
28	Male	Upper	No
55	Female	Lower	No
22	Male	Lower	Yes
38	Female	Upper	Yes
18	Female	Lower	No
50	Male	Upper	Yes

- Since age offers the best split, it is selected as the root node.
  Proceed to the next level if Age is equal to or less than 25. Proceed to the following level if your age is more than 25.
- The next best feature for age under or equal to 25 is gender.
  Don't survive if your gender is male (Leaf node: No). Survive if the gender is female (Leaf node: Yes).
- Class is the next best attribute for those over 25.
  Don't survive if your class is lower (Leaf node: No). Survive if the class is Upper (Leaf node: Yes).

Decision trees are inverted, meaning multiple nodes split off from the root at the top. All decision trees are comprised of a series of if-else statements. It moves on to the next node connected to that decision after verifying that the condition is true.
The decision tree table displayed above is illustrated in the diagram below.



It is simple to predict the result because the data are processed in a tree-like structure. Decision trees are thought to be the best algorithms when compared to all others.

9.7.Random Forest

Random forest is made from multiple decision trees. The ability of the Random Forest Algorithm to handle data sets with both continuous variables as in regression and categorical variables as in classification is one of its most crucial properties. In tasks involving regression and classification, it performs better.

9.7.1. This is an explanation of Random Forest's operation

- Decision Trees
- Ensemble Learning
- Randomization
- Bootstrap Aggregating
- Feature Randomization
- Voting (Classification) / Averaging (Regression)

9.7.2. Naive Bayes

The Naïve Bayes algorithm is a straightforward yet effective classification method that relies on the independence of predictors and the Bayes theorem. It is referred regarded as "naïve" since it makes the assumption that a feature's existence in a class has nothing to do with the existence of any other feature. Naïve Bayes has proven to be highly effective in many real-world applications, especially in text classification and spam filtering, despite this oversimplified assumption.

Here is a summary of each Naïve Bayes component:

A foundational statement in probability theory, Bayes' statement expresses the likelihood of an event based on past knowledge of potential confounding variables. In mathematics, it is expressed as:

$$P(A/B)=\frac{P(B/A)\times P(A)}{P(B)}$$

Where:

- The chance that event A will occur given that event B has occurred is expressed as $P(A|B)$.
- The chance that event B will occur given that event A has occurred is expressed as $P(B|A)$.
- The odds that occurrences A and B will occur are denoted by $P(A)$ and $P(B)$, respectively.

9.7.3. Conditional Independence Assumption

Given the class label, Naïve Bayes assumes that every characteristic is conditionally independent. This implies that the existence of a single feature has no bearing on the existence of any other feature.

9.7.4. Classification Process

In order to classify an instance, Naïve Bayes first determines which class has the highest probability by calculating the posterior probability of each class given the input data. Typically, the term "maximum a posteriori" (MAP) estimate refers to the class that has the highest posterior probability.

9.7.5. Naïve Bayes Types

Multinomial Naïve Bayes:

Appropriate for discrete feature classification (word counts, for text classification, for example).

Gaussian Naïve Bayes:

Predicted on the Gaussian distribution of continuous features.
Bernoulli naïve Bayes:
Similar to multinomial naïve Bayes, Bernoulli naïve Bayes is best suited for binary characteristics (such as the presence or absence of a word in a document).
In fact, Naïve Bayes typically performs remarkably well despite its simplicity and strong independence assumption, notably in text classification applications where the features are word counts or frequencies. Even with enormous datasets, it can be taught rapidly and with computational efficiency. However, if characteristics are highly correlated or the independence condition is broken, its performance might suffer.

10. Challenges in Natural language processing:

Artificial intelligence has permeated every aspect of our daily lives, from chatbots for customer support to Alexa and Siri, text and email autocorrection, and more. They all process, "understand," and react to spoken and written human language using machine learning algorithms and Natural Language Processing (NLP).

10.1. Homonyms and context-relevant words and phrases

The context of a sentence can change the meaning of the same words and phrases. Particularly in the English language, certain words may sound the same but have quite distinct meanings. As we have seen earlier
Sentence 1: Hang him not, leave him.
Sentence 2: Hang him, not leave him.
Humans can easily understand these because we can read the sentence in its whole and are familiar with all of the definitions. It's challenging for a machine, though. The primary obstacle in natural language processing is thus this.

10.2. Textual and vocal errors

Words that are misused or misspelled can cause issues for text analysis. While autocorrect and grammar checkers are capable of handling common errors, they are not always able to interpret the author's purpose. Mispronunciations, accents, stutters, and other spoken language can be challenging for a machine to comprehend. These problems can be reduced, though, as language databases expand and personal users educate intelligent assistants.

10.3. Practice data

NLP is fundamentally about interpreting language through analysis. Even the finest AI needs to spend a lot of time reading, listening to, and using a language in order to become fluent in it; humans need to be immersed in a language continuously for years. An NLP system's capabilities are determined by the training data that is given to it. The system will learn incorrectly or in an ineffective manner if it is fed inaccurate or dubious data.

10.4. Syntax and Semantics

Because of the intricacy of sentence patterns and the wide range of possible interpretations, it can be difficult to comprehend the syntax and semantics of human language. Tasks such as sentiment analysis, semantic role labelling, and parsing are made more difficult by its complexity.

10.5. Lack of Data

In order to properly train machine learning models, many NLP jobs require a significant amount of annotated data. But getting such information can be costly and time-consuming, particularly for specialized fields or languages with little resources.

10.6. Out-of-Vocabulary Words (OOV)

Words that are not in the training data that natural language processing (NLP) models frequently encounter are known as Out-of-Vocabulary Words (OOV) and can cause errors in tasks such as machine translation, speech recognition, and language modeling. Robust NLP systems must be able to handle terms that are not commonly used. These are the some of the challenges that are faced by NLP.
To tackle these obstacles, a blend of algorithmic breakthroughs, data-centric strategies, and cross-disciplinary cooperation amongst scholars in computer science, languages, and associated domains is

needed. Furthermore, developing reliable and inclusive systems depends on addressing ethical considerations and justice in NLP research and implementations.

11. Conclusion

The rapidly developing discipline of natural language processing(NLP), holds great promise for improving human-computer interaction, deciphering human language, and resolving This paper gives an insights of the various important terminologies of NLP , and it can be useful for the readers to start their work in NLP and its relevant applications.

References:

[1] Khurana, D., Koli, A., Khatter, K. et al. Natural language processing: state of the art, current trends and challenges. Multimed Tools Appl **82**, 3713–3744 (2023). https://doi.org/10.1007/s11042-022-13428-4.

[2] Yue Kang, Zhao Cai, Chee-Wee Tan, Qian Huang & Hefu Liu (2020) Natural language processing (NLP) in management research: A literature review, Journal of Management Analytics, 7:2, 139-172, DOI: 10.1080/23270012.2020.1756939

[3]**Diksha Khurana , Aditya Koli Kiran Khatter,Sukhdev Singh**,” Natural language processing state of art,current trends and challenges”, Multimedia Tools and Applications (2023) 82:3713–3744

[4]**Ronan Collobert, Princeton NJ. Jason, NY. L´eon Bottou ,Michael Karlen Koray Kavukcuoglu, Pavel Kuksa,”** Natural Language Processing (almost) from Scratch”, Journal of Machine Learning Research 1 (2000) 1-48

[5] Natural language processing 2004,8 lectures by **Anna Copestake**

[6]Introduction to natural language processing by **R.Kibble**

[7]Natural language processing by **Jacob Eisenstein**

[8]Natural language processing recipes(unlocking text data with machine learning using python)by **AkshayKulkarniAdarshaShivananda**

[9]Natural language processing with python by **Steven bird Ewan Klein & Edward loper**

[10]Speech and language processing (An introduction to to natural language processing, computational linguistics,and speech recognition Third edition) by **Daniel Jurafsky and James H.Martin**

[11]Machine learning & pattern recognition series (series edition) **by Ralf-Herbrich and ThoreGraepel**