# NATURAL LANGUAGE PROCESSING FRAMEWORK FOR TEXT ANALYSIS

**[1]Rajan Kumar Jha, [2]Anju Rajput, [3]Priyanka Mitra, [4]B. Umamaheswari**

[1234]*Assistant Professor*

[134]*Department of Computer Science & Engineering*

[2]*Department of Electronics & Communication Engineering*

*Jaipur Engineering College & Research Centre, Jaipur*

[1]*rajanjha.cse@jecrc.ac.in*, [2]*anjurajput.ece@jecrc.ac.in*, [3]*priyankamitra.cse@jecrc.ac.in*
[4]*umamaheswari.cse@jecrc.ac.in*

## ABSTRACT

***Natural Language Processing (NLP) plays a pivotal role in extracting meaningful insights from unstructured text data. In this paper, we propose a NLP framework designed specifically for text analysis tasks. A NLP framework enables efficient identification of entities, sentiment, and topics. Researchers and businesses can leverage this advantage to gain insights from large volumes of textual content, such as customer reviews, social media posts, or research articles. Our study delves into the advantages, such as enhanced accuracy and efficiency in text analysis, as well as the limitations, including computational resource requirements and domain-specific challenges. This research contributes to a deeper understanding of the adoption and utilization of NLP frameworks in the development community, informing future advancements and best practices in the field of text analysis. Additionally, this paper presents a study on the development and evaluation of a NLP framework for sentiment analysis, focusing on the assessment of its effectiveness using the Student Feedback dataset as a benchmark.***

**Keywords - Natural Language Processing (NLP), Text Analysis, Frameworks, User Experience, Sentiment Analysis.**

## [1] INTRODUCTION

Natural Language Processing (NLP) is a subfield of computer science and artificial intelligence that focuses on interactions between computers and human (natural) languages. Its primary goal is to enable machines to understand, analyze, and generate human language. NLP plays a crucial role in various applications, especially when dealing with textual data. The significance of natural language processing can be defined in such a way that it helps the computers for communicating with the humans in their own languages. If we take illustration, many languages are very complex and different in nature. We've been expressing ourselves in so numerous types of ways and both verbally and while writing also. Natural Language Processing framework for textual analysis or sentiment analysis (opinion mining) is defined as a method to perform analysis which specifically aimed at counting, grouping and bunching words (known as keywords) for rooting the structure of meaningful content from larger quantity of content provided as input. The textual analysis is being used for exploring the textual content and inferring new type of variables from raw data which might be imaged, filtered, visualized and further being used in the form of inputs for the models and other types of statistical approaches. The following objectives of the study were chosen:

1. To delve in the evolution of Natural Language Processing (NLP) ways, outlining their evolution and advancements over time.
2. To explore and interpret various concepts related to NLP, gaining perceptivity into the different approaches and methodologies employed in the field.
3. To examine and outline the steps involved in the NLP procedure, encompassing data preprocessing, feature extraction, model training, and evaluation, giving a comprehensive understanding of NLP workflow.
4. To assess and estimate the practical applications and use of NLP across different areas and industries, correlating its effectiveness, extents, and possible areas for refinement.

## II. LITERATURE REVIEW

The literature review of Natural Language Processing (NLP) encompasses a comprehensive examination of research studies, methodologies, and advancements in the field. A thorough literature review provides insights into the evolution of NLP techniques, from early rule-based systems to modern deep learning approaches. It also explores steps involved in NLP and its application domains such as sentiment analysis, machine translation, and information extraction.

### *A. NLP over the years: Evolution*

The concise journey of Natural Language Processing through its development:

<u>1950s</u> - Early Beginnings: The roots of NLP trace back to the 1950s when researchers began exploring language processing using rule-based methods. Early efforts focused on tasks like machine translation and information retrieval.

1960s - Linguistic Theories and Formal Grammars: In the 1960s, researchers incorporated linguistic theories and formal grammars into NLP. Chomsky’s transformational grammar influenced this era, emphasizing syntax and structure.

1970s - Statistical Approaches: Statistical methods gained prominence. Hidden Markov Models (HMMs) and n-grams were used for language modeling. Parsing algorithms improved sentence structure analysis.

1980s – 90s - Corpus Linguistics and Machine Learning: The focus shifted to corpus linguistics and machine learning. Part-of-speech tagging, named entity recognition, and word embeddings gained traction. Stochastic methods became popular.

2000s – 2010s - Deep Learning and Neural Networks: Deep learning transformed NLP. Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs) improved sequence modeling. Attention mechanisms (e.g., BERT, GPT, XLNet) achieved state-of-the-art results.

2020s - Present - Transfer Learning and Beyond: Transfer learning dominates NLP. Pre-trained models fine-tuned for specific tasks yield remarkable performance. Multimodal NLP integrates text, images, and other modalities.

*B. NLP: Theory & Concept*

Natural Language Processing (NLP) is a fascinating field that combines data science, computer science, and linguistics to enable machines to understand and process human language. There are two main elements of Natural Language Processing:

Natural Language Understanding: NLU focuses on comprehending and interpreting human language. Tasks involved in NLU are Tokenization, Part-of-Speech Tagging, Named Entity Recognition (NER), Syntax Parsing, Sentiment Analysis for determining the emotional tone of text (positive, negative, neutral). NLU forms the basis for understanding user queries, chatbots, and information retrieval.

Natural Language Generation: NLG creates human-like text from structured data or other forms of input. Tasks involved in NLG are Text Summarization, Machine Translation, Content Creation (i.e. Crafting articles, emails, or creative writing). NLG aims for coherence, fluency, and context-awareness and enhances communication, personalization, and content generation.

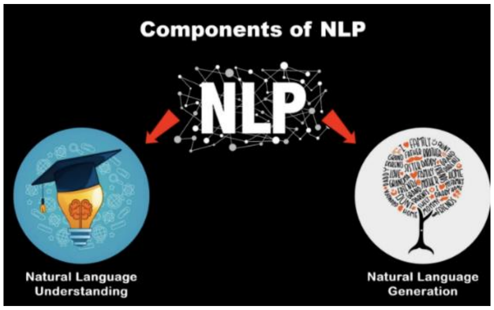


**Fig 1:** Components of NLP

Different Approaches to NLP:

Rule-Based Approach: Based on linguistic rules and patterns. Involves predefined rules for language processing.

Machine Learning Approach: Relies on statistical analysis. Trained models learn from data to perform tasks like text classification, sentiment analysis, and named entity recognition.

Neural Network Approach: Utilizes artificial neural networks (e.g., recurrent and convolutional neural networks). Effective for tasks like sequence-to-sequence modeling, machine translation, and text generation.

## III. METHODOLOGY: NLP PIPLINE

The following figure shows basic steps involved in any NLP framework.

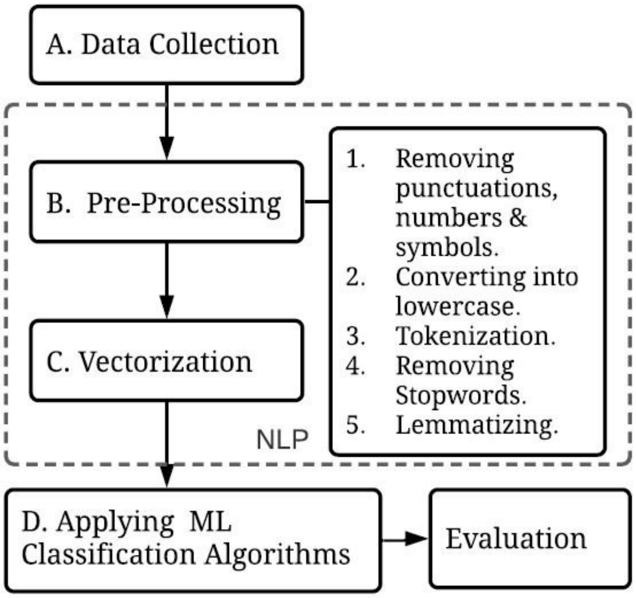


**Fig 2:** NLP Framework

Data Collection

Data collection is the foundation of any NLP framework. It is the initial process which involves gathering relevant textual data from various sources which is referred to as dataset or bag of words. For instance, you might collect social media posts, news articles, or customer reviews. This data will serve as the basis for providing as input to NLP model.

Pre-processing

Before feeding the data into machine learning models, we need to pre-process it. Here are the essential steps:

1. Cleaning and Removing Noise: We remove any irrelevant information, such as HTML tags or special characters, punctuation, numbers, and symbols are stripped from the text. The text is converted to lowercase for consistency as our ML classifiers are generally trained on lowercase words.
2. Tokenization: Tokenization breaks down the text into individual words or tokens. For example, the sentence "I like research!" would be tokenized or split into a vector like this: ["I", "like", "research"].
3. Stopword Removal: Stopwords (common words like "the," "and," or "in") don't carry much meaning and can be removed. This step reduces noise and improves model performance.
4. Lemmatization: Lemmatization reduces words to their base form (lemma). For instance, "running" becomes "run," "better" becomes "good," etc.

Applying ML Classification Algorithms

Now that we have pre-processed our data, let's apply machine learning classification algorithms. Some popular algorithms include:

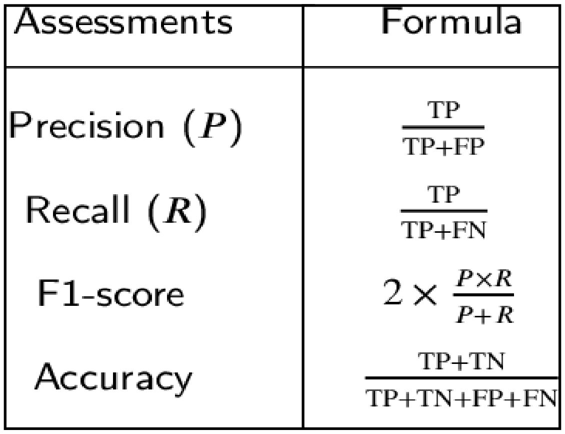
Support Vector Machines (SVM): SVMs are effective Supervised algorithms for performing binary classification tasks. They try to find a hyperplane (separation boundary) that best separates positive and negative instances.

Random Forest: Random Forest is an ensemble learning method that combines the classification from multiple decision trees. It's robust and handles noisy data well. It excels in handling imbalanced datasets and achieving high accuracy.

Evaluation

To evaluate our models, we use metrics such as accuracy, precision, recall, and F1-score. Let's assume we've trained our models on our dataset which have given some predictions as output on providing an input and we want to compare its performance. The following table shows various evaluation metrics.

**Table 1:** Evaluation Metrics



| Assessments | Formula |
|---|---|
| Precision $(P)$ | $\frac{\text{TP}}{\text{TP+FP}}$ |
| Recall $(R)$ | $\frac{\text{TP}}{\text{TP+FN}}$ |
| F1-score | $2 \times \frac{P \times R}{P+R}$ |
| Accuracy | $\frac{\text{TP+TN}}{\text{TP+TN+FP+FN}}$ |

*Accuracy:* It estimates the overall correctness of predictions made by model. It is a measure of how well a machine learning model is performing on given input. It provides the ratio of correct classifications made by the model to total classifications made. It is commonly represented as a value between 0 and 1 (or between 0% and 100%).

*Precision and Recall:* Precision tells how often a ML model is correct while predicting the output target class. Recall tells whether a ML model can find all objects/outputs of a particular target class. Precision is the ratio of true positive predictions to all positive predictions while Recall is the ratio of true positive predictions to all actual positive instances.

*F1-score:* Harmonic mean of precision and recall. It provides a balance between above both metrics (i.e. precision and recall). The two metrics contribute equally to the F1 score, ensuring that the F1 metric correctly indicates the reliability of a model.

## IV. IMPLEMENTATION/SIMULATION

To assess the effectiveness of the proposed methodology for text analysis, we conducted experiments on student feedback dataset, the same dataset we used in our project. The dataset consists of student's feedback labeled as positive, negative or neutral sentiments, providing a suitable testbed for evaluating sentiment analysis algorithms.

Prior to applying the proposed methodology, we conducted standard preprocessing steps on the text data, including tokenization, removal of punctuation, and stemming. Let's consider an example statement, statement "I like this research paper". Providing this statement as input to NLP model in order to analyze the sentiment of the statement using NLP steps we have following internal processing:

Lowercase: Convert the given text into its lowercase form- "i like this research paper".

Tokenization: Split the text into individual words - ["i", "like", "this", "research", "paper"].

Stop word removal: Remove common stop words like "this" - ["i", "like", "research", "paper"].

Lemmatization: Reduce words to their base form - ["i", "like", "research", "paper"].

Sentiment Analysis: Utilize a pre-trained sentiment analysis model to classify the sentiment. In this case, the sentiment would likely be classified as positive.

Result: The sentiment classification would be **Positive**.

## V. RESULTS & DISCUSSIONS

The experimental results demonstrated that the proposed methodology achieved good performance in sentiment classification on the given input feedback. It exhibited high accuracy, precision, and recall values, indicating its effectiveness in accurately predicting the sentiment polarity of student reviews. Additionally, the analysis of misclassified instances provided insights into the challenges faced by the sentiment analysis models, such as ambiguous language or context-dependent sentiments. This information can guide future improvements in model training and feature engineering.

While the results are promising, it's essential to acknowledge the limitations of the study, such as getting low accuracy when multiple keywords of opposite polarity occur in same sentence. Future research could explore advanced techniques, such as deep learning architectures and ensemble methods, to further improve the performance of sentiment analysis models.

## VI. CONCLUSION

The paper has provided a comprehensive overview of the evolution and steps of Natural Language Processing (NLP). The evolution of NLP has been characterized by a progression from rule-based systems to statistical methods and, more recently, to deep learning and neural network approaches. Furthermore, our exploration of the steps involved in NLP, including data preprocessing, feature extraction, model training, and evaluation, has offered valuable insights into the intricacies of NLP workflow. Empirical evaluations on benchmark datasets, such as the student feedback dataset for sentiment analysis, demonstrated the effectiveness of the proposed methodology.

Natural language processing and text analysis are essential components in various applications, encompassing:

Investigation discovery: These technologies aid in identifying patterns and clues within emails or reports, assisting in the detection and resolution of crimes.

Subject matter expertise: They are employed to classify content into meaningful topics, enabling individuals to discern trends and take informed actions.

Social media analytics: Leveraging these tools allows for the tracking of awareness and sentiment surrounding specific topics, as well as the identification of key influencers.

Sentiment analysis: It involves assessing the emotional tone or polarity expressed in textual content (like positive, negative or neutral), providing insights into public opinion and attitudes.

## REFERENCES

[1] Lehnert, Wendy, and Beth Sundheim. "A performance evaluation of text-analysis technologies." AI magazine 12, no. 3 (1991): 81-81

[2] Ahonen, Helena, Oskari Heinonen, Mika Klemettinen, and A. Inkeri Verkamo. Applying data mining techniques in text analysis. Report C-1997-23, Dept. of Computer Science, University of Helsinki, 1997.

[3] Fisher, Ingrid E., Margaret R. Garnsey, Sunita Goel, and Kinsun Tam. "The role of text analytics and information retrieval in the accounting domain." Journal of Emerging Technologies in Accounting 7, no. 1 (2010): 1-24.

[4] Day, M. Y. (2020). Artificial Intelligence for Text Analytics.