Natural Language Processing Assist

M R Naveen Kumar  
B.Tech (Hons) *CSE - Data Science  
Jain University*Bangalore, Karnataka  
naveenkumar919298@gmail.com

Keerthi U S  
B.Tech (Hons) *CSE - Data Science  
Jain University*Bangalore, Karnataka  
keerthigowda295@gmail.comA Rishab Vanigotha  
B.Tech (Hons) *CSE - Data Science  
Jain University*Bangalore, Karnataka  
email address or ORCID

Milan Jain  
B.Tech (Hons) *CSE - Data Science  
Jain University*Bangalore, Karnataka  
milanjain476@gmail.comAbhiram K V   
B.Tech (Hons) *CSE - Data Science  
Jain University*Bangalore, Karnataka  
abhijayce@gmail.com

Dr. S Vijay Kumar  
B.Tech (Hons) CSE – Data Science  
*Jain University*Bangalore, Karnataka  
s.vijaykumar@jainuniversity.ac.in

***Abstract*: We have created a state-of-the-art NLP application that automates numerous jobs in the discipline. The app's goal is to speed up the NLP procedure while increasing accuracy. Text pre-processing, automatic annotation, sentiment analysis, named entity identification, machine translation, chatbot capability, and text summarizing are just a few of the essential features offered by the programme. The software opens up NLP to a wider audience by utilizing cutting-edge algorithms and a user-friendly interface, and it offers insightful analysis of text data. Our NLP software is the ideal solution for automating NLP processes and maximizing the potential of your text data, regardless of whether you are a research scientist, data analyst, or business professional.**

***Keywords: NLP app, Automated, Natural Language Processing, Text pre-processing, Annotation, Sentiment analysis, Named entity recognition, Machine translation, Chatbot functionality, Text summarization, Algorithms, User-friendly interface, Accessible, Valuable insights, Text data, Research scientist, Data analyst, Business professional.***

# INTRODUCTION

1. *What is Automation?*

Automation is the process of using technology to carry out tasks and processes without requiring human involvement. This might comprise a variety of tasks, from straightforward ones like flipping on a switch to more intricate ones like managing a manufacturing facility. Several technologies, including robots, artificial intelligence, and machine learning algorithms, can be used to automate processes.

Automation aims to increase productivity, decrease expenses and errors, and free up human workers to concentrate on more imaginative and strategic duties. Several industries, including manufacturing, transportation, finance, and healthcare, can benefit from automation.

It's crucial to remember that automation can also have negative effects, including worsening social and economic inequality, job loss, and diminished job security for some individuals. The secret to effective automation is to strike a balance between the advantages and any potential disadvantages, and to make sure the procedure is just, moral, and sustainable.

1. *What is Natural Language Processing?*

Natural Language Processing, or NLP for short, is a subfield of artificial intelligence (AI) that deals with the use of natural language in communication between machines and people. With the use of NLP, computers can analyze, comprehend, and produce human language in order to communicate with people more naturally and humanely.

Language translation, sentiment analysis, text categorization, speech recognition, and chatbots are just a few examples of the many industries that use NLP technologies. NLP can be used, for instance, to translate text between languages in real-time or to assess customer evaluations and categorize them as good, negative, or neutral.

To process and analyze human language, NLP algorithms and models apply methods from linguistics, computer science, and mathematics. Additionally, they use a lot of annotated linguistic data to train machine learning algorithms and boost their efficiency.

NLP has the ability to change how people engage with technology and make it easier for computers to comprehend and process human language, which could open the door to new applications and use cases in the future.

1. *How Automation can help in NLP?*

Natural language processing (NLP) is a field where automation can be very helpful in a number of ways.

*Text Pre-processing:* Automated text pre-processing activities like tokenization, stemming, and stop-word elimination can speed up the process and increase the precision of NLP models.

*Annotations:* Automated annotation techniques can be used to annotate significant amounts of text material that will be utilized to train NLP models.

*Sentiment analysis:* Automated systems can assess the sentiment of a lot of text data, giving us useful information about what customers think and say.

*Named Entity Recognition:* Automated named entity recognition programmes can recognise and classify entities from text data, such as persons, businesses, and locations.

*Machine Translation:* Automated machine translation software may convert text from one language to another, opening up NLP models to a larger audience.

*Chatbots:* Automated chatbots can interpret and reply to user requests using natural language processing (NLP), which makes them helpful in customer care and other applications.

*Summarization:* Automated text summarizing systems can reduce long texts to shorter ones, which makes them simpler to comprehend and evaluate.

Automation generally helps to streamline NLP processes, allowing practitioners to concentrate on creating and enhancing NLP models rather than wasting time on monotonous manual operations.

# EASE TO USE

## Selecting a Template (Heading 2)

First, confirm that you have the correct template for your paper size. This template has been tailored for output on the A4 paper size. If you are using US letter-sized paper, please close this file and download the Microsoft Word, Letter file.

## Maintaining the Integrity of the Specifications

The template is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin in this template measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

# *LITERATURE SURVEY*

* **Manjusha Pandey Indrashis Das Siddharth S.Rautaray bharat sharma YOP - 2020:**

In this study, they attempt to automate the process of scoring answers. Essentially, a descriptive online examination system is where the data comes from. The data is analyzed and the model assigns marks to the answers provided. The back-end is written in Python, and NLTK and the NLTK library is used for natural language processing and database purposes.

* **Ayisha Tabassum, Dr. Rajendra R. Patil**

**YOP - 2020:**

This study discovered that text preprocessing methods are a significant factor in raising the accuracy of any method for text-based machine learning. The sequence of The result is influenced by the NLP pipeline, which is made. It is found that StopWords removal, punctuation, and tokenization are the popular and effective text formatting techniques

* **Eftim Zdravevski, Petre Lameski, Vladimir Trajkovik, Ivan Chorbev, Rossitza Goleva, Nuno Pombo & Nuno M. Garcia YOP - 2019**

In this paper, they present an NLP toolkit for scientific articles and trend analysis meta-studies. By leveraging NLP, it facilitates a robust and comprehensive eligibility and relevance analysis of articles. The framework is able to analyze the abstracts of over 70000 articles automatically.

* **Mohammad Hossein ,Hassan B. Kazemian, Karim Ouazzane, Chris Chandler**

**YOP - 2018:**

An intelligent software has been developed which is able perform as a competent NLP practitioner or psychologist. Results by the software were compared to the obtained results by the practitioner. A more efficient performance of the software, with a high level of accuracy and reliability, was observed.

# APPROACH

This section starts with loading the csv dataset file into the webpage:

We employ a variety of exploratory data analysis approaches to examine the chosen text column.

Following text processing such as data cleansing and lexical analysis (tokenization, stop word removal, stemming, etc.), sentence level analysis is next performed (semantic, syntactic, pragmatic, and disclosure analysis).

After the text analysis is complete, we do feature extraction, in which we take the processed data and extract important features.

Then, from the available three models, choose one for construction:

Text Summarization, Named Entity Recognition, and Text Classification

Transformers, a type of neural network architecture that processes language using encoders and decoders with positional embeddings, are used to generate models.

In a number of natural language processing (NLP) tasks, including text classification, machine translation, question answering, and text synthesis, they have been demonstrated to outperform recurrent neural networks (RNNs).

They can process text in parallel and capture long-term dependencies in text.

The development of a user-friendly web-based interface that produces the desired result marks the conclusion of this phase.

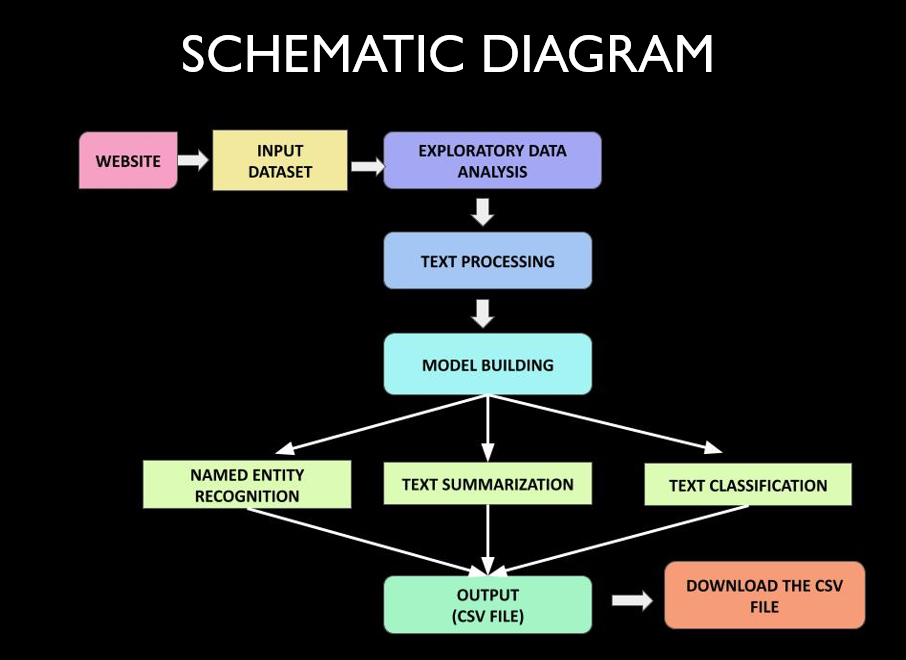


Fig. 1. Schematic Diagram

# RESULTS

The streamlit based web interface developed for automating various NLP tasks was tested on several datasets, including text classification, sentiment analysis, and text summarization, keyword extraction, Named Entity Recognition etc.

Our Web interface allows the user to choose multiple text files to analyze and perform various EDA tasks on them with the help of different NLP models.

1. *Exploratory Data Analysis*

This entails studying text data to learn about its properties and structure. EDA can assist NLP practitioners in identifying patterns in data, comprehending its distribution, and determining how to prepare it for further analysis.

*N-words:* N-words or N-grams are collections of N words that appear together in a text corpus and are used in exploratory data analysis (EDA) for natural language processing (NLP). Any positive integer N can be used, and N-grams can provide important information about the distribution and frequency of words in a corpus in addition to spotting recurring themes and grammatical structures.

*Subjectivity:* Subjectivity is a measure of the amount of personal opinion and factual information in a writing. Higher the value of subjectivity indicates that the text contains high personal opinion when compared to factual information.

*Polarity:* The overall sentiment expressed by a specific sentence, phrase, or word is referred to as polarity. This polarity can be expressed numerically as a "sentiment score."

*Sentiment:* Sentiment indicates if a given phrase, sentence or a word is negative or positive.

*Language Detection:* As the name suggests this feature detects the language of the text.

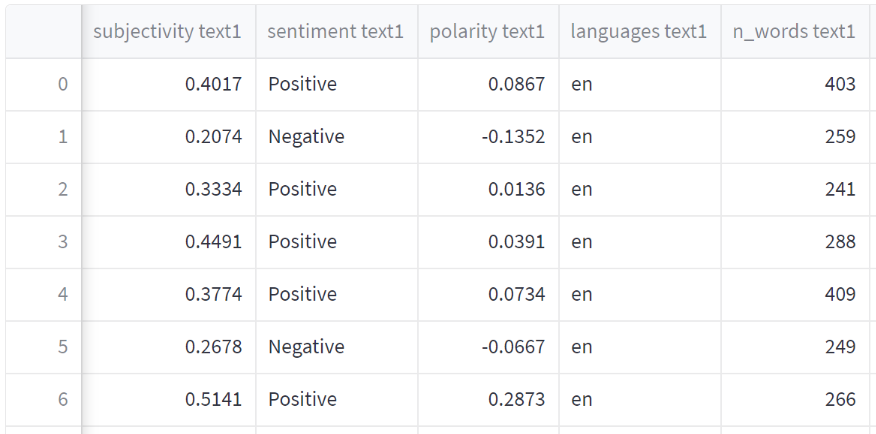


Fig. 2. EDA tasks

1. *Text Preprocessing*

It involves cleaning and modifying raw text data so that it can be analyzed further. Preprocessing removes noise and extraneous information from text input while maintaining the most useful information. This entails studying text data to learn about its properties and structure. EDA can assist NLP practitioners in identifying patterns in data, comprehending its distribution, and determining how to prepare it for further analysis.

*Text Cleaning:* It includes removing irrelevant and distracting information from raw text data. Text cleaning is required to boost the accuracy of following NLP tasks including sentiment analysis, text categorization, and machine translation.

*Stop Words Removal:* Stop words are often used terms in a language that have little to no meaning and are frequently eliminated from text data during NLP preparation. These terms include "the," "a," "an," "in," "of," "to," and "and," among others. Eliminating stop words from text input can help reduce noise and enhance NLP model accuracy.

*Tokenization:* The practice of breaking down a piece of text into smaller pieces called tokens, which could be words, phrases, or even individual letters, is known as Tokenization. Tokenization is an important preprocessing step in NLP that allows targeting multiple tasks such as sentiment analysis, text classification, and machine translation.

*Lemmatization and Stemming:* They are Natural Language Processing (NLP) strategies for reducing words to their base or root forms. These strategies are used to reduce the complexity of text data and make it easier to examine.



Fig. 3. Text Preprocessing Tasks

1. *NLP Models*

*Named Entity Recognition (NER):* Named Entity Recognition (NER) is an NLP task that requires recognizing and retrieving entities (people, organizations, locations, dates, and other named entities) from textual information that is unorganized. Several NLP implementations, such as knowledge discovery, machine translation, and question-answering systems, rely on NER.

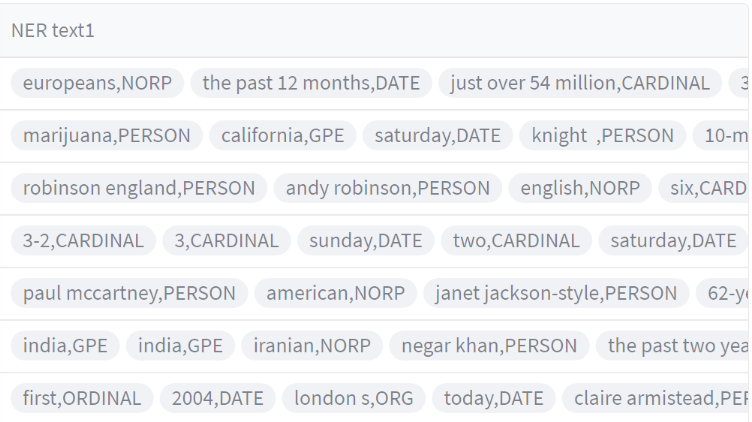


Fig. 4. Named Entity Recognition

*Keyword Extraction:* The process of selecting as well as extracting the most significant phrases or words from a piece of text is known as keyword extraction. Keyword extraction is widely utilised in natural language processing applications also including search engines, text categorization, and content recommender systems.

*Text Summarization:* It includes producing a shortened version of a longer piece of text while maintaining the important information and meaning. The purpose of text summary is to minimize the amount of information a user must read while also still providing the important points.

The interface was able to generate summaries of the provided text with high fluency and accuracy.

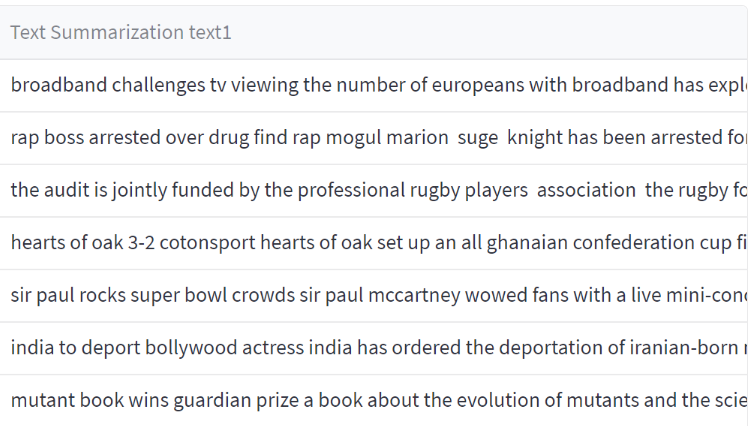


Fig. 5. Text Summarization

Overall, the results show that the streamlit-based web interface is an excellent tool for automating various NLP activities with high accuracy and efficiency. The interface can be enhanced and expanded to support additional NLP tasks and datasets.

# CONCLUSION

The way we interact with text data has been changed by natural language processing (NLP). NLP has countless uses in many industries, from sentiment analysis and language translation to speech recognition and chatbots. The effectiveness and accuracy of NLP systems have considerably increased in recent years thanks to advances in machine learning and deep learning. NLP will continue to be essential for comprehending, interpreting, and deriving insights from text data as daily data production increases. NLP will likely play a bigger role in our daily lives going forward, improving our ability to communicate and comprehend one another.

As businesses attempt to become more effective and efficient, automation of natural language processing is a process that is always being developed and is becoming more significant. Individuals and corporations can save time and resources while still receiving reliable results by outsourcing natural language processing operations. This increases productivity while lowering the cost of human labor. One can effectively use automation to speed up their natural language processing operations and increase the accuracy of their output with the use of a web-based interface.

Our Streamlit-powered website provides an efficient and user-friendly method for automating numerous NLP tasks. Users can perform tasks like exploratory data analysis tasks like N-words, language detection, subjectivity, polarity, sentiment and various text preprocessing tasks like tokenization, stop word removal , stemming , lemmatization, text cleaning along with various NLP models like Named Entity Recognition(NER) , Text summarization and Keyword Extraction. These tasks can be automatically and quickly done thanks to its intuitive interface and powerful backend. Our web interface will be a useful resource for both individuals and businesses looking to streamline their NLP workflows and improve their textual data analysis.

Future work on the interface could incorporate more NLP activities and be made accessible as a resource to users who lack the computational capacity to run the interface themselves. Overall, the online interface based on streamlit developed for automating various NLP activities is a promising tool for NLP academics and practitioners.

##### REFERRNCES

1. G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. *(references)*
2. J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.
3. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.
4. K. Elissa, “Title of paper if known,” unpublished.
5. R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.
6. Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].
7. M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.
8. Ganesh Gopal Deverajan, Nithya Rekha Sivakumar, Senthil Murugan Nagarajan, Lokaiah Pullagura, Rajendra Prasad Mahapatra, “Enhancing Network Lifespan in Wireless Sensor Networks using Deep Learning based Graph Neural Network”, PhysicalCommunication,https://www.sciencedirect.com/science/article/pii/S1874490723000794
9. L. Pullagura, N. Kittad, G. Diwakar, V. Sathiya, A. Kumar and M. S. Yalawar, "ML based Parkinson’s Disease Identification using Gait Parameters," 2022 International Conference on Automation, Computing and Renewable Systems (ICACRS), Pudukkottai, India, 2022, pp. 561-566, doi: 10.1109/ICACRS55517.2022.10029281.
10. V. Suresh Kumar, Lokaiah Pullagura, Nalli Vinaya Kumari, S. Pooja Nayak, B. Padmini Devi, Adnan Alharbi, Simon Atuah Asakipaam, "Internet of Things-Based Patient Cradle System with an Android App for Baby Monitoring with Machine Learning", Wireless Communications and Mobile Computing, vol. 2022, Article ID 1140789, 11 pages, 2022. https://www.hindawi.com/journals/wcmc/2022/1140789/ https://doi.org/10.1155 /2022/ 1140789.
11. Rakesh Kancharla, Venkata Rao Maddumala, T. V. N. Prasanna, Lokaiah Pullagura, Ratna Raju Mukiri, M. Viju Prakash, "Flexural Behavior Performance of Reinforced Concrete Slabs Mixed with Nano- and Microsilica", Journal of Nanomaterials, vol. 2021, Article ID 1754325, 11 pages, 2021. https://www.hindawi.com/journals/jnm/2021/1754325/ https://doi.org/10.1155/ 2021/1754325.
12. Lokaiah Pullagura, Dr. Anil Kumar, “Analysis of Train Delay Prediction System based on Hybrid Model”, Journal Of Advanced Research in Dynamical & Control Systems, Vol. 12, 07SpecialIssue,2020.https://www.jardcs.org/abstract.php?id=5860 https://doi.org/ 10.5373/JARDCS/V12SP7/20202433.