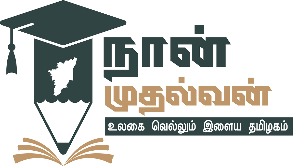
NAAN MUDHALVAN PROJECT REPORT



# PROJECT TITLE

# Web Page Summarization with NLP and Emotion Analysis

# PROBLEM STATEMENT

# Current web page summarization techniques often lack the ability to capture the emotional tone of the content, resulting in summaries that may fail to convey the intended sentiment and engagement level. There is a need for a summarization system that not only extracts key information but also analyzes the emotional context of the text, providing users with concise summaries that reflect the emotional nuances of the original content.

# OBJECTIVES

1. Develop a Natural Language Processing (NLP) algorithm capable of extracting key sentences and phrases from web pages to generate informative summaries.

2. Integrate emotion analysis techniques to identify and classify the emotional tone expressed within the text.

3. Design and implement a user-friendly interface or application allowing users to input web page URLs and receive summarized content enriched with emotional insights.

4. Evaluate the summarization system's performance in terms of accuracy, coherence, and emotional relevance compared to existing methods.

5. Explore potential applications of the system, such as enhancing user engagement, aiding in content curation, and improving information retrieval processes

# PROPOSED WORK

1. Web Scraping: Utilize web scraping techniques to extract the text content from the provided web page URLs.

2. Preprocessing: Clean the extracted text by removing HTML tags, special characters, and irrelevant content such as advertisements or navigation links.

3. Tokenization and Word Frequency Calculation: Tokenize the cleaned text into words and calculate the frequency of each word. This step helps identify the most common words in the text, which can be indicative of its main topics.

4. Sentence Frequency Calculation: Calculate the frequency of each sentence in the text based on the occurrence of important words. This step helps identify sentences that contain key information.

5. Summarization: Implement an extractive summarization technique (e.g., TextRank or TF-IDF) to select the most relevant sentences based on their frequency and importance. These sentences will form the summary of the web page content.

6. Sentiment Analysis: Utilize a sentiment analysis model (e.g., VADER or a trained deep learning model) to analyze the emotional tone of the summarized content. The sentiment analysis model will classify the summary into positive, negative, or neutral sentiment categories based on the polarity of the expressed emotions.

7. Output Presentation: Present the summarized content along with the sentiment analysis results in a user-friendly interface or application. Users can input web page URLs and receive summarized content enriched with emotional insights, helping them quickly grasp the main points and emotional tone of the original text.

8. Evaluation: Evaluate the performance of the summarization and sentiment analysis components in terms of accuracy, coherence, and emotional relevance. This evaluation can involve comparing the generated summaries and sentiment classifications against manually created summaries and human-perceived sentiments.

# IMPLEMENTATION

## 

## 

