# ML Problem Statement

## Mentorness Internship Program

**Problem Statement:**
News Classification and Analysis using Natural Language Processing

**Data Description:**
- Headline: The headline or title of the news article.
- Content: The textual content of the news article.
- Category: The category or topic of the news article, indicating the subject matter it covers.
- Date (of scraping): The date when the news article was scraped and added to the dataset.

**Background:**
The dataset comprises scraped news articles from various topics, sourced from inshorts.com, a news aggregation platform. With the rapid growth of digital media and news consumption, analyzing and classifying news articles have become crucial for understanding public discourse, tracking emerging trends, and monitoring events across different domains. Leveraging natural language processing (NLP) techniques, this project aims to classify news articles into predefined categories and extract valuable insights from the textual content.

**Objective:**
The objective of this internship project is to perform news classification and analysis using natural language processing (NLP) techniques to categorize news articles into predefined topics and extract actionable insights. By automating the classification process and analyzing news content, the project aims to facilitate efficient information retrieval and trend identification in the rapidly evolving news landscape.

**Key Components:**
1. Data Collection and Preprocessing: Understand the dataset of news articles from inshorts.com, preprocess the textual content by removing noise, such as special characters and stopwords, and tokenize the text for further analysis.

2. Topic Classification: Develop NLP classification models to categorize news articles into predefined topics or categories based on their headlines and content. Explore techniques such as text classification algorithms, including logistic regression, random forests, or deep learning architectures like convolutional neural networks (CNNs) or transformers (e.g., BERT).

3. Model Training and Evaluation: Train the classification models on the labeled dataset of news articles and evaluate their performance using appropriate evaluation metrics, such as accuracy, precision, recall, and F1-score. Fine-tune the models to improve classification accuracy and robustness.

4. Topic Analysis and Insights: Analyze the classified news articles to gain insights into the distribution of topics, emerging trends, and patterns in news coverage. Identify popular topics, recurring themes, and changes in public discourse over time.

5. Visualization and Reporting: Create visually informative representations, such as topic distribution plots, word clouds, and trend graphs, to present the findings. Generate comprehensive reports summarizing the analysis, insights, and recommendations for further research or decision-making.

**Expected Outcomes:**
- Trained NLP classification models capable of accurately categorizing news articles into predefined topics.
- Insights into the distribution of topics and trends in news coverage across different categories.
- Recommendations for improving news classification accuracy and enhancing understanding of public discourse.

**Deliverables:**
- Preprocessed dataset of news articles with labeled categories.
- Trained classification models with evaluation results.
- Report summarizing findings, insights, and recommendations for further analysis and application.

**Conclusion:**
This project aims to leverage natural language processing techniques to perform news classification and analysis, providing valuable insights into the distribution of topics and trends in news coverage. By automating the classification process and extracting actionable insights from news content, this project contributes to enhancing information retrieval and understanding in the dynamic news landscape.