# 

# **Data Visualization and Analysis**

# **Instructor:** Ma’am Eisha tur Razia

# **Project Proposal**

# **Section:** BDS 5B

## 

|  |  |
| --- | --- |
| **Roll No** | **Name** |
| 22L-7478 | Abdullah Maqsood |
| 22L-7465 | Omar Bandial |
| 22L-7541 | Eman Atiq |
| 22L-7555 | Ameer Tufail |

## 

# Comparative Linguistic Analysis on Narratives of Mainstream International Media Outlets

## Problem and data set description

In this project, we aim to collect articles from various international media outlets on specific topics to explore potential biases or differing narratives in their reporting. By analyzing the language and framing used across different news sources such as BBC, CNN, Al Jazeera, TRT World, and Fox News, we hope to identify patterns that indicate partiality or distinct editorial perspectives. These articles will cover the same events or topics, allowing for a transparent comparative analysis. The key is to observe how differences in tone, word choice, and structure may reveal underlying ideologies or regional influences within the media landscape.

The dataset will consist of several features extracted from each article, including the source, headline, description, timestamp, region, tags, author, article content, and more. These attributes will enable a detailed examination of how each outlet constructs its narrative. The analysis will attempt to uncover linguistic trends that highlight the subtle ways media sources shape their coverage. Through this, we aim to better understand how the same event can be interpreted differently depending on the outlet presenting it.

## Preliminary Ideas on How to Address the Problem

At this stage, we are considering a range of natural language processing (NLP) techniques and machine learning models to analyze the articles. First, we will preprocess the text by removing stopwords, punctuation, and performing tokenization and lemmatization to prepare the data for analysis. For feature extraction, we'll likely use Term Frequency-Inverse Document Frequency (TF-IDF) to represent the articles based on word importance.

For bias detection and sentiment analysis, we are interested in using pre-trained models like BERT, as it excels in understanding context and nuances in language. However, we may explore other sentiment analysis tools depending on the complexity of the data. When it comes to classifying articles based on their source or detecting bias, we plan to try several models, including Naive Bayes, Logistic Regression, and Decision Trees. Naive Bayes could be useful due to its efficiency in text classification, while Logistic Regression may provide more interpretability. Decision Trees, on the other hand, could help us identify key features, like specific words or phrases, that contribute to differences between outlets.

Ultimately, we will experiment with multiple models and techniques and evaluate their performance with different metrics. Our goal is to identify the best approach based on how well each model captures patterns in the data and helps us uncover differences in narrative style across media outlets.

## Software tools

We will use Python as the primary scripting language throughout the project. For web scraping and data extraction, we will use BeautifulSoup and Selenium to gather articles from different media outlets. For data manipulation and analysis, Pandas will be our main tool. To visualize trends and patterns, we will rely on Matplotlib and Seaborn. Additionally, for machine learning tasks such as text classification and model building, we will use Scikit-learn and some other relevant libraries depending on the models we explore.

## Expected Results and Evaluation

Our analysis aims to reveal distinct differences in how various international media outlets report on the same topics, highlighting potential biases in their narratives. By examining sentiment and language use, we aim to show which outlets frame issues more positively or negatively and identify specific words tied to each source. Visualizations will help present these findings in a clear and understandable way.

To evaluate our models, we will use metrics like accuracy, precision, recall, and F1 score, along with cross-validation to ensure reliable results. We’ll also focus on the interpretability of our models, especially Logistic Regression and Decision Trees, by analyzing the features that influence our predictions. Finally, we will manually review a sample of articles to confirm that our findings align with the actual content and provide an understanding of how different media outlets shape narratives around key topics.

## Preliminary Results and Data Set Explored

In our initial exploration of the dataset, we have collected articles from various international media outlets, including BBC, CNN, Al Jazeera, TRT World, and Fox News. Our dataset includes articles covering key topics and attributes such as source, headline, publication date, region, and article content. So far, we have analyzed a subset of this data to understand the general sentiment and language patterns.

Preliminary results indicate that different outlets often use varying tones and word choices when reporting on the same events. For instance, we observed that some sources tend to use more emotionally charged language, while others adopt a more neutral tone. This early analysis sets the stage for deeper investigations into biases and narrative styles, providing a foundation for our ongoing work.

## Outline of the Work-to-Do

1. **Data Collection**: Gather articles from selected international media outlets on specified topics.
2. **Data Preprocessing**: Clean and prepare the data, including text normalization and feature extraction.
3. **Exploratory Data Analysis (EDA)**: Conduct preliminary analysis to identify trends and patterns in language use and sentiment.
4. **Model Development**: Implement and evaluate various models, including Naive Bayes, Logistic Regression, and Decision Trees, for classification and bias detection.
5. **Results Analysis**: Analyze model performance using evaluation metrics and feature importance.
6. **Final Review**: Manually review a subset of articles to validate findings and ensure alignment with quantitative results.

## 

## References

<https://www.bbc.com/>

[https://edition.cnn.com](https://edition.cnn.com/)

[https://www.aljazeera.com](https://www.aljazeera.com/)

<https://www.trtworld.com/>

<https://www.foxnews.com/>