BanglaNewsClassifier: A Machine Learning Approach for News Classification in Bengali Newspapers.

Group: A21

Objective:

The main objective of this study is to critically analyze Bangla media on various issues including international sports, technology, country, economy, and entertainment by using large datasets collected through web scraping techniques. Perform in-depth content analysis to identify diversity, explore emotional differences expressed in media to identify possible emotional tones and biases, comparatively analyze categories to identify similarities, differences, its unique characteristics in terms of content, emotion, and linguistic structures linguistic features, and stylistic elements employed in Ngala news articles Investigate, and explore how objectivity and potential bias in news coverage across groups and locations. To achieve these objectives, this study seeks to provide a comprehensive, data-driven understanding of the Bangla media landscape, which helps support media studies, linguistics, and cultural research.

Design and Architecture Selection:

To address the diverse range of objectives, we employed an ensemble of machine learning models from the scikit-learn library:

To accomplish objectives, we used several machine learning examples from the scikit-learn library:

Decision Trees: Leveraging their ability to process statistical and classification data, decision trees have been used for content classification and sensitivity analysis tasks

Naive Bayes classifiers: Known for their simplicity and efficiency, naive Bayes models were used for text classification tasks, such as title identification and sentiment analysis.

Random Forests: Cluster models with multiple decision trees were used to improve the robustness and accuracy of content classification and sensitivity analysis predictions

This model was trained using a hierarchical approach to ensure a balanced set of categories and sensitivities in the training data. Cross-validation methods were used to check the model performance and prevent overfitting.

Data Collection

First, we collected data sets by scraping websites of Bangla newspapers using Selenium, a web automation tool. To prevent bots from being detected from the web, we used a random interval method during scraping. The data set contains news stories from six different categories: International, Sports, Technology, Country, Economy, and Entertainment.

Dataset Overview

The scraped data was saved in a CSV file, with 98,884 rows. The CSV file has two columns: one for the class and the other for the media description. Because newspaper articles tend to be high-quality, well-written articles, minimal data processing was necessary.

Data Preprocessing

Despite the fact that the newspaper articles are usually of high quality, we did some preprocessing to further improve their quality. First, we downloaded a collection of Bangla reference words from an Excel file. Then using the bangla-stemmer Python library, we extracted these stop words from the dataset. Additionally, we removed punctuation marks from the media to ensure the cleanliness of the text.

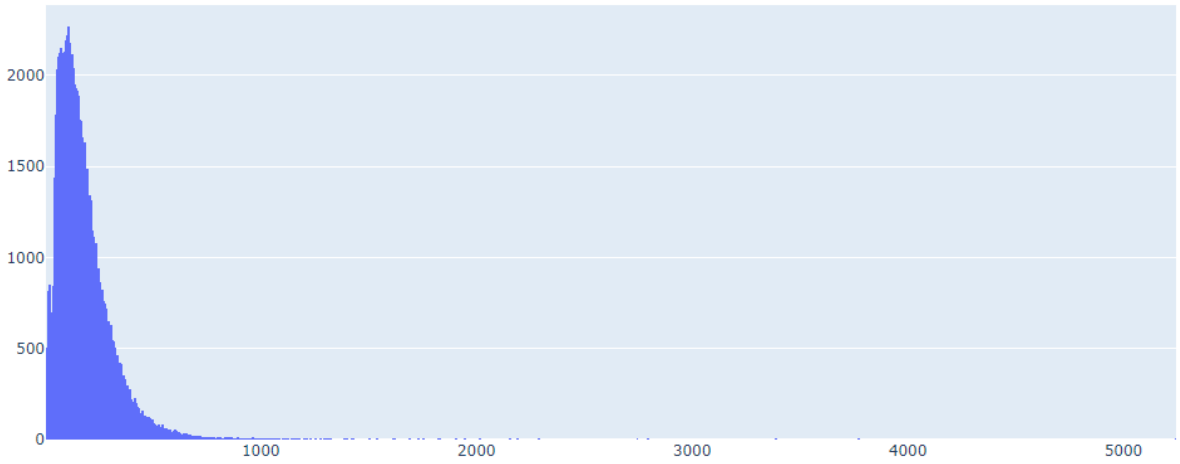
Exploratory Data Analysis:

In this subpart, we will explore the dataset's characteristics, such as:

1. **Category Distribution**: We Analyze the distribution of news articles across the six categories to understand the dataset's balance or skewness.

2. **Text Length Analysis**: Investigate the length of news descriptions by calculating summary statistics and visualize the distribution using histograms or box plots.

Text Length Histogram of news articles in the dataset



Text Preprocessing and Vectorization:

This subsection will cover additional steps for preprocessing text and converting text data into numerical representation for analysis or further modeling:

Tokenization: We then divided the news descriptions into individual words or tokens.

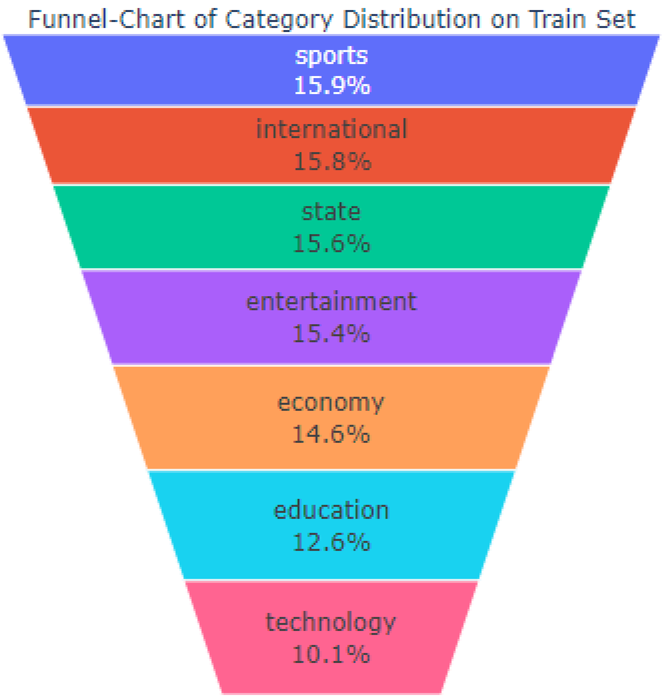
Stemming or Lemmatization: Also, word bases or root forms are reduced to eliminate inflections.

Vectorization: Preprocessed texture data were converted into numerical vectors using methods such as TF-IDF

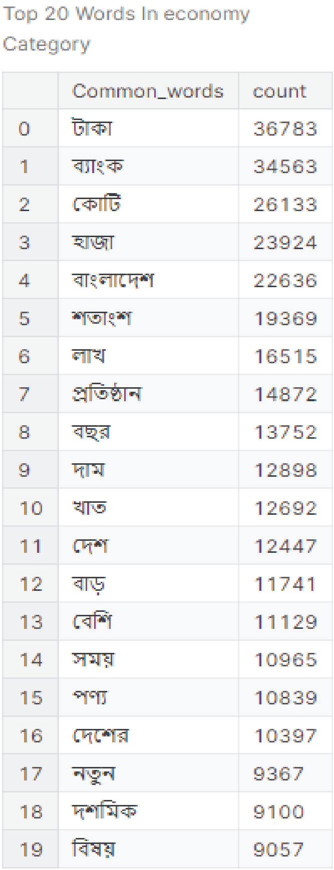
Exploratory Visualization:

In this subpart, we will visualize the preprocessed and vectorized data to gain deeper insights:

This is the distribution of our dataset based on the category. The dataset is pretty much close to balance that's why we did not try balancing our dataset.

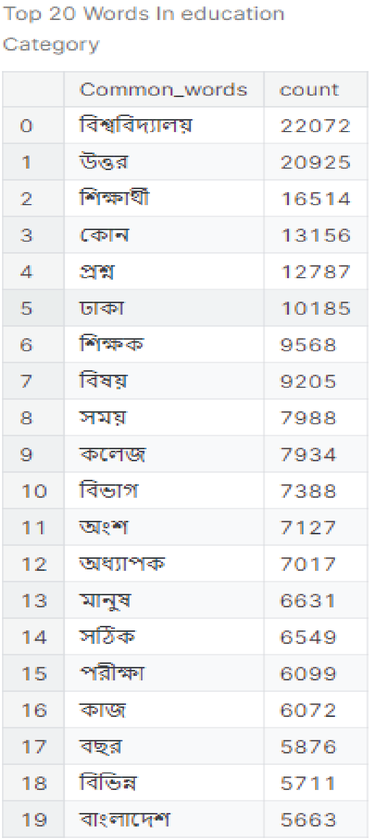


We have also done an analysis of the subcategories where we looked for the most frequently used words in the dataset.



Top 20 Words In economy Category

	Common_words	count
0	টাকা	36783
1	ব্যাংক	34563
2	কোটি	26133
3	হাজা	23924
4	বাংলাদেশ	22636
5	শতাংশ	19369
6	লাখ	16515
7	প্রতিষ্ঠান	14872
8	বছর	13752
9	দাম	12898
10	খাত	12692
11	দেশ	12447
12	বাড়	11741
13	বেশি	11129
14	সময়	10965
15	পণ্য	10839
16	দেশের	10397
17	নতুন	9367
18	দশমিক	9100
19	বিষয়	9057



Top 20 Words In education Category

	Common_words	count
0	বিশ্ববিদ্যালয়	22072
1	উত্তর	20925
2	শিক্ষার্থী	16514
3	কোন	13156
4	প্রশ্ন	12787
5	ঢাকা	10185
6	শিক্ষক	9568
7	বিষয়	9205
8	সময়	7988
9	কলেজ	7934
10	বিভাগ	7388
11	অংশ	7127
12	অধ্যাপক	7017
13	মানুষ	6631
14	সঠিক	6549
15	পরীক্ষা	6099
16	কাজ	6072
17	বছর	5876
18	বিভিন্ন	5711
19	বাংলাদেশ	5663