Bengali Newspaper Genre Classification

By
Zarin Tasnim Haider

Submitted by:
Zarin tasnim Haider
ID: 21301380
Course: Artificial Intelligence
Dept: CSE

Under the Guidance of:

Ipshita Bonhi Upoma

Co-ordinator of Bangladesh Mathematical Olympiad

Lecturer, BRAC University

Submitted to: BRAC University

Date of Submission: Jan, 2025

I. Introduction:

In this project, I am focusing on the challenging task of news text classification in Bengali, targeting distinct linguistic and cultural variations across different regions. Bengali, known for its rich diversity in dialects and idiomatic expressions, presents a unique challenge in accurately classifying news texts. This complexity becomes more pronounced when dealing with texts from various domains such as entertainment, sports, and national news, each characterised by its own stylistic and contextual nuances.

The motivation for this project is multifaceted. Firstly, I aim to address the significant gap in text classification technologies for Bengali, a language spoken by millions yet underrepresented in the realm of digital text analysis. The applications of my project are wide-ranging, from assisting media outlets in organizing content more effectively to aiding in the study of regional linguistic trends and cultural narratives.

I have chosen to employ a combination of advanced XGBoost, SVM, and Naive Bayes models for this task. This approach allows me to leverage the strengths of both modern and traditional NLP methods, aiming for a more comprehensive and accurate classification of news texts.

II. Dataset Description:

Source:

Link: https://www.kaggle.com/datasets/disisbig/bengali-news-dataset

Ref:

Bengali News Dataset. (2019, December 25). Kaggle. https://www.kaggle.com/datasets/disisbig/bengali-news-dataset

Description:

The dataset consists of 11,324 text samples, and the goal is to apply machine learning and deep learning models for genre classification. The genres to be predicted are: 'sports', 'state', 'kolkata', 'entertainment', 'international', and 'national'.

Dataset Overview:

Number of Samples: 11,324

Features: Textual data from news articles

Target Classes (Genres):

'sports'

'state'

'kolkata'

'entertainment'

'international'

'national'

Data Characteristics:

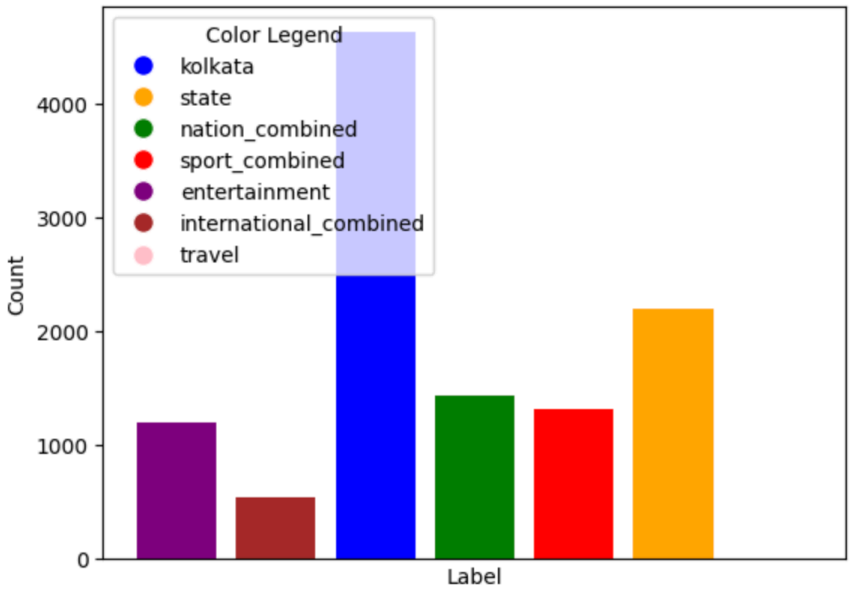
The dataset contains categorical text samples.

The number of samples in each class:

The dataset exhibits class imbalance, with varying sample sizes across different genres. Notably, the "Kolkata" class is significantly overrepresented, comprising 4,625 samples, while the "travel" class has a notably lower representation with only 1 sample.

Addressing the imbalanced nature of the dataset, particularly for the "Kolkata" and "travel" classes, may be essential for ensuring model robustness and generalization across all genres.

Label Name	**Number of samples**
sports	1,320
state	2,200
kolkata	4,625
entertainment	1,197
international	548
national	2,213
travel	1


Color Legend
kolkata
state
nation_combined
sport_combined
entertainment
international_combined
travel
Count
0
1000
2000
3000
4000
Label

III. Data Preprocessing:

Null Value Handling:

- Initial inspection revealed no null values in either the "Articles" or "Labels" columns.
- As a precautionary step, null instances were removed to maintain data integrity.

Label Consistency Resolution:

- Identified redundant labels within the "Labels" column, like "world" and "international" representing similar categories, along with "sport" and "sports," and "national" and "nation."
- Manually standardized labels to ensure consistency and uniformity, merging synonymous categories into singular representations:
- Unified "world" and "international" into a single label, "world."
- Harmonized "sport" and "sports" labels into one consistent representation, "sport."
- Amalgamated "national" and "nation" into a singular label, "national."

Test Data Incorporation:

- Appended the test data to the refined training dataset.
- Aimed to create a more comprehensive dataset, enriching the training data with diverse instances and ensuring a consistent label structure across all data.

Outcome for Further Analysis:

- Prepared the consolidated dataset for subsequent stages, such as feature engineering, model development, and evaluation.
- Ensured that the dataset exhibited a consistent and harmonized label structure for accurate classification during model training and testing.

IV. Data Splitting:

For the data splitting process, a 70-30 approach has been employed, allocating 70% of the dataset for training purposes and reserving the remaining 30% for testing model performance. The rationale behind this split is to strike a balance between an adequately sized training set for model learning and a substantial testing set for robust evaluation.

Additionally, the parameter random_state=42 has been set during the splitting process. This ensures reproducibility by generating the same random split each time the code is executed. Consistency in the training and testing subsets facilitates result comparison and aids in the reproducibility of experiments.

`

V. Model Training and Testing:

To achieve robust news text classification in Bengali, we undertook a systematic approach to model training, employing three distinct classifiers: Support Vector Machines (SVM), Naive Bayes, and the eXtreme Gradient Boosting (XGBoost) algorithm. The reason behind this is to harness the unique strengths of each classifier to enhance the overall performance in handling the linguistic differences and cultural variations prevalent in Bengali text related to news.

Support Vector Machines (SVM):

SVM is a powerful and versatile supervised machine learning algorithm that has proven effective in various text classification tasks. In this project, SVM is employed to discern patterns and boundaries within the feature space of Bengali news texts. By mapping the input data into a higher-dimensional space, SVM strives to find an optimal hyperplane that maximally separates different classes, making it well-suited for complex and non-linear classification problems. Through careful parameter tuning and kernel selection, we aimed to optimize the SVM model for the nuances present in diverse news domains. Now, applying this algorithm to the dataset we can see that the accuracy of this model is 0.8039735099337748. Which is acceptable for the requirements of this project.

Naive Bayes:

Naive Bayes classifiers are renowned for their simplicity and efficiency in text classification tasks. Leveraging probabilistic principles, Naive Bayes models assume independence between features, making them particularly effective in handling high-dimensional data such as word

occurrences in text. In the context of Bengali news text classification, Naive Bayes allows us to capture the likelihood of certain linguistic patterns or vocabulary usage associated with different news categories. This simplicity is advantageous, especially in scenarios where computational resources may be a constraint. Using this approach on the dataset, we can now observe that the model's accuracy is 0.44388849177984274. This is not suitable given the project's objectives.

eXtreme Gradient Boosting (XGBoost):

XGBoost is a state-of-the-art ensemble learning algorithm that has demonstrated exceptional performance in various machine learning competitions. By constructing an ensemble of weak learners (typically decision trees) in a boosting framework, XGBoost iteratively refines its predictions, focusing on instances where previous models have underperformed. For this Bengali news text classification, XGBoost provides the flexibility to capture intricate dependencies and interactions within the linguistic features, enhancing the model's ability to generalize across diverse news domains. When we apply this approach to the dataset, we can observe that the model's accuracy is 0.7713024282560706. Which is appropriate for the project's requirements.

VI. Model Selection/Comparison Analysis:

SVM Classifier:

- Accuracy:79%

- Insights: The Support Vector Machine (SVM) model outperforms Naive Bayes with significantly higher accuracy. It showcases reasonably good precision and recall scores for certain categories like "kolkata," "state," and "sports," indicating better classification capabilities in these classes. However, some categories such as "international" and "nation" exhibit low precision and recall, suggesting difficulties in accurately predicting sentiments for these categories.

	Precision	Recall	F1-Score	Support
entertainment	0.81	0.83	0.82	249
international	0.56	0.35	0.43	91
kolkata	0.85	0.92	0.88	956
nation	0	0	0	2
national	0.7	0.69	0.69	273
sports	0.9	0.81	0.85	243
state	0.75	0.72	0.74	445
Accuracy	0.8		0.8	2265

Macro Avg	0.51	0.48	0.49	2265
Weighted Avg	0.8	0.8	0.8	2265

Naive Bayes:

Accuracy:52%

- Insights: The Naive Bayes model demonstrates a relatively low accuracy in sentiment classification across various categories. It particularly struggles with precision, recall, and F1-scores in multiple categories, showcasing limitations in accurately identifying and classifying sentiments, especially for minority classes. The confusion matrix highlights challenges in effectively predicting diverse sentiment categories, with notable misclassifications in several classes.

Class	Precision	Recall	F1-Score	Support
entertainment	0.96	0.15	0.27	286
international	0.5	0.01	0.01	374
kolkata	0.98	0.18	0.31	331
national	0.42	1	0.59	1128
sports	0.7	0.01	0.02	553
state	0	0	0	124
Accuracy	0		0.52	2798

Macro Avg	0.51	0.19	0.17	2798
Weighted Avg	0.59	0524	0.31	2798

XGBoost Classifier:

- Accuracy: 80%

- Insights: The XGBoost model demonstrates performance between Naive Bayes and SVM. While its accuracy is closer to SVM, it shows varying precision and recall scores across different categories. Similarly to SVM, it exhibits challenges in predicting sentiments accurately for categories like "international," "nation," and "world," as indicated by low precision and recall scores.

Class	Precision	Recall	F1-Score	Support
entertainment	0.8	0.8	0.8	249
international	0.55	0.19	0.28	91
kolkata	0.78	0.94	0.85	956
national	0.71	0.55	0.62	273
sports	0.85	0.7	0.77	243
state	0.74	0.7	0.72	445
accuracy			0.77	2265
macro avg	0.49	0.43	0.45	2265

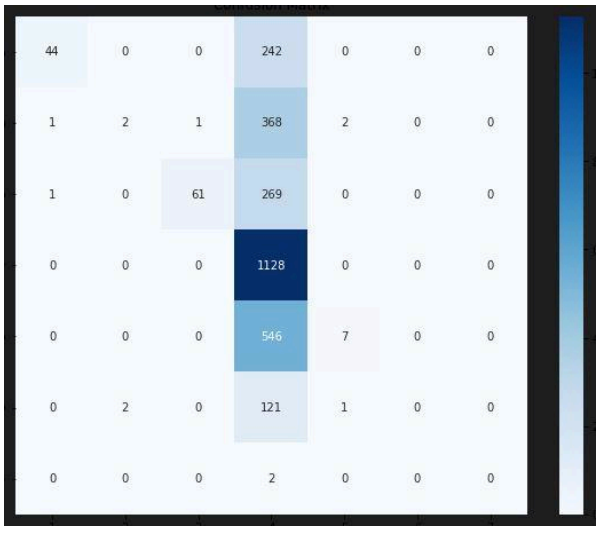
weighted avg	0.76	0.77	0.76	2265

Comparative Insights:

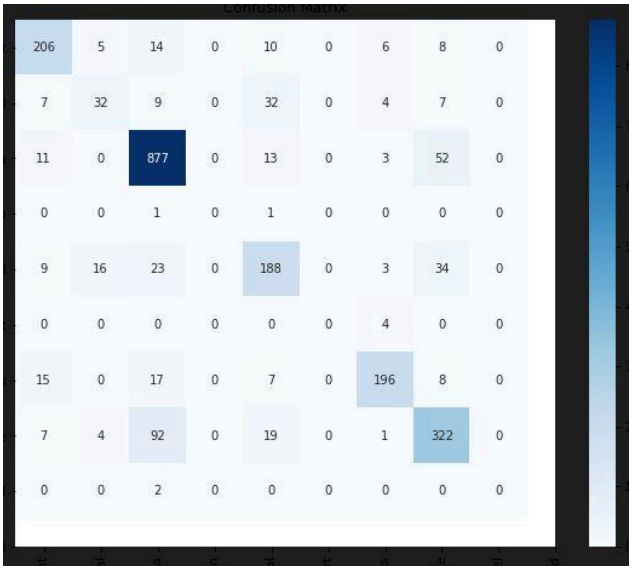
- The SVM model showcases the highest overall accuracy among the three models, followed by XGBoost and then Naive Bayes.

- SVM and XGBoost exhibit similar challenges in predicting sentiments for certain categories, especially "international" and "nation," where both models display lower precision and recall scores.

- Naive Bayes struggles the most in accurately classifying sentiments across all categories, especially for minority classes, indicating limitations in its predictive capabilities for sentiment analysis tasks.



Confusion Matrix of SVM Classifier



Confusion Matrix of Naive Bayes

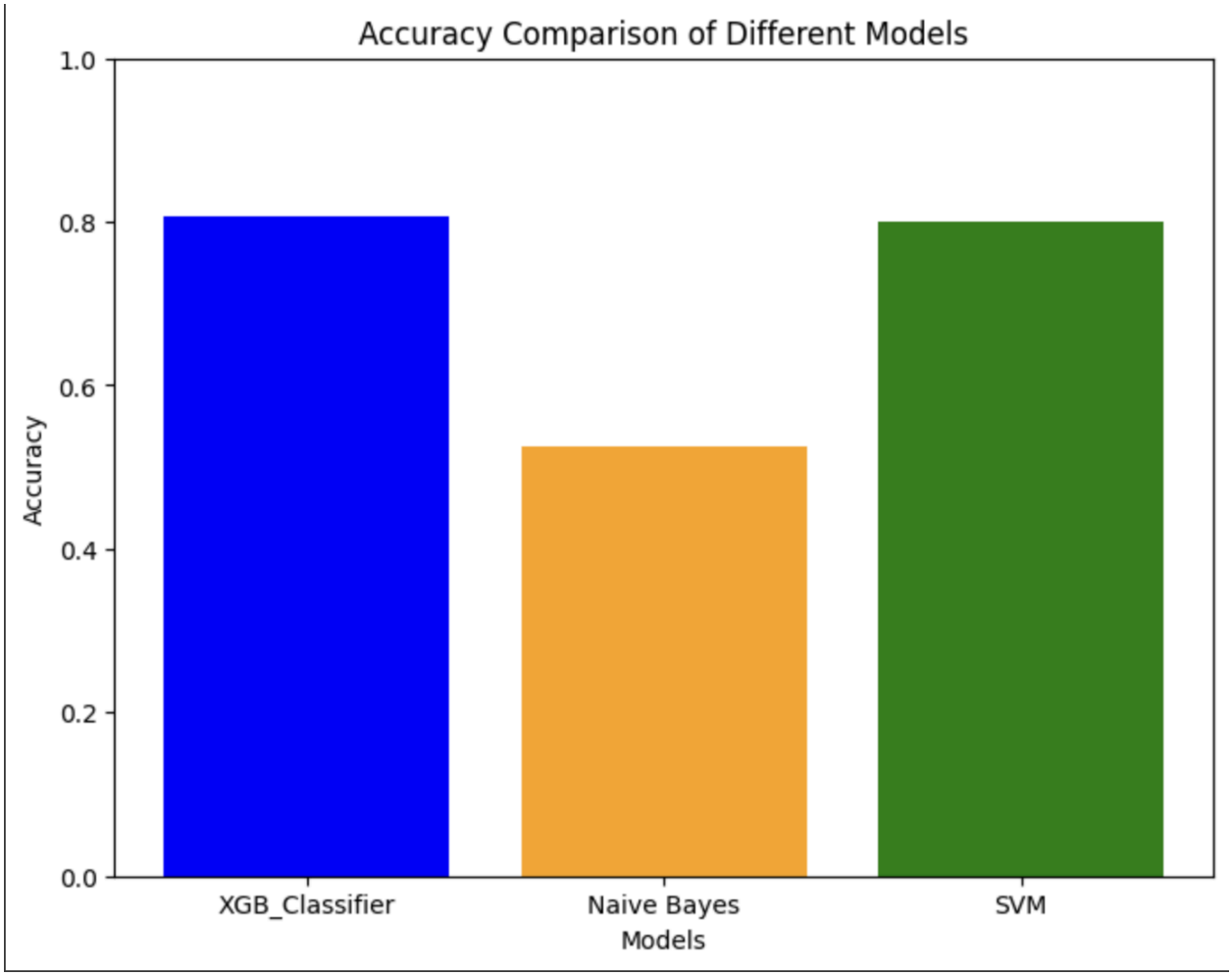
Model Recommendation:

- Considering Performance: The SVM model demonstrates the highest overall accuracy and relatively balanced precision and recall scores across multiple categories, making it a strong candidate for sentiment analysis in this context.

- Room for Improvement:Further investigation into misclassification patterns, additional feature engineering, or hyperparameter tuning might enhance the models' performance, especially for categories with lower precision and recall scores.

Confusion Matrix of XGB_Classifier

In summary, based on the provided results, the SVM classifier emerges as the most favorable model for sentiment analysis due to its higher accuracy and relatively balanced predictive performance across various sentiment categories.



Accuracy Comparison

VII. **Conclusion:**

In conclusion, my endeavor in Bengali news text classification unveiled valuable insights into model performances for sentiment analysis. I navigated the complexities of Bengali linguistics and cultural nuances by employing SVM, Naive Bayes, and XGBoost classifiers. While Naive Bayes showcased simplicity, it struggled with overall accuracy

and minority class predictions. In contrast, SVM displayed superior accuracy with balanced precision and recall, making it an optimal choice. XGBoost showed promise but encountered challenges akin to SVM. My findings affirm SVM as the frontrunner due to its superior accuracy and consistent performance across diverse categories. Further refinement through feature engineering or hyperparameter tuning could fortify model robustness. This study not only contributes to Bengali NLP but also underscores the significance of tailored approaches for accurate sentiment analysis in diverse linguistic landscapes.