GROUP 5

NEWSNET PUBLISHING COMPANY

- OLIVE MULOMA
- ABIGAIL MWENDWA
- HAWKINS MURITHI
- HARRY ATULAH

Table of contents

01

OVERVIEW

In today's digital age, news consumption has skyrocketed, and with it, the challenge of efficiently managing and categorizing vast amounts of content. Automated text classification offers a solution by enabling efficient sorting of news articles into relevant categories such as Politics, Technology, Entertainment, and Business. This enhances user experience by ensuring that readers have quick access to content that interests them and assists in organizing and managing content for media companies.

02

BUSINESS UNDERSTANDING

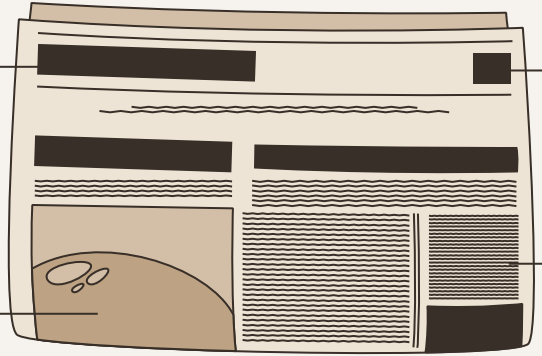
STAKEHOLDERS

Editorial Team

Responsible for reviewing and publishing news articles.

Authors

Provide news articles to the publishing company.



Content Management Team

Oversee the workflow of news content from submission to publication.

Customers/Readers

Consume the news content provided by the publishing company.

BUSINESS OBJECTIVES

- ★ **Develop an Automated Categorization Model:**Build a robust machine learning model to accurately classify news articles into Politics, Technology, Entertainment, or Business.

- ★ **Enhance Editorial Efficiency:**Automate article categorization to reduce the time and effort needed by editors, allowing them to concentrate on improving content quality.

- ★ **Improve Accuracy:**Minimize misclassification rates to ensure each article is correctly reviewed by the relevant editorial team, enhancing content relevance and quality.

BUSINESS PROBLEM

Newsnet Publishing receives a large volume of news content from various authors on a daily basis. Currently, the manual process of categorizing each article into one of four categories Politics, Technology, Entertainment, and Business is time-consuming and prone to errors.

This inefficiency leads to delays in publication, increased workload for editors, and potential misclassification of articles, which affects content quality and reader engagement.

03

DATA UNDERSTANDING

DATASET

The training dataset (`data_train`) contains 7,628 rows and 2 columns:

★ STORY: This column holds the text of news articles, which will be used to train the classification model.

★ SECTION: This column contains numerical labels representing different categories (e.g., 0 for Politics, 1 for Technology) that serve as the target labels for classification.

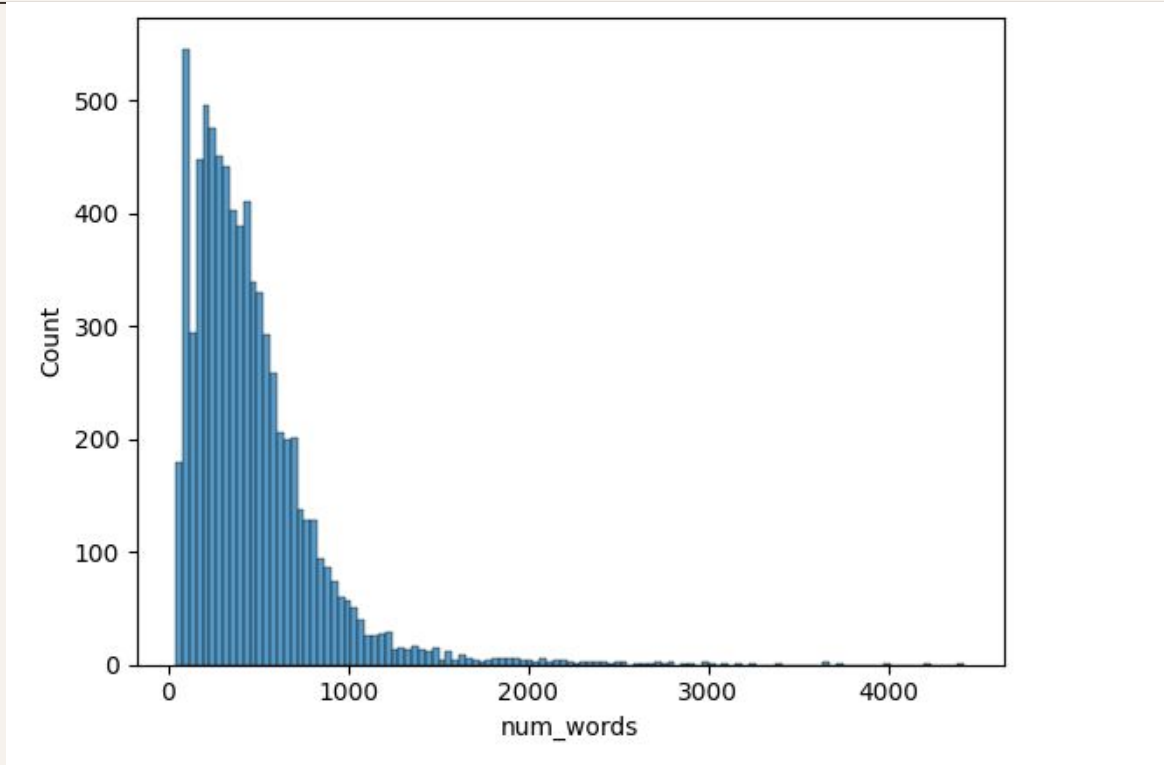
The test dataset (`data_test`) has 2,748 rows and 1 column:

★ STORY: This column contains news article texts similar to the training set but lacks associated category labels. It is used to evaluate the performance of the trained model by predicting the categories for these articles.

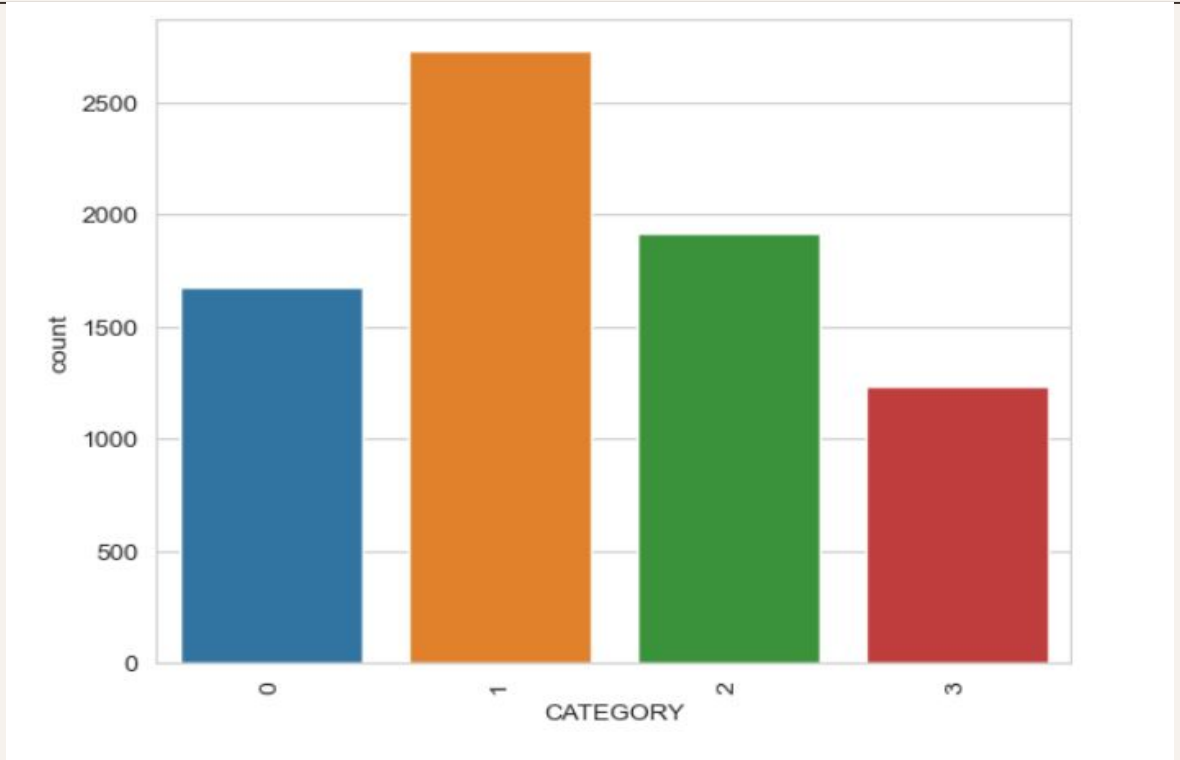
DATA ANALYSIS

- ★ Data analysis was done to prepare the dataset for effective model training and evaluation.
- ★ The data cleaning process identified and removed 77 duplicate rows from the training data, ensuring the uniqueness of entries.
- ★ The columns were then renamed and new labels were added to better reflect the data's content. Text preprocessing involved converting all text to lowercase, removing punctuation, and filtering out stopwords, followed by tokenization and lemmatization to standardize the text data. This resulted in a refined column of processed text.
- ★ EDA provided insights into the distribution of article lengths and category frequencies, highlighting class imbalances and the lack of strong correlation between text length and category. Visualizations were used to understand the data better, revealing patterns and the prominence of key terms in each category.

Distribution of article length

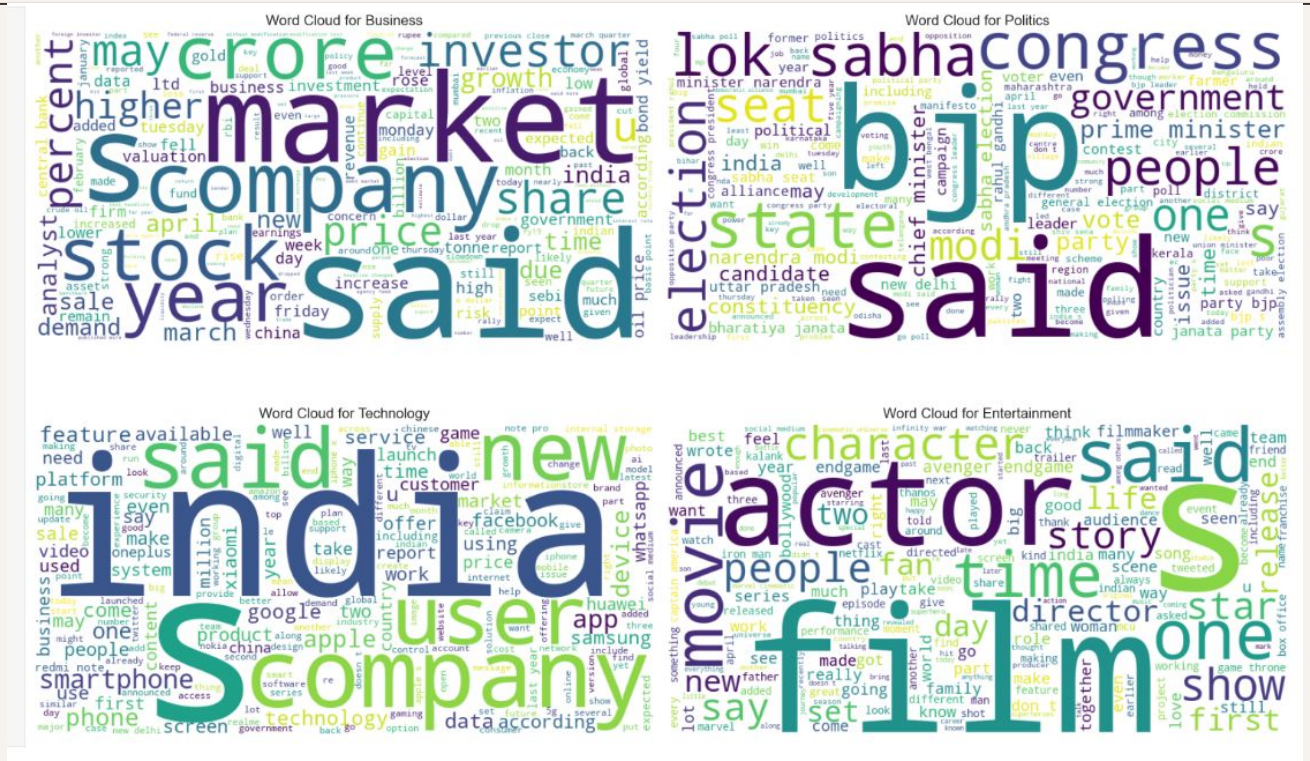


Distribution of categories



Text Length vs. Category

Word Cloud


Word Cloud for Business
market
company
said
stock
year
share
price
crore
investor
higher
percent
Word Cloud for Politics
said
bjp
congress
election
lok
sabha
government
people
state
seat
minister
Word Cloud for Technology
india
company
said
new
user
Word Cloud for Entertainment
film
actor
movie
said
character
time
one
people
story

04

MODELLING

Word2Vec with Logistic Regression model

We used the Word2Vec to create detailed word representations based on their usage in the text. These word vectors help capture the meaning and context of words more effectively.

The Logistic Regression model then used these word vectors to classify news articles into different categories.

This approach achieved a high accuracy of 94%, reflecting its ability to understand and categorize text based on nuanced word meanings.

TF-IDF Vectorization and Multinomial Naive Bayes model

We used TF-IDF Vectorization to transform text into numerical features by evaluating the importance of each word in relation to the entire document collection.

The Multinomial Naive Bayes model uses these features to classify text.

This combination also achieved a high accuracy of 93%, demonstrating its effectiveness in handling text classification by focusing on word frequency and relevance.

05

EVALUATION

Logistic Regression model

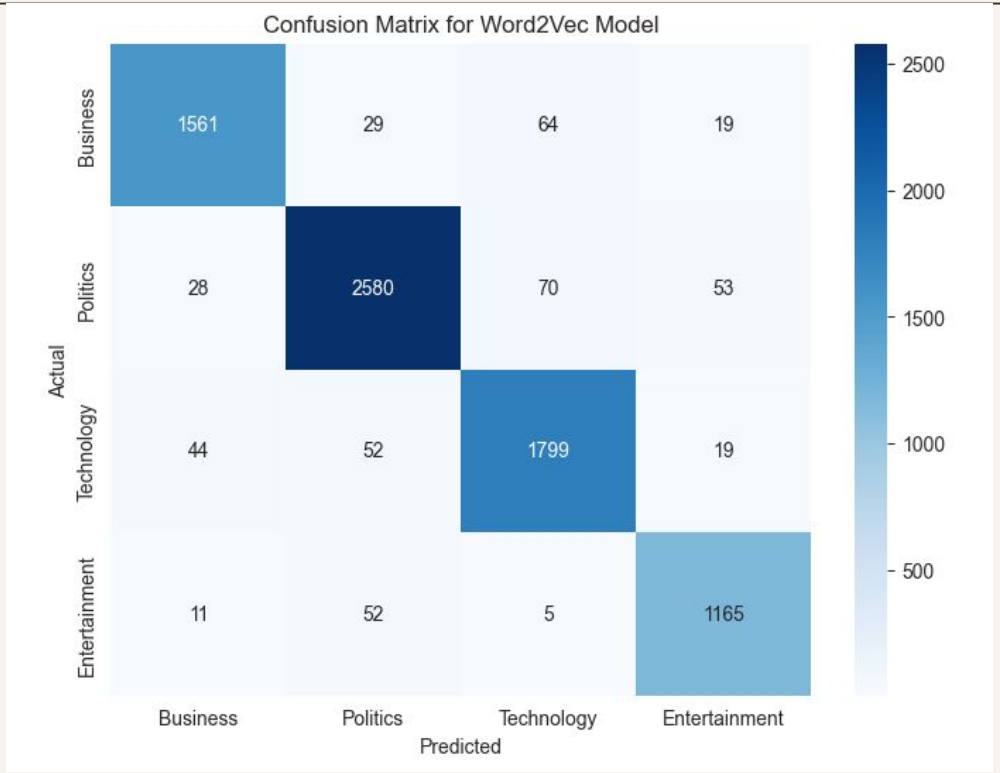
The Logistic Regression model, trained on Word2Vec embeddings, achieved an overall accuracy of 94%.

Precision and recall were high across all classes, with a slight imbalance in the number of instances per class.

Class 2 had the lowest F1-score at 0.93 due to a marginally lower recall. The model effectively classified most instances but struggled slightly with distinguishing between Politics and Technology.

Overall, the model performs well, though further improvements could enhance the precision and recall for Class 2.

Logistic Regression model



Multinomial Naive Bayes model

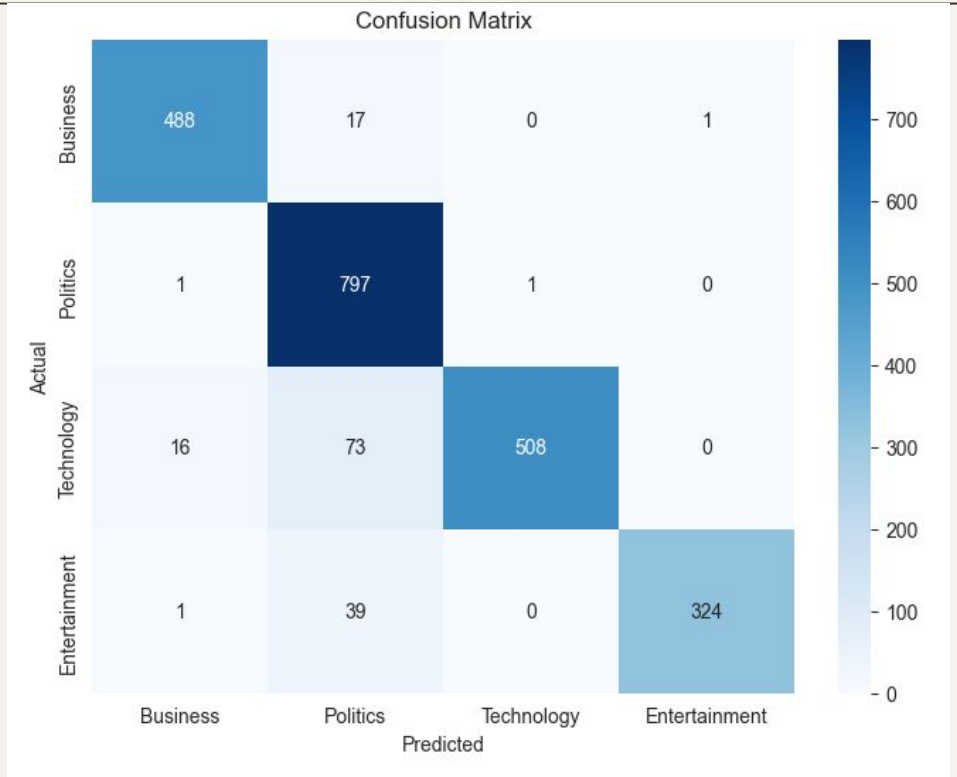
The Multinomial Naive Bayes (MultinomialNB) model achieved an overall accuracy of 93%.

The classification report indicated high precision and recall for most classes, but Class 2 (Technology) had a notably lower recall, meaning that many actual instances of this class were not correctly identified.

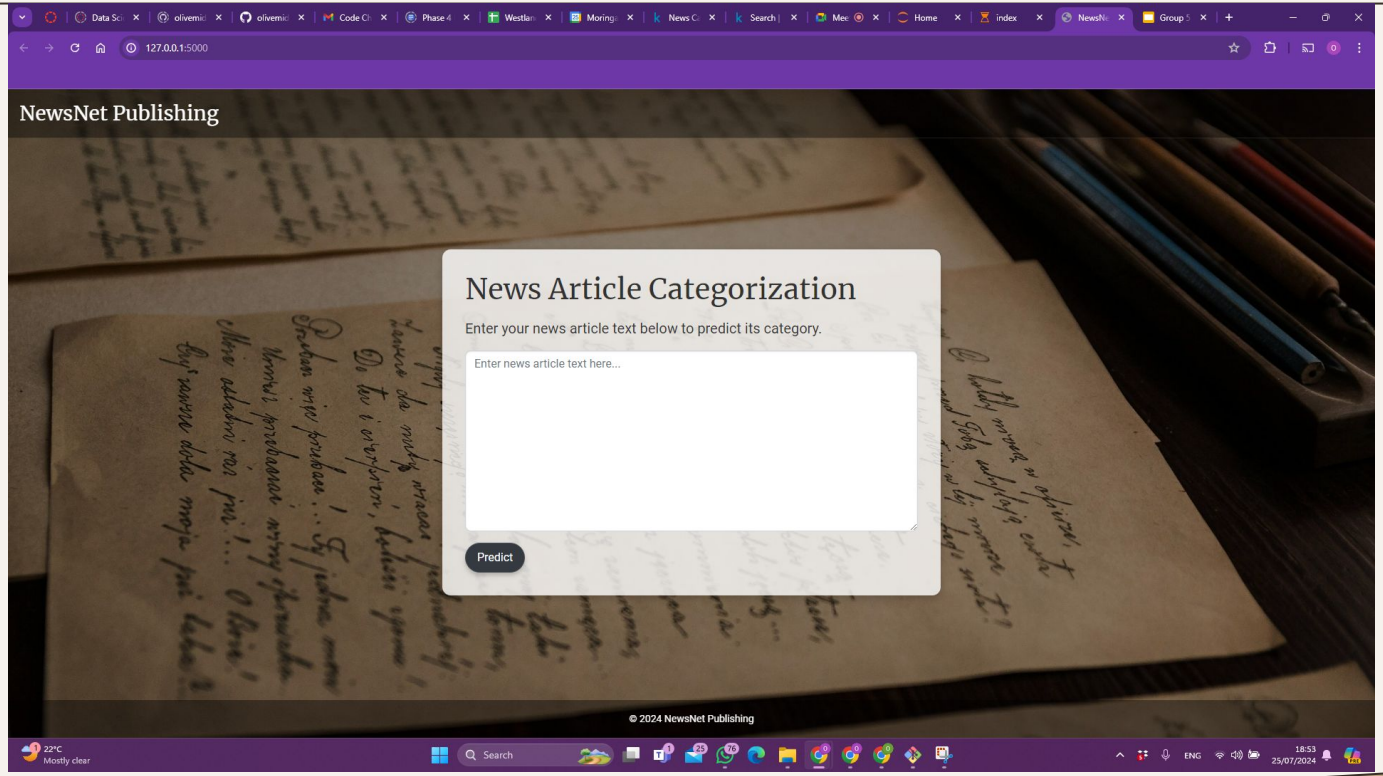
The F1-score, which balances precision and recall, was also lower for Class 2, highlighting an area for potential improvement.

The confusion matrix revealed that while the model was generally effective, it struggled with distinguishing between certain categories, particularly Technology and Entertainment, with some misclassifications noted across categories.

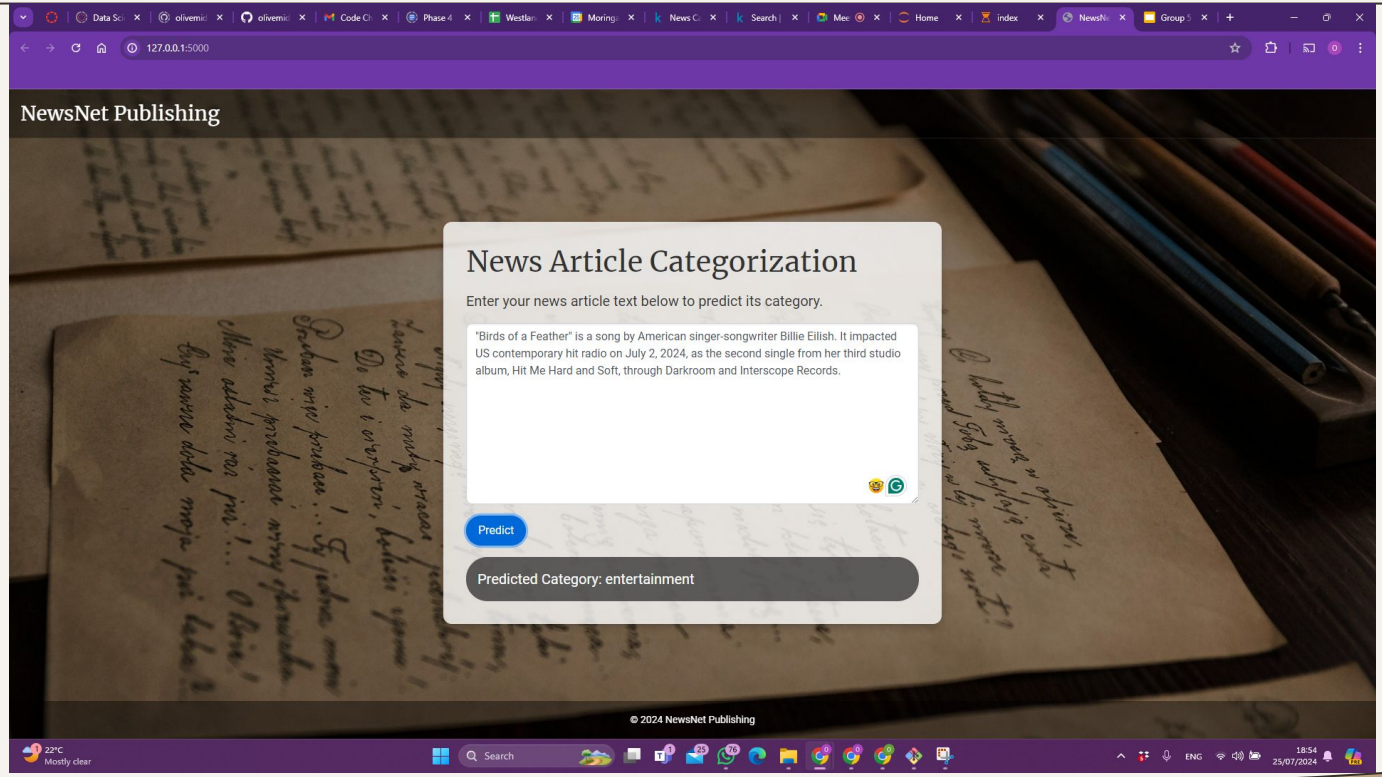
Multinomial Naive Bayes model



DEPLOYMENT


127.0.0.1:5000
NewsNet Publishing
News Article Categorization
Enter your news article text below to predict its category.
Enter news article text here...
Predict
© 2024 NewsNet Publishing

DEPLOYMENT


127.0.0.1:5000
NewsNet Publishing
News Article Categorization
Enter your news article text below to predict its category.
"Birds of a Feather" is a song by American singer-songwriter Billie Eilish. It impacted US contemporary hit radio on July 2, 2024, as the second single from her third studio album, Hit Me Hard and Soft, through Darkroom and Interscope Records.
Predict
Predicted Category: entertainment
© 2024 NewsNet Publishing

06

RECOMMENDATION

RECOMMENDATIONS

- ★ **Write on Trending Topics:** Focus on high-interest areas to attract readers.
- ★ **Incorporate Feedback:** Improve writing based on reader feedback.
- ★ **Prioritize High-Impact Articles:** Use categorization to streamline content review.
- ★ **Update Guidelines:** Align with reader preferences and trends.
- ★ **Organize Content Efficiently:** Use advanced categorization for better management.
- ★ **Personalize Content Delivery:** Enhance reader engagement with personalized recommendations.
- ★ **Offer Personalized Recommendations:** Enhance user experience with tailored content.
- ★ **Improve Discoverability:** Help readers find relevant articles quickly.

THANK YOU!

Do you have any questions?