#### 

****

**AMERICAN UNIVERSITY OF BEIRUT**

**OLAYAN SCHOOL OF BUSINESS**

**MSBA 315 – Machine Learning and Predictive Analytics -- Spring 2024**

**Final Year Project**

**Presented to: Dr. Wael Khreich**

**Team Members Group 2:**

**Ali Annan 202475973**

**Kinan Morad 202471895**

**Sasha Nasser 202473486**

**Romanos Rizk 202471561**

**Rita Salloum 202371596**

#### 

**Table of Contents**

[**Abstract 2**](#_th98jru1gm9v)

[**Introduction 3**](#_uz6cfnoikxgd)

[**Literature Review 3**](#_ex26s3eyi5rr)

[**Methodology 5**](#_5gx0pr9w3pny)

[Preprocessing and Feature Engineering 5](#_s0uvvx8hgm0p)

[Pipeline 7](#_6ry9bcn6r3ut)

[Models 8](#_enu67ahmr1lj)

[Random Forest Classifier 8](#_bph3gm4ttvvp)

[Logistic Regression 9](#_mo7pmr2ptc9m)

[KNN 9](#_9hjpv89lpm99)

[XGBoost 9](#_mvtdyhe3xdas)

[Neural Network 9](#_pyv0zkeme2rv)

[Utilizing BERT for Deep Learning 10](#_1bk28cgzw5c6)

[Best Model and Comparative Analysis with Related Work 11](#_w2stg9t9jn5r)

[**Results & Discussions 11**](#_6ka19kozqj11)

[Answering the research question 12](#_w1a6s41sgqnm)

[Error Analysis 12](#_ew14opgjrbqz)

[Obstacles and Solutions in Model Development 13](#_9jadrpiqvu9r)

[**Conclusion 13**](#_ies2431vessb)

[Summary of Solution, Findings, and Recommendations: 13](#_3pve4lg5q4qk)

[Data Handling, Business Recommendations, Strategic Insights, and Future Work 14](#_ha34d3cwb6ro)

[**Appendices 15**](#_bgpwb537aa7m)

[**Reference 26**](#_wad983bz5n1j)

# 

# Abstract

This report presents the development and deployment of a sophisticated text classification system using a comprehensive dataset of articles from The New York Times spanning from January 1, 2000, to the present day. The dataset, titled "The New York Times Articles Metadata," encompasses over 2.1 million articles, forming the cornerstone of our investigation aimed at understanding and categorizing news articles based on their section names. Initial data preprocessing involved removing columns with less than 10% null values, including abstract, snippet, headline, and keywords, while extraneous columns like 'web\_url,' 'print\_section,' and 'source' were discarded to streamline the dataset. A selection of 24 relevant section names was identified for focused analysis. The clean dataset was then split into a training set and a validation set using a 70/30 ratio to facilitate comprehensive model learning and evaluation. Text columns such as headlines and keywords were transformed into dictionary objects and aggregated into a 'combined\_text' column, simplifying their usage in the model. Preprocessing steps included converting all text to lowercase, removing special characters, numeric values, stop words, and performing stemming. The 'combined\_text' underwent TF-IDF vectorization to convert it into a suitable numerical representation for machine learning algorithms. Recursive feature elimination was employed for feature selection, identifying the most influential features for classification. Several classification models, including Logistic Regression, Random Forest Classifier, KNN, Gradient Boosting Classifier, Sequential Neural Network and Bert, were evaluated using measures such as accuracy and confusion matrices. The Bert model demonstrated superior performance across multiple evaluation metrics, emerging as the most effective classifier. This project not only enhances our understanding of content patterns and writing styles employed by The New York Times but also contributes valuable insights to the broader field of text classification by showcasing effective methodologies and advanced machine learning techniques applied to a large and intricate real-world dataset.

# 

# 

# 

# Introduction

The news industry has undergone substantial changes in the digital age, transitioning from traditional print to primarily online platforms, resulting in a global shift in the way information is received. The New York Times has attracted a diversified readership by providing comprehensive coverage and addressing a wide range of topics. This move has not only enhanced the convenience of access but also heightened the intricacy of comprehending reader preferences and the dynamics of news consumption.

As The New York Times shifted its content to the internet, the number of articles produced each day increased significantly, creating a dataset that contains a wide range of information. The collection titled 'New York Times Items Metadata' contains more than 2.1 million items published since the year 2000. This dataset offers a chance to examine patterns, preferences, and modifications in news coverage over twenty years. Due to the extensive size and variety of the material, it is essential to correctly categorize and analyze these pieces to effectively serve and expand the reader base. Machine learning is utilized to classify articles based on their content, providing an effective tool for this purpose. This project seeks to utilize advanced algorithms to predict the classification of articles into certain sections of The New York Times. The goal is to improve the discoverability of information and increase reader engagement. The research utilizes advanced methods such as natural language processing (NLP), feature engineering, and several machine learning models, including Random Forest, Gradient Boosting classifiers, and BERT, to assess and forecast the sections of articles based on their textual content. By optimizing the classification process, The New York Times can maintain its position as a frontrunner in providing personalized content in an ever-growing digital landscape. How can machine learning and NLP techniques be optimized to accurately predict the classification of articles into their respective sections, to improve the relevance and accessibility of content for diverse audiences, considering the dynamic nature of news topics and the varied interests of readers?

# Literature Review

The study conducted by Jeelani Ahmed and Muqeem Ahmed, titled "Online News Classification Using Machine Learning Techniques, investigates the use of several machine learning models to classify online news items. Their objective is to improve how organizations explore, analyze, and manage the large quantities of unstructured data accessible on the internet, emphasizing news categorization. They highlight that 90% of the data available on the internet is in an unorganized manner and underscore the importance of classification in transforming this data into organized and valuable information. Notably, a Bayesian classifier was used, resulting in an accuracy rate of 93%. The notable degree of precision demonstrated underscores the promise of machine learning methodologies in the domain of text categorization in the context of online news material.

In their scholarly study titled "Article Classification using Natural Language Processing and Machine Learning," Dien, Loc, and Nguyen Thai-Nghe explore the creation of an automated system that employs machine learning methodologies to categorize articles into pertinent subjects. The study employed Support Vector Machines (SVM), Naïve Bayes, and k-Nearest Neighbors (kNN) algorithms to analyze two datasets of articles. The results demonstrated significant accuracy, notably with SVM, which achieved a classification rate above 91% for scientific articles. The research emphasizes the significance of preliminary data processing, normalization of words, elimination of stop words, and subsequent vectorization utilizing TF-IDF.

In their scholarly essay titled "Twenty Years of Machine-Learning-Based Text Classification: A Systematic Review," Palanivinayagam, El-Bayeh, and Dama̡evi̯ius present a thorough examination of the use of machine learning techniques in the field of text categorization spanning twenty years. This study examines a total of 224 scholarly articles, with a specific emphasis on the development and utilization of different machine learning models within the field. The analysis highlights the adoption of SVM, Naive Bayes, and k-Nearest Neighbors as prominent models. The SVM reached high accuracy rates, which reached as high as 98.88% on some datasets. The paper examines significant obstacles encountered in the field of text classification, including the issues of feature selection and overfitting. The review utilized a systematic methodology, adhering to the PRISMA principles, to provide a methodical examination.

The paper titled "SrBERT: automatic article classification model for systematic review using BERT" authored by Sungmin Aum and Seon Choe (2021) presents a novel methodology for article classification in the context of systematic reviews. This approach utilizes the srBERT model, which incorporates the BERT algorithm. The present study aims to mitigate the labor-intensive nature of systematic review procedures by the implementation of automated article classification, thereby improving efficiency and accuracy.

Disayiram and Rupasingha (2022) conducted a comparative study on the classification of English news articles using machine learning algorithms. The study focused on categorizing news items into domain categories such as business, health, politics, technology, and sports. The researchers conducted experiments on five distinct machine learning algorithms: Decision Tree, Random Forest, Naïve Bayes, Multilayer Perceptron (MLP), and SVM. The SVM model demonstrated exceptional performance. However, when employed with an ensemble learning technique, the classification accuracy increased to 92.75%, indicating a more sophisticated approach with a reduced error rate in comparison to the performance of individual algorithms. Furthermore, the researchers evaluated different algorithms using the same dataset to determine the most efficient approaches for classification tasks.

# Methodology

## Preprocessing and Feature Engineering

This dataset titled "NYT Articles: 2.1M + (2000-Present) Daily Updated" was extracted from Kaggle and features over 2.1 million articles from The New York Times, spanning from January 1, 2000, to the present. It is updated daily and covers a wide range of topics. Furthermore, it includes certain metadata such as the abstract, web URL, snippet, print page, print section, source, multimedia, lead paragraph, headlines, publication dates, keywords, document type, news desk, subsection name, section name, and byline. Due to the dataset being too large, we decided to take a subset of 10,000 rows.

We conducted Exploratory Data Analysis (EDA) to gain insight into the distributions of our data collection. The Appendices contain all the visuals produced throughout the project. Our exploratory data analysis provided insightful revelations about the trends of news content over the years. The analysis of frequent words in headlines revealed that words like "new", "paid", "notice", and "deaths" dominate. The distribution of articles across different sections showed that 'Business Day', 'U.S.', and 'World' are the most prolific sections, suggesting robust coverage in these areas. In contrast, sections like 'College', 'Parenting', and 'Climate' had significantly fewer articles, indicating more niche or specialized coverage. Within subsections, 'Arts' was the most populated, followed by significantly lesser counts in 'Food', 'Travel', and 'Sports'. A heatmap of articles by year and section highlighted how certain sections have fluctuated over time. For example, articles in the 'Fashion & Style' and 'Real Estate' sections saw peaks and troughs, which could correlate with economic cycles and changing public interests. The distribution of word counts over the years, particularly within the U.S. section, showed a general consistency with occasional outliers. A heatmap of articles by print section and page number revealed that certain sections consistently appear on specific pages. The vast majority of the content was categorized under 'articles', with a minimal presence of 'multimedia' and other forms.

We started with Null Value Handling, columns with less than 10% null values, such as the abstract and keywords columns, had these null values eliminated. The subsection name column had more than 70% null values, it was dropped. In addition to that, the ‘print\_section’ and ‘print\_page’ columns were also dropped as they had almost 35% of null values. Our investigation focused exclusively on textual data to leverage natural language processing tools. The initial preprocessing stage consisted of eliminating irrelevant columns, such as 'web\_url', ‘pub\_date’, 'source', 'multimedia', 'document\_type', 'news\_desk', 'byline', 'type\_of\_material', '\_id', ‘subsection\_name’, ‘word\_count’ and 'uri'', which do not provide significant information for text categorization purposes. Furthermore, the 'pub\_date' column, since the models tried, would not fit and take datetime objects as input.

After careful consideration, we selected 24 section names that aligned and made sense relative to our data-cleaning process. The section names are: 'Arts', 'Automobiles', 'Blogs', 'Books', 'Business Day', 'College', 'Climate', 'Education', 'Fashion & Style', 'Food', 'Health', 'Home & Garden', 'Job Market', 'Movies', 'Parenting', 'Podcasts', 'Real Estate', 'Science', 'Sports', 'Technology', 'Theater', 'Travel', 'U.S.' and 'World'.

The dataset was split into training (70%) and testing (30%) sets to ensure a thorough validation process. This division enables the model to acquire knowledge from a significant section of the data and ensures that there is a sufficient amount of data for accurate model evaluation.

The 'Headline' column, which once consisted of a collection of dictionary objects, was modified to enhance simplicity and user-friendliness. We extracted the principal dictionary containing the main headline from each list and stored this data in a new column called 'main\_headline'. The initial 'Headline' column was subsequently eliminated from the dataset to optimize the feature set. The 'Keywords' column, which is also a list of dictionaries, was streamlined to decrease intricacy and concentrate on irrelevant textual material. We eliminated less informative keys, such as 'rank' and 'majors', from these dictionaries. The keyword data was streamlined and transformed into a single string for each article. This string was then saved in a newly constructed column called 'keyword\_sentences'. The initial 'Keywords' column was removed after the transformation process. In order to generate a thorough textual representation of each article, we merged the contents of the 'main\_headline', ‘keyword\_sentences’, 'snippet', 'lead\_paragraph', and 'abstract' columns into a unified column called 'combined\_text'. This offers a compact and enlightening feature, encompassing the fundamental aspects of each previous column. After the generation of 'combined\_text', the individual contributing columns were eliminated to avoid repetition and concentrate the dataset exclusively on the most influential attributes.

Standardization was applied to all text columns to enhance the quality of input data for text analysis and in order to maintain consistency. The process entailed transforming all text to lowercase, eliminating special characters, numeric values, and excessive whitespace in order to properly cleanse the data. In addition, we removed commonly used stop words from the text in order to prioritize more significant words for subsequent analysis. We employed the Porter-Stemmer method to decompose the text into its fundamental components (tokens) and condense each word to its base form (stemming). This approach facilitates the reduction of textual complexity of related word variations into a single base form, hence boosting the efficacy of text analysis. The ‘combined\_text’ column was subsequently subjected to a TF-IDF vectorizer, which converted the text into a weighted vector format that emphasizes the significance of terms based on their frequency across documents while reducing their value if they are overly prevalent. The TF-IDF matrix was subsequently transformed into a DataFrame format and merged with the original dataset to incorporate these additional numerical characteristics. After vectorization to enhance the feature set, we carried out a feature selection procedure in which non-numeric columns were eliminated, and numeric characteristics were filtered based on their document frequency.

We utilized Label Encoding to convert the text labels in the 'section\_name' column into a distinct integer value for each class. This procedure is crucial for the machine learning model to make predictions on categorical data. The newly formed 'section\_name\_encoded' column after label encoding led to the removal of the ‘section\_name’ column from the dataset to prevent duplication. All columns except 'section\_name\_encoded' were included as independent variables in both 'X\_Train' and 'X\_Test'. The variable 'section\_name\_encoded', which is the objective for classification, was present in both 'Y\_Train' and 'Y\_Test'. Upon collecting all essential features and converting them into a format appropriate for our models, we assessed the usefulness of the 'combined\_text' column. While this column was important during feature engineering, especially for generating TF-IDF vectors, it was considered unnecessary after vectorization. Therefore, to optimize our dataset and concentrate just on the most influential characteristics, we made the decision to exclude the 'combined\_text' column. By reducing the dimensionality of our feature space, this operation improves the efficiency of the model by eliminating any possible noise or redundancy.

## Pipeline

In order to achieve efficient preparation and modification of the features in our dataset, we have created a thorough pipeline using Python's scikit-learn module. The purpose of this pipeline is to optimize the process of preparing textual data for machine learning modeling. It specifically emphasizes text normalization, vectorization, and encoding.

The preprocessing and vectorization components are combined in a pipeline, guaranteeing that each step is performed in order and that the output of one phase smoothly transitions to the next.

* Pipeline Construction: A Pipeline object integrates the TextPreprocessor and TfidfVectorizer to provide a cohesive process flow. This pipeline is responsible for converting unprocessed text into a format that is appropriate for machine learning models.

*Framework for transforming features*

* To perform these changes exclusively to the 'combined\_text' column and efficiently manage other data within the dataset, we utilize a ColumnTransformer:
* Column-Specific Transformation: The pipeline is integrated into a ColumnTransformer, which exclusively applies the text processing pipeline to the 'combined\_text' column. This technique guarantees that the customized modifications designed for text data do not unintentionally impact other data types in the collection.
* Non-transformed columns are excluded by dropping all columns that are not specifically indicated for transformation. This ensures that the focus is maintained on the features that are considered most valuable for the classification task.

## Models

To enhance text classification techniques, we employed various cutting-edge machine learning and deep learning models to efficiently categorize articles from a large dataset. Every model was carefully selected based on its pertinence and capacity to process textual material. We have established a set of pipelines to accommodate conventional machine learning models, such as Random Forest, Logistic Regression, K-Nearest Neighbors (KNN), and XGBoost. The pipelines were meticulously constructed to guarantee that preprocessing, including tokenization and text normalization, was uniformly implemented across all models, thus standardizing the input data.

To properly evaluate our four models, we decided to run each model three times with different parameters and check the performance of each. We first started by running the base models with their default parameters. This provides us with a benchmark on how the model is expected to perform. We then decided to run our models using the optimal parameters that yielded superior results based on the literature and their performance in the context of article classification. To further improve the model performance, we decided to run a grid search for each model. We explored a predefined set of hyperparameters for each model to identify the combination that yields the best results. Each model underwent cross-validation to assess its performance and prevent overfitting. A 5-fold cross-validation was applied to each model.

### Random Forest Classifier

The Random Forest Classifier was chosen for its resistance to overfitting and its capacity to efficiently generalize to unknown data. At first, when the classifier was used with its default parameters, the model attained an accuracy of 71.88%. By incorporating suggestions from a literature review, the model was further improved, resulting in a minor increase in accuracy to 72.16%. This improvement was achieved by using a configuration of 3,000 estimators and setting the maximum depth to 100. Afterwards, the model was further improved by optimization using Grid Search, resulting in an increased accuracy of 72.77%. The parameters that were found to be the most effective were: **n\_estimators: 300, max\_depth: None (allowing unlimited tree growth) min\_samples\_split: 5, min\_samples\_leaf: 2, max\_features: 'sqrt', bootstrap: False.**

### Logistic Regression

The Logistic Regression model initially achieved a baseline accuracy of 66.01% when employing the default settings. To improve the performance of the model, modifications were made based on a thorough analysis of existing literature. These adjustments resulted in an enhanced accuracy of 71.49%. The adjustments were made based on previous research and established best practices in the field, indicating an initial fine-tuning of hyperparameters. Additional improvement was accomplished by conducting a thorough Grid Search to efficiently optimize many hyperparameters. The optimal parameters derived from the Grid Search were:

**C: 10, class\_weight: balanced, Penalty: L2, Solver: LBFGS**

By utilizing these adjusted settings, the model attained its peak accuracy of 74.73%. This enhancement highlights the efficacy of focused hyperparameter adjustment in improving the performance of Logistic Regression models, especially for intricate classification tasks.

### KNN

Using the classifier with its default parameters, yielded an accuracy of 62.77%. Taking into account the suggestions from a literature review, the model performance was slightly improved to reach an accuracy of 63.83%. This improvement was achieved by using 9 n\_neighbors. We tried to run a grid search to try and further optimize the parameters which yielded an increase in the accuracy up to 66%. The parameters that were found to be the most effective were: **clf\_algorithm: auto, clfmetric: euclidean, clfn\_neighbors: 10, clf\_weights: distance.**

Taking into account that KNN yielded the lowest accuracy among all the models, it is clear that the model is not appropriate for handling large-scale datasets or high-dimensional feature spaces due to its computational inefficiency and reliance on distance calculations. However, it can still be a feasible choice for smaller-scale article classification tasks or as a basic model for comparison with more sophisticated algorithms.

### XGBoost

The XGBoost Classifier was chosen for its robustness in handling classification tasks, particularly in the context of news article categorization. Initially, the model achieved an accuracy of 71.49% with default parameters. After incorporating insights from the literature, the model was further improved to an accuracy of 72.61% through these parameters: use\_label\_encoder=False, eval\_metric='mlogloss', and setting max\_depth to 5. In an attempt to further enhance the model, we fitted the XGBoost Model on the following Grid Search: This optimization resulted in a refined model with an accuracy of 72.55%, and the parameters that contributed to this improvement were the following: **'clf\_eval\_metric': 'mlogloss', 'clflearning\_rate': 0.2, 'clfmax\_depth': 4, 'clfnum\_class': 24, 'clf\_objective': 'multi:softmax'.**

### Neural Network

We integrated a Neural Network model using TensorFlow to enhance the accuracy of our predictions. This model employs a special arrangement of layers, including embedding, LSTM, and thick layers, which are designed to effectively handle the intricacies of textual data. We converted the section names in our dataset into numeric representations using LabelEncoder. This made the classification process easier and ensured compatibility with neural network outputs. The dataset was partitioned into separate sets for training and testing purposes. At first, we allocated 20% of the data for testing. This portion was then evenly divided to create the final test set. We used a consistent random state to guarantee reproducibility. A textual input was transformed using an Embedding Layer to represent it as a dense vector in a higher-dimensional space. This layer employed a predetermined embedding matrix that was designated as non-trainable to maintain the previously acquired word associations. To mitigate overfitting, the SpatialDropout1D technique was employed after the embedding layer. This technique involves discarding complete 1D feature maps. A (LSTM) layer consisting of 128 units is employed to effectively capture temporal dependencies in the data. A Dense Layer is added to the model with 64 units and the (ReLU) activation function. This introduces non-linearity into the model, which improves its ability to learn. A Dropout Layer is implemented with a dropout rate of 0.5 to mitigate overfitting. This is achieved by randomly deactivating input units and setting them to 0 during the training process. The ultimate output layer employed a Softmax Activation function, which is well-suited for multi-class classification, resulting in a probability distribution over the different classes. The model was created using the Adam optimizer and sparse categorical cross-entropy loss, which is suitable for multi-class classification jobs where each class is represented by a single label. The model underwent training for 10 epochs using a batch size of 32. The model's performance was monitored and overfitting was mitigated by using testing data. The model revealed an accuracy rate of roughly 71.86%. This statistic demonstrates a commendable level of performance, while there is potential for enhancement, possibly through refining the model's structure or training methodology.

### Utilizing BERT for Deep Learning

Incorporating deep learning, we added the BERT model, which is widely recognized for its exceptional performance in natural language comprehension. The dataset was divided into two parts, with 20% of the data put aside for testing purposes. This was done to ensure that the evaluation of the model's performance is fair and unbiased. The BERT tokenizer was utilized to tokenize and encode the text into formats that are appropriate for input into the model. The model configuration involved using the 'bert-base-uncased' model with modifications to accommodate the specific amount of distinct labels in our dataset. This configuration guarantees that the model is meticulously optimized for our precise classification requirements. The training process for BERT involved three epochs, during which the loss was closely monitored to modify the learning rates and prevent overfitting. ***Training the Initial Model using Validation Set:*** Initially, a BERT model was developed utilizing a division of training, validation, and test sets. The purpose of including a validation set was to mitigate the overfitting of the model by periodically assessing its performance on data that was not used for training. Although this strategy aimed to create a more broad model, it led to a reduced accuracy rate of 79%. The decrease in performance may be ascribed to the validation and test sets being more arduous or dissimilar to the training set, a common occurrence when the validation technique effectively mitigates overfitting. ***Model Optimization without Validation Set:***After evaluating the performance and attributes of the first model, we refined our technique and modified the model to train just on the training and test sets, omitting the validation step. Presumably, this was a reaction to the particular requirements of our dataset and the observed attributes during the initial training. The optimized model attained a superior accuracy of 88% on the test set. This suggests that the model may have been more finely adjusted to the particular characteristics of the training and test data, resulting in an enhancement in its ability to handle real-world test scenarios or a more accurate alignment with the distribution of the data. ***Conclusion on Best Result:*** The superior accuracy achieved in the second configuration indicates that, for our particular dataset and task, the model without the validation set outperformed the model in test situations. This result suggests that either the test set closely resembled the attributes of the training set, or that the model parameters were well-matched to the characteristics of the data. Nevertheless, it is crucial to take into account the balance between the risk of overfitting the training data and the goal of attaining high accuracy on the test set. When a model needs to perform well on fresh, unseen data, it can be beneficial to reintroduce and fine-tune the model using a validation set.

### Best Model and Comparative Analysis with Related Work

Among the various models tested, the BERT model, optimized without a validation set, emerged as the most effective, achieving an impressive accuracy of **88%** on the test set. This model, leveraging the deep learning capabilities of the 'bert-base-uncased' configuration, was specifically adapted to our dataset's unique characteristics. By focusing exclusively on training and test data, the model demonstrated superior performance, suggesting an optimal fit to the dataset's distribution and nuances. This configuration's success underscores BERT's robustness in text classification tasks, highlighting its ability to extract and learn from complex textual relationships effectively, making it the standout choice for enhancing article categorization in The New York Times dataset. Studies on news article classification often yield accuracies between 70% and 85% when employing traditional machine learning techniques. Our solution not only surpasses the previous limits of these outcomes but also offers valuable insights into successful ways for utilizing BERT in a real-world dataset containing diverse and intricate text data.

# Results & Discussions

|  |  |  |  |
| --- | --- | --- | --- |
| **Basic Models** | **Base** | **Literature Review** | **Grid Search** |
| Random Forest Classifier | 0.718837 | 0.721632 | 0.727781 |
| Logistic Regression | 0.660145 | 0.714925 | 0.747345 |
| KNN | 0.627725 | 0.638345 | 0.665176 |
| XG Boost | 0.714925 | 0.726104 | 0.725545 |
| Neural Network | 0.718593 | | |

|  |  |  |  |
| --- | --- | --- | --- |
| **BERT** | **Training** | **Testing** | **Validation** |
| 3 splits | 0.919031 | 0.790369 | 0.798755 |
| 2 splits | Test Accuracy: 0.886336 | | |

## Answering the research question

In our research, we meticulously outlined key preprocessing steps including null value handling, text standardization, and TF-IDF vectorization to properly prepare the dataset for machine learning analysis. To improve prediction accuracy, we tested and refined multiple machine learning models such as Random Forest, Logistic Regression, KNN, XGBoost, and BERT through rigorous grid search methods. Further enhancing our capability to handle complex text classification tasks, we integrated BERT, a state-of-the-art deep learning model. We conducted a comprehensive analysis of our results, showcasing the accuracies of our models and the significant impact of our optimizations in enhancing content relevance and accessibility. Additionally, by utilizing a broad dataset from The New York Times and employing models effective in diverse contexts, we have successfully optimized our machine learning and NLP techniques to accurately classify articles into their respective sections. This capability allows us to address the dynamic nature of news topics and cater to the varied interests of readers, ensuring that our methods meet the complex demands of modern news dissemination and enhance the relevance and accessibility of content.

## Error Analysis

Please refer to the appendix section for illustrations of the errors.

***The confusion matrix*** showed that the model performed well in differentiating categories with distinct content like "Sports", but struggled with categories having overlapping themes such as "U.S." and "World." Misclassifications between "U.S." and "Business Day" suggested confusion in articles involving political and business contexts. ***The classification report*** reveals a similar pattern: strong performance in clearly defined categories like "Sports'' but issues in categories such as "Blogs'' and "Technology," where precision and recall varied significantly. Also, categories with less training data or broader thematic coverage exhibited lower performance metrics. ***Error inspection*** highlighted specific instances where the model confused categories due to overlapping content themes. For example, a text involving Bill Clinton's memoir was misclassified from "Business Day" to "Arts," likely due to the focus on a cultural figure rather than the business aspect of the book deal. Analysis of the top***misclassified classes*** indicated that "Blogs," "Arts," and "U.S." had the highest number of errors. These categories likely suffer from thematic overlaps with other categories, causing higher confusion in the classification process. This might also be due to a higher number of texts under these subjects in the dataset. *The text lengt*h analysis showed that shorter texts were more prone to misclassification, suggesting that the model may require more context to make accurate predictions, which longer texts tend to provide. ***Word and Phrase analysis*** highlighted the presence of specific bigrams (two-word combinations) that frequently appeared in misclassified articles. For example, bigrams like "New York" and "Los Angeles" were prevalent in the misclassified texts. These geographical terms might be confusing because they are common across various news categories that deal with local events, cultural stories, or business news, making them non-discriminative features. Thus, the model appeared to struggle with distinguishing contextual nuances in texts that contained common phrases across multiple categories. As a last step, we performed ***sentiment analysis*** on misclassified texts to check if there is a correlation between the sentiment of the text and its predicted class. The analysis revealed that the subtle emotional undertones of the text did not impact its classification by the model.

## Obstacles and Solutions in Model Development

Throughout our project, we faced numerous obstacles, mostly with model optimization and overfitting. At first, the dataset's diverse and intricate textual data posed a challenge in attaining high accuracy across all models. To tackle these problems, we conducted thorough experiments with hyperparameter tuning and implemented techniques such as Grid Search to identify the most effective configurations. This led to an enhancement in the performance of the model. Overfitting was a notable obstacle, especially when dealing with deep learning models. To address this issue, we made modifications to the model structures, integrated dropout layers, and improved our training methodology by implementing cross-validation. By following these methods, we were able to ensure that our models exhibited strong generalization capabilities when faced with new data, thereby improving their resilience and dependability in practical scenarios.

# Conclusion

## Summary of Solution, Findings, and Recommendations:

The project effectively utilized a range of sophisticated machine learning and deep learning models, specifically the BERT model, to improve the classification of text in the extensive dataset of The New York Times articles. The results of our study suggest that although classic models such as Random Forest and Logistic Regression performed well, the BERT model, especially when optimized without a validation set, achieved much better results compared to other models, with an astounding accuracy rate of 88%. This exemplifies the efficacy of customizing cutting-edge models to particular datasets, highlighting the capabilities of BERT in comprehending and analyzing intricate textual information.

Nevertheless, this strategy has disadvantages, particularly the potential for overfitting when the validation set is excluded. While this approach achieved a high level of accuracy, its ability to effectively apply to completely unfamiliar or fresh data may be limited. From a technical perspective, it is recommended that future work should prioritize the inclusion of a validation phase to achieve a balance between model fit and generalizability. Furthermore, investigating other developing models in Natural Language Processing (NLP) such as Transformer-based architectures could potentially improve both the precision and effectiveness of the system.

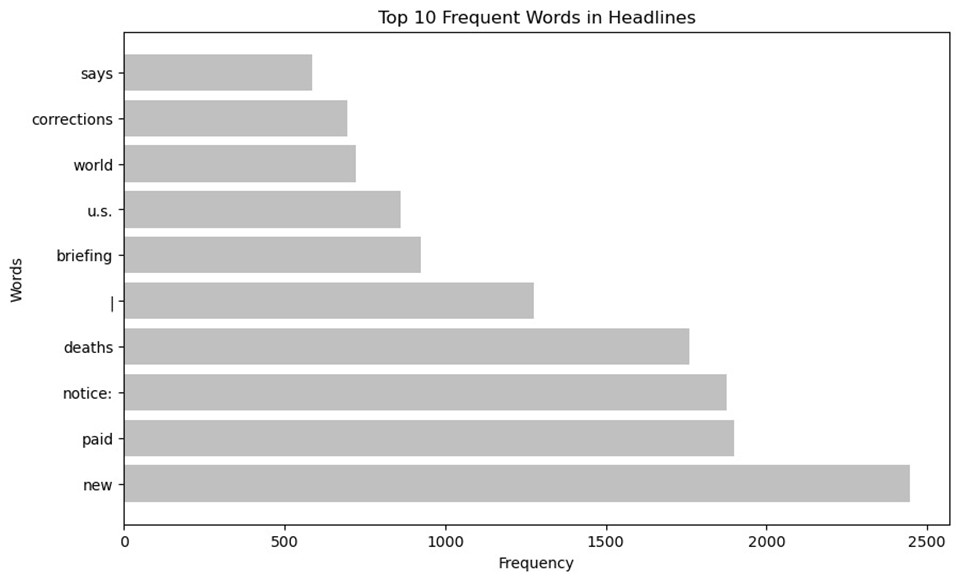
From a business standpoint, the integration of these sophisticated text categorization algorithms has the potential to greatly enhance content targeting and personalization strategies for The New York Times. This might result in a notable boost in reader engagement and subscription rates. Moreover, similar approaches can be customized for different digital media platforms aiming to improve their content distribution and recommendation systems. Further research could investigate the incorporation of real-time categorization systems to aid editorial teams in dynamically categorizing information as it is generated, promoting a more flexible newsroom setting.

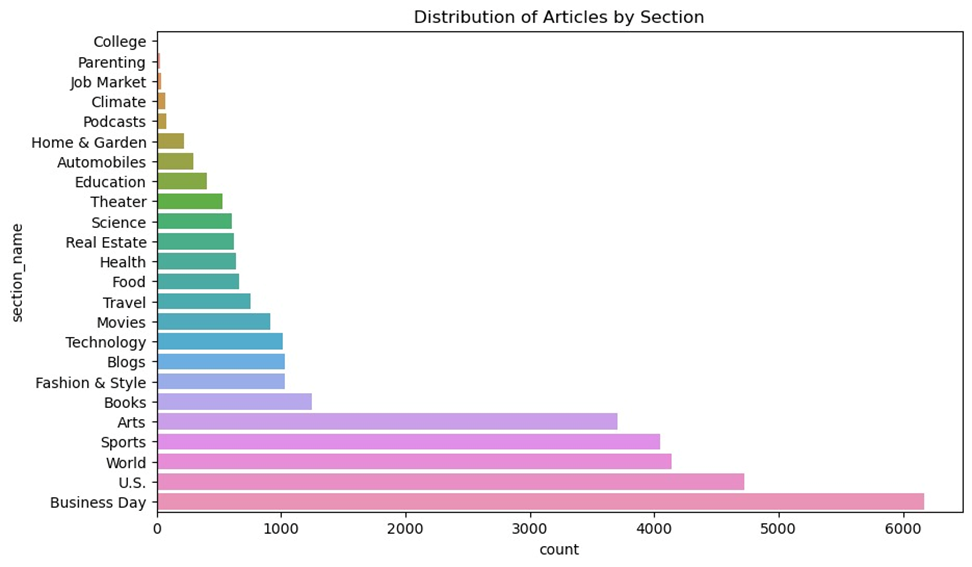
## Data Handling, Business Recommendations, Strategic Insights, and Future Work

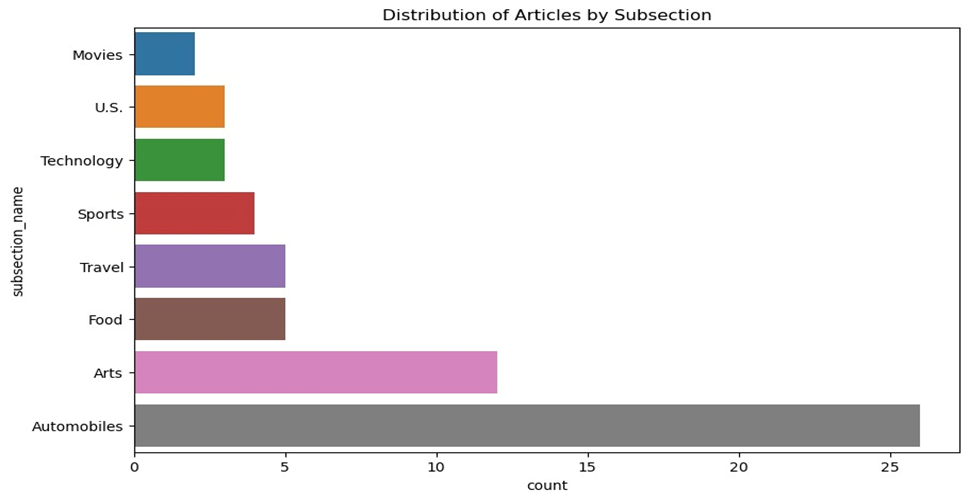
Our experiment has shown that using advanced machine learning algorithms for data handling in The New York Times article dataset is effective. This ensures that articles are appropriately categorized, leading to improved efficiency in content delivery. From a business standpoint, we suggest that The New York Times utilize our categorization algorithms to improve user experience by customizing information streams according to expected interests. This would result in higher user engagement and potentially more subscriptions. From a strategic standpoint, these insights have the potential to guide the editorial team's decision-making process by identifying popular subjects and areas where content is lacking, which could ultimately attract a larger readership. In our future study, we suggest investigating the use of real-time analytics to identify and dynamically recommend material as it is being created. This might be used with user interaction data to iteratively enhance model predictions by leveraging real-time reader engagement. In addition, the utilization of unsupervised learning methods could reveal hidden subjects and emerging patterns without pre-established classifications, offering The New York Times novel approaches to excel in content generation and dissemination in the digital age.

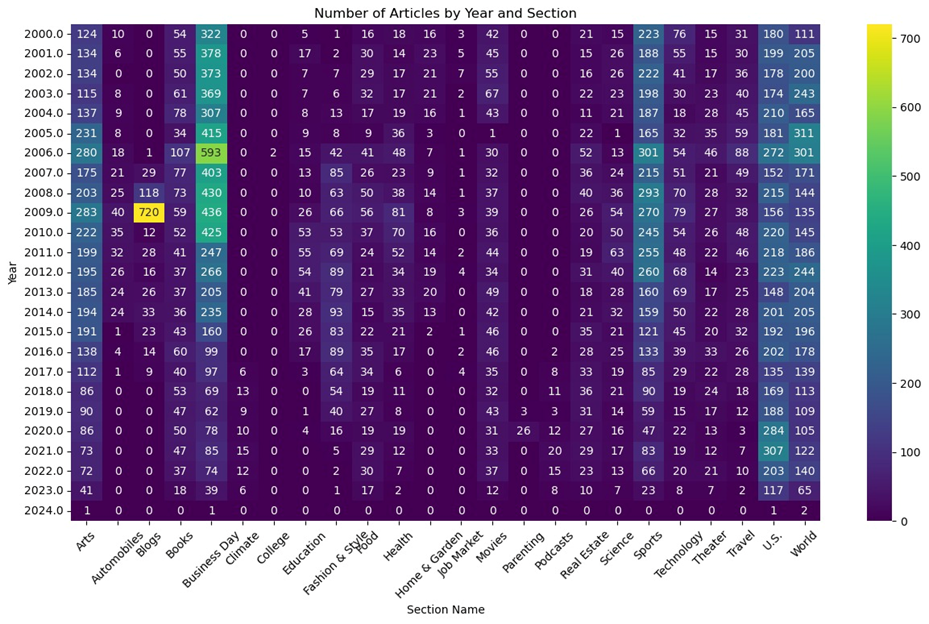
# Appendices

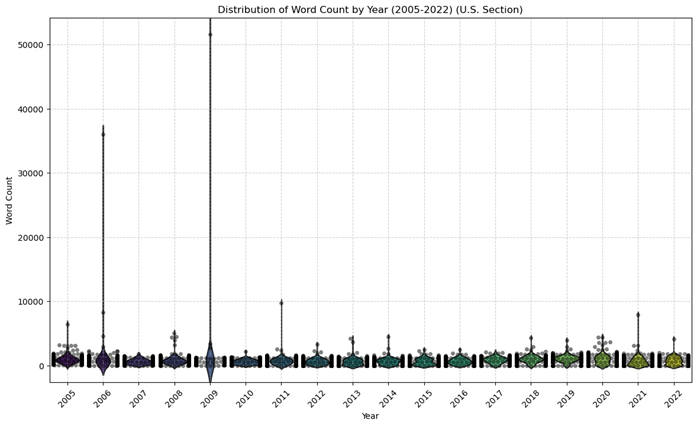
***EDA***

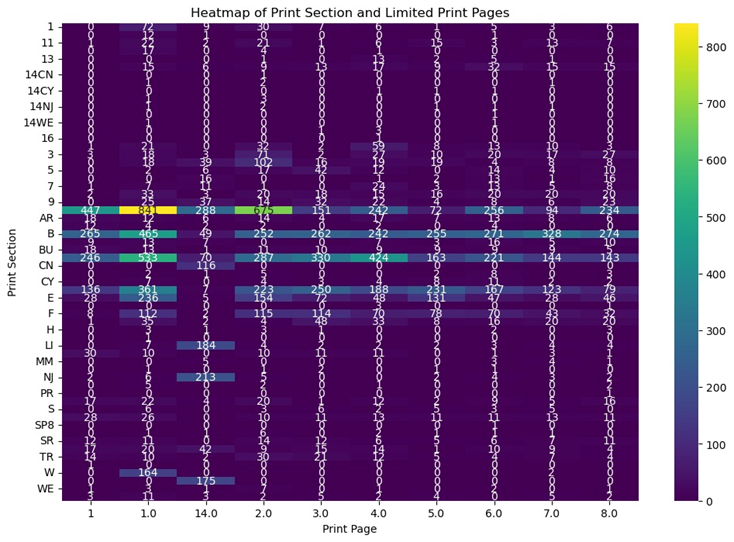




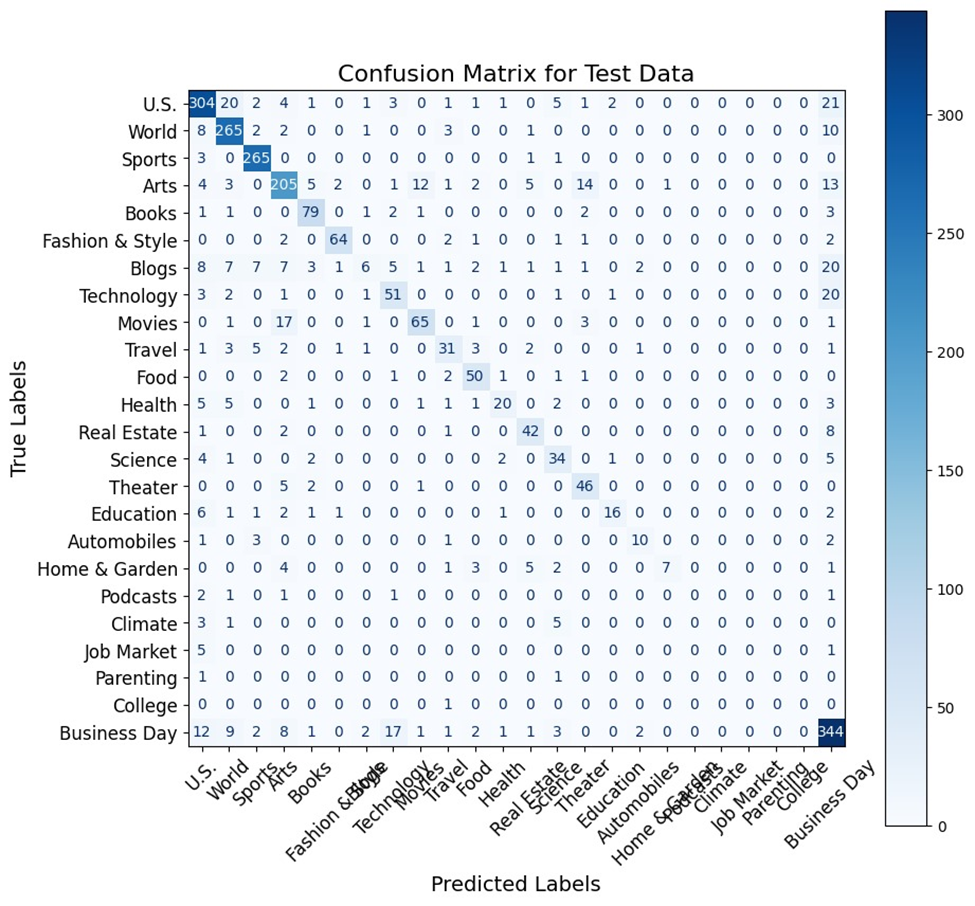


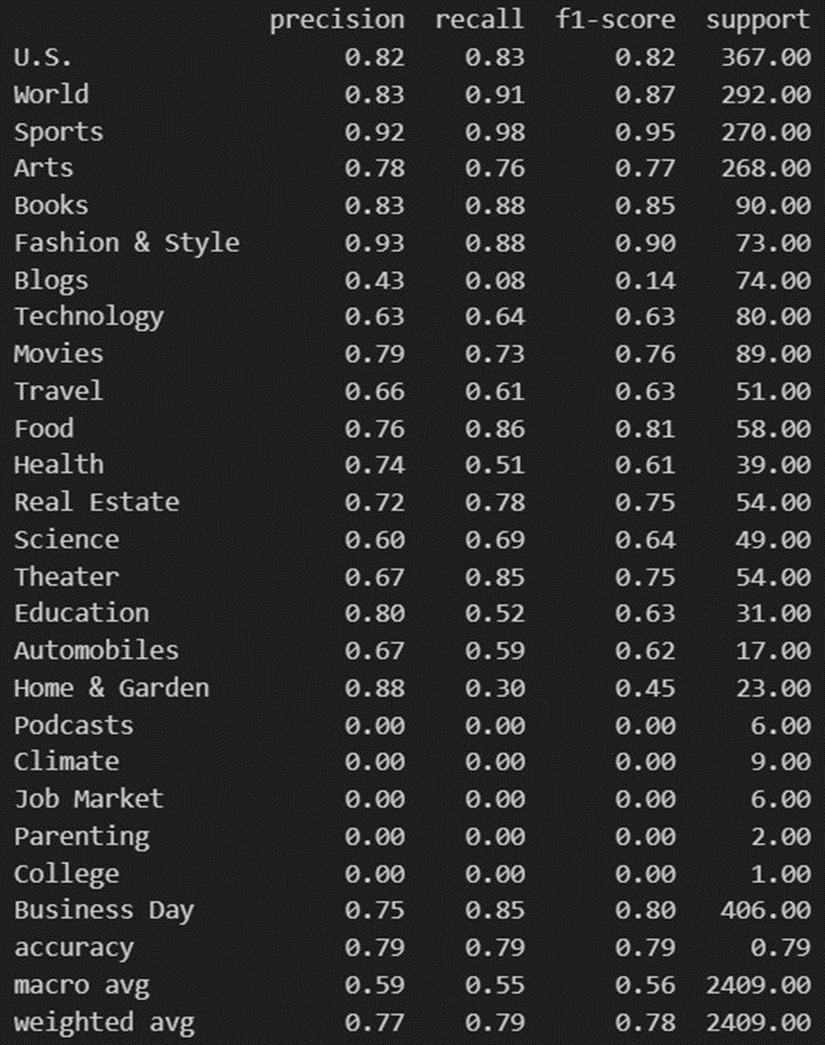












Sample for error analysis (1):

Text: ‘clinton is said to be close to book deal bill clinton said to be close to deal to sell rights to publish memoir for advance that will probably exceed his wife's near-record payment of $8 million for her memoir; former president may decide to accept offer and negotiate contract without au... bill clinton said to be close to deal to sell rights to publish memoir for advance that will probably exceed his wife's near-record payment of $8 million for her memoir; former president may decide to accept offer and negotiate contract without auction; that would allow him to avoid discussing his plans with publishers and soliciting bids, as his wife, sen hillary rodham clinton, did; publishing executives say any offer would be contingent on nature of book--that it be personal memoir rather than treatise on public policy; it is not clear how much clinton has disclosed about his plans for memoir, including whether he would discuss monica s lewinsky or whitewater investigation (m) bill clinton is close to a deal to sell the rights to publish a memoir for an advance that will probably exceed his wife's near-record payment, people close to mr. clinton said last week.

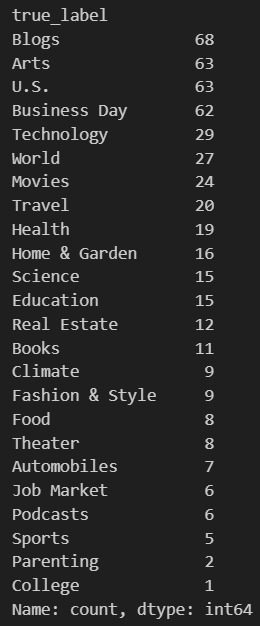
Persons: Clinton, Bill, persons: Clinton, Hillary Rodham, persons: Lewinsky, Monica S,

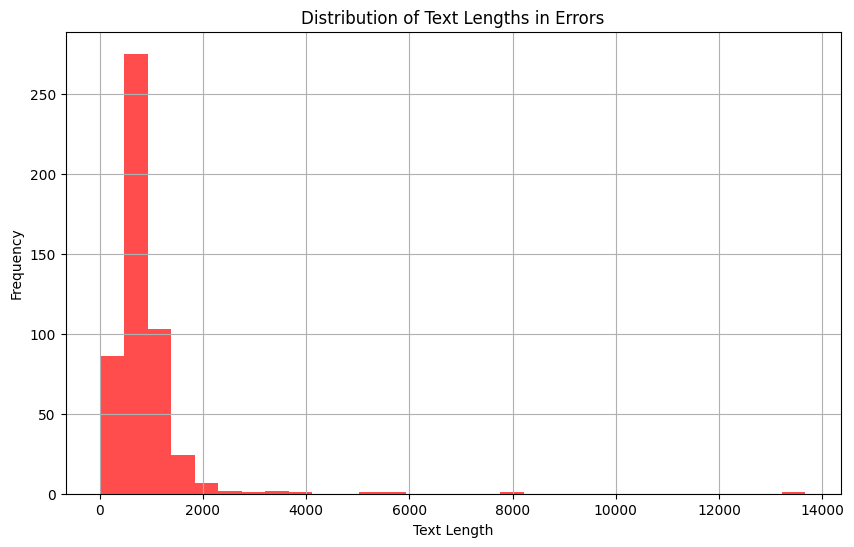
Subject: United States politics and government, subject: books and literature, subject: biographical information, subject: ethics

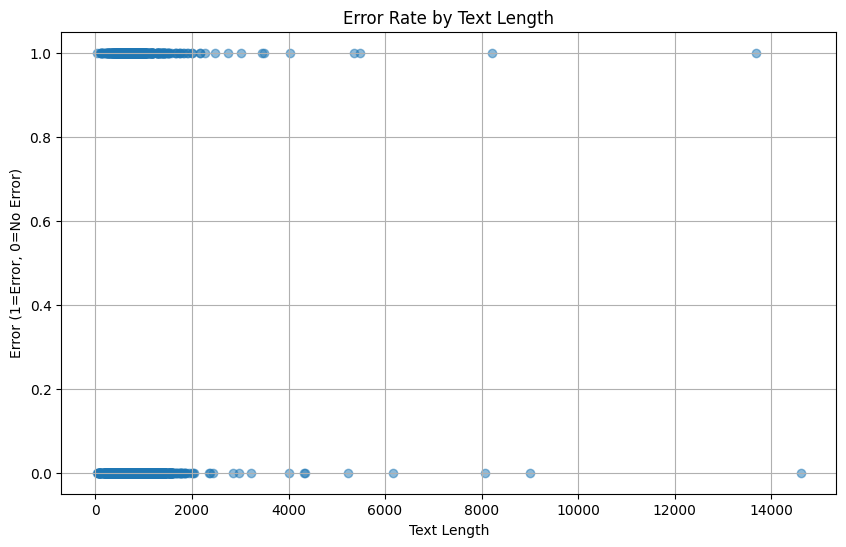
true\_label: Business Day

predicted\_label: U.S.

Name: 16505, dtype: object







Sample for error analysis (2):

Text: Public Enemy to use a digital distributor Tunecore, a digital music distributor, is expected to announce that Public Enemy, one of the seminal hip-hop groups, will use its service for its new album. TuneCore, a digital music distributor, is expected to announce that Public Enemy, one of the seminal hip-hop groups, will use its service for its new album. Jeff Price, the founder and chief executive of Tunecore, a digital music distributor, has a simple pitch for musicians: “For $30, the cost of a pizza and a six-pack, you can get your album on iTunes, the third-largest music store in the country.”

Subject: advertising and marketing, persons: Chuck D (1960- ), persons: Winehouse, Amy, subject: music, persons: Marley, ziggy, organizations: public enemy

True Label: Business Day

Predicted Label: Arts

Sentiment: -0.017373737373737375

----------------------------------------------------------------------------------------------------

Sample for error analysis (4):

Text: U.S. pledges to pay family of those killed in botched Kabul drone strike the pentagon offered unspecified amounts to relatives of the 10 civilians who died in Aug. 29 attack and agreed to help relocate those who want to move to the U.S. The Pentagon offered unspecified amounts to relatives of the 10 civilians who died in aug. 29 attack and agreed to help relocate those who want to move to the u.s. Washington — the pentagon offered unspecified condolence payments this week to the family of the 10 civilians, including seven children, who the military has acknowledged was mistakenly killed on aug. 29 in the last u.s. drone strike before American troops withdrew from Afghanistan.

Subject: United States Defense and military forces, subject: afghanistan war (2001- ), subject: drones (pilotless planes), subject: civilian casualties, persons: ahmadi, zemari (d 2021), glocations: kabul (afghanistan)

True Label: U.S.

Predicted Label: World

Sentiment: -0.1222222222222222

----------------------------------------------------------------------------------------------------

Sample for error analysis (5):

Text: The gleaming landscape of debt around 30 new condo projects in Fort Greene and Clinton hill is in financial trouble, according to assemblyman Hakeem Jeffries, who wants to turn vacant luxury dwellings into affordable housing. around 30 new condo projects in Fort Greene and Clinton Hill are in financial trouble, according to assemblyman Hakeem Jeffries, who wants to turn vacant luxury dwellings into affordable housing. update &#124; 5:51 p.m., aug. 28. none

True Label: Blogs

Predicted Label: U.S.

Sentiment: 0.03409090909090909

----------------------------------------------------------------------------------------------------

# Reference

Ahmed, J., & Ahmed, M. (2021). Online News Classification Using Machine Learning Techniques. IIUM Engineering Journal, 22(2). <https://doi.org/10.31436/iiumej.v22i2.1662>

Dien, T. T., Loc, B. H., & Thai-Nghe, N. (2019). Article Classification using Natural Language Processing and Machine Learning. 2019 International Conference on Advanced Computing and Applications (ACOMP).

Tan, H. (2021). Machine Learning Algorithm for Classification. Journal of Physics: Conference Series, 1994, 012016. <https://doi.org/10.1088/1742-6596/1994/1/012016>

Daud, S., Ullah, M., Rehman, A., Saba, T., Damaševičius, R., & Sattar, A. (2023). Topic Classification of Online News Articles Using Optimized Machine Learning Models. Computers, 12(16). <https://doi.org/10.3390/computers12010016>

Palanivinayagam, A.; El-Bayeh, C.Z.; Damaševičius, R. Twenty Years of Machine-Learning-Based Text Classification: A Systematic Review. Algorithms 2023, 16, 236. <https://doi.org/10.3390/a16050236>

Aum, S., & Choe, S. (2021). srBERT: automatic article classification model for systematic review using BERT. Systematic Reviews, 10, 285. <https://doi.org/10.1186/s13643-021-01763-w>.

Disayiram, N., & Rupasingha, R. A. H. M. (2022). A Comparative Study of Classifying English News Articles Using Machine Learning Algorithms. 2022 Trends in Electrical Electronics Computer Engineering Conference (TEECCON). DOI: 10.1109/TEECCON54414.2022.9854832