horizontal line

Effective Classification of Cancer Documents in Biomedical Research

**Vritihik Sibbadi**

**6th May 2024**

# Introduction

In the rapidly evolving field of biomedical research, text classification plays a crucial role in organizing vast amounts of unstructured text data, enabling researchers and healthcare professionals to derive meaningful insights rapidly. By automating the categorization of text into predefined classes, this technology facilitates quicker decision-making and enhances the accessibility of medical literature and research findings.

The business problem addressed in this project focuses on the classification of extensive biomedical documents into three specific types of cancer: Thyroid Cancer, Colon Cancer, and Lung Cancer. Effective classification supports our organization's goals by improving information retrieval processes, enhancing diagnostic research, and ultimately contributing to personalized medicine strategies.

# Data Source and Justification

This project utilizes a real-world dataset sourced from Kaggle, specifically designed for the classification of biomedical texts related to cancer research. The dataset can be accessed here:[kaggle.com/datasets/falgunipatel19/biomedical-text-publication-classification/data](https://www.kaggle.com/datasets/falgunipatel19/biomedical-text-publication-classification/data)

This dataset is deemed real-world data due to its basis in actual academic and clinical documents that discuss various aspects of cancer-related studies, making it highly relevant and representative of ongoing research in the field. The dataset's focus on long research papers, which are more detailed and comprehensive than shorter articles, provides a robust challenge typical of real-world data applications.

Statistical Overview:

* Total Publications: 7569
* Category Distribution:
  + Thyroid Cancer: 2810 samples
  + Colon Cancer: 2579 samples
  + Lung Cancer: 2180 samples

This data, with its detailed content and significant volume across three major cancer categories, offers an excellent basis for developing and refining machine learning models intended for real-world application in biomedical document classification.

# Methodology

**Data Preprocessing:**

The dataset underwent several preprocessing steps to ensure optimal model performance:

1. Cleaning: Columns were renamed for clarity, and unnecessary index columns were removed. Text data was standardized by converting to lowercase, removing punctuation, numbers, and extra spaces to reduce noise.
2. Duplicate Removal: Duplicate entries were identified and removed to ensure the integrity of the dataset, leaving us with a distinct set of documents.
3. Stop Words Removal: Common English stop words were removed to focus on more meaningful words in the texts.
4. TF-IDF Vectorization: The cleaned text was transformed using TF-IDF vectorization, setting a maximum of 2000 features. This technique emphasizes important words which are frequent in a document but not across documents, enhancing the model's ability to distinguish between cancer types.

**Exploratory Data Analysis:**

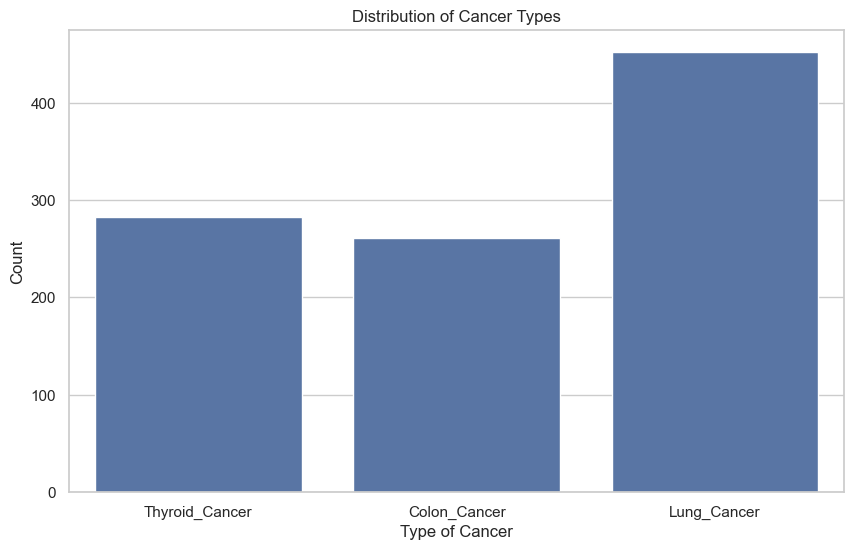


Fig 1 - cancer type distribution after removing duplicates

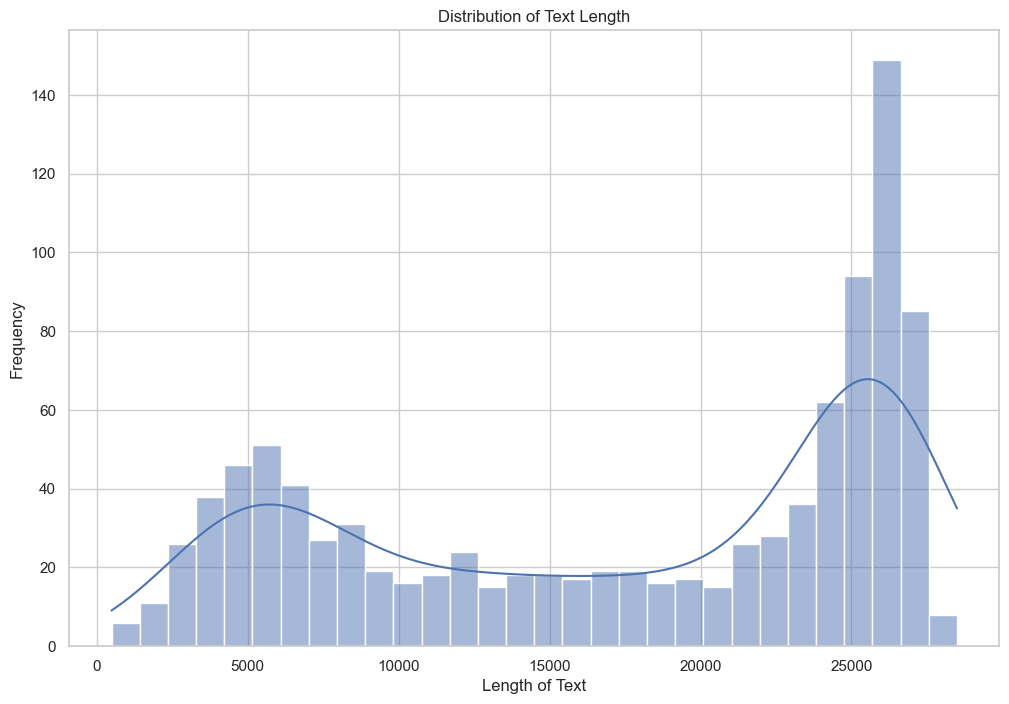


Fig 2 - distribution of text length

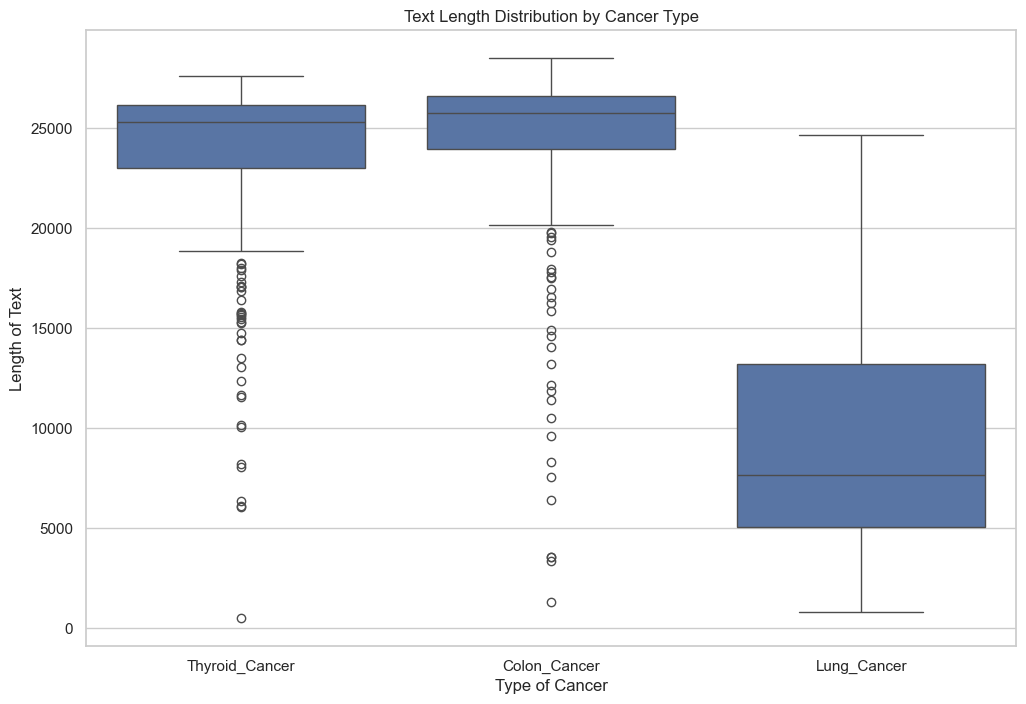


Fig 3 - distribution of text length by cancer types

**Model Selection:**

Three machine learning models were employed to classify cancer documents: Logistic Regression, Random Forest, and Support Vector Machine (SVM). These models were chosen due to their effectiveness in handling classification tasks with high-dimensional data.

**Model Training and Hyperparameter Tuning:**

* The dataset was split into training (80%) and testing (20%) sets to evaluate model performance.
* Each model was trained on the training set, and hyperparameters were tuned using GridSearchCV, particularly for the Random Forest model. This process involved adjusting parameters like the number of estimators and the depth of the trees to find the optimal settings for best performance.

The models' performance was compared based on accuracy, precision, recall, and F1-score. Random Forest showed improved performance through hyperparameter tuning, achieving an accuracy of 80%, making it the selected model for deployment.

This methodology ensures that the analysis is robust, providing reliable and actionable insights for classifying cancer types in biomedical texts.

# Results

**Model Performance Metrics**

The analysis involved three primary classification models: Logistic Regression, Support Vector Machine (SVM), and Random Forest. Below is a comparative summary of their performance metrics evaluated on the test set:

| **Metric** | **Logistic Regression** | **Random Forest** | **SVM** | **Improved Random Forest** |
| --- | --- | --- | --- | --- |
| **Accuracy** | 74% | 79.50% | 73.50% | 80% |
| **Precision (avg)** | 73% | 79% | 73% | 80% |
| **Recall (avg)** | 69% | 75% | 68% | 76% |
| **F1-Score (avg)** | 69% | 75% | 68% | 76% |

Best Model Identification

* The Random Forest model initially showed a good balance between precision and recall, achieving an accuracy of 79.5%.
* After hyperparameter tuning via GridSearchCV, the Improved Random Forest model further enhanced its performance, achieving the highest accuracy of 80%, with notable improvements in precision and recall across all classes.

Model Specifics and Confusion Matrix Insights:

* Logistic Regression struggled with distinguishing between 'Thyroid Cancer' and other types, resulting in lower precision.
* SVM displayed excellent results for 'Lung Cancer' but underperformed for 'Colon Cancer' and 'Thyroid Cancer' in terms of precision and recall.
* The Improved Random Forest model, with tuned parameters, demonstrated consistent improvements across all metrics, particularly in correctly predicting 'Lung Cancer' with nearly perfect precision and recall.

This comparative analysis indicates that the Improved Random Forest model, after parameter tuning, stands out as the most effective model for classifying the biomedical texts into the respective cancer categories. This model provides a robust solution for aiding in the diagnosis and categorization of cancer types based on textual data from biomedical publications.

# Conclusion and Recommendations

**Key Findings**

The project successfully applied machine learning techniques to classify biomedical texts into three cancer categories: Thyroid Cancer, Colon Cancer, and Lung Cancer. The Improved Random Forest model emerged as the most effective, achieving an accuracy of 80% with enhanced precision and recall. These results underscore the potential of machine learning in enhancing diagnostic accuracy and categorization in biomedical research.

**Implications**

The use of real-world, detailed biomedical text data for training enabled the models to perform with high relevance, suggesting that further enrichment of data could continue to improve accuracy and applicability in real-world scenarios.

**Recommendations**

1. Further Data Enrichment: Integrate more diverse datasets, including papers from additional cancer types and sub-types, to broaden the model's applicability and robustness.
2. Model Optimization: Continue to explore advanced machine learning techniques and hyperparameter tuning to further enhance model performance.
3. Implementation: Pilot the use of the Improved Random Forest model in a real-world clinical setting to evaluate its effectiveness in aiding decision-making processes.
4. Ongoing Research: Extend the research to include multi-label classification scenarios where documents may address multiple cancer types, reflecting more complex real-world conditions.

These steps are recommended to leverage the findings of this project to improve and expand the scope of machine learning applications in biomedical research and healthcare.

# References:

<https://www.kaggle.com/code/maushamjha/cancer-detection-via-nlp>

<https://www.kaggle.com/datasets/falgunipatel19/biomedical-text-publication-classification/data>