Northeastern University

Project 3: NLP

INFO7390.14011.202410

Fall 2022

Presented by Nupoor Korde - 002792237
Sarthak Jasani - 002728489

Outline

- **Dataset:** alldata_1_for_kaggle.csv
- **Goal:** The goal of the Natural Language Processing group project is to perform text classification on a dataset of long research papers related to cancer, with a focus on classifying them into one of three categories: 'ThyroidCancer,' 'ColonCancer,' and 'Lung_Cancer.' This project aims to implement and compare various machine learning and deep learning models, evaluate their performance using relevant metrics, and provide a comprehensive analysis of the models' strengths, weaknesses, and potential improvements.
- **Result:** The project's result involves the successful classification of cancer research papers into three categories using various models, accompanied by a thorough analysis of model performance and potential enhancements.

Tools used

- Jupyter

Dataset

- The project involves working with a dataset consisting of extensive research papers, each with more than 6 pages. The dataset encompasses cancer-related documents to classify them into three distinct categories: 'ThyroidCancer,' 'ColonCancer,' and 'Lung_Cancer.' With a total of 7569 publications, the dataset contains varying numbers of samples in each category: 2579 for colon cancer, 2180 for lung cancer, and 2810 for thyroid cancer.

- The project's approach includes training the model twice:

- 1) Initially, the model is trained without dropping duplicate entries.
- 2) Then, the model is trained after removing duplicate entries.

- The project aims to analyze and compare the results obtained from these two training approaches, providing insights into the impact of duplicate data on model performance and classification accuracy.

Methodology

- **Problem Domain Selection**:

Cancer Research Classification: The problem domain centers around categorizing extensive research papers that focus on various aspects of cancer. These papers cover in-depth studies, findings, and insights related to cancer, which is a critical area of medical research.

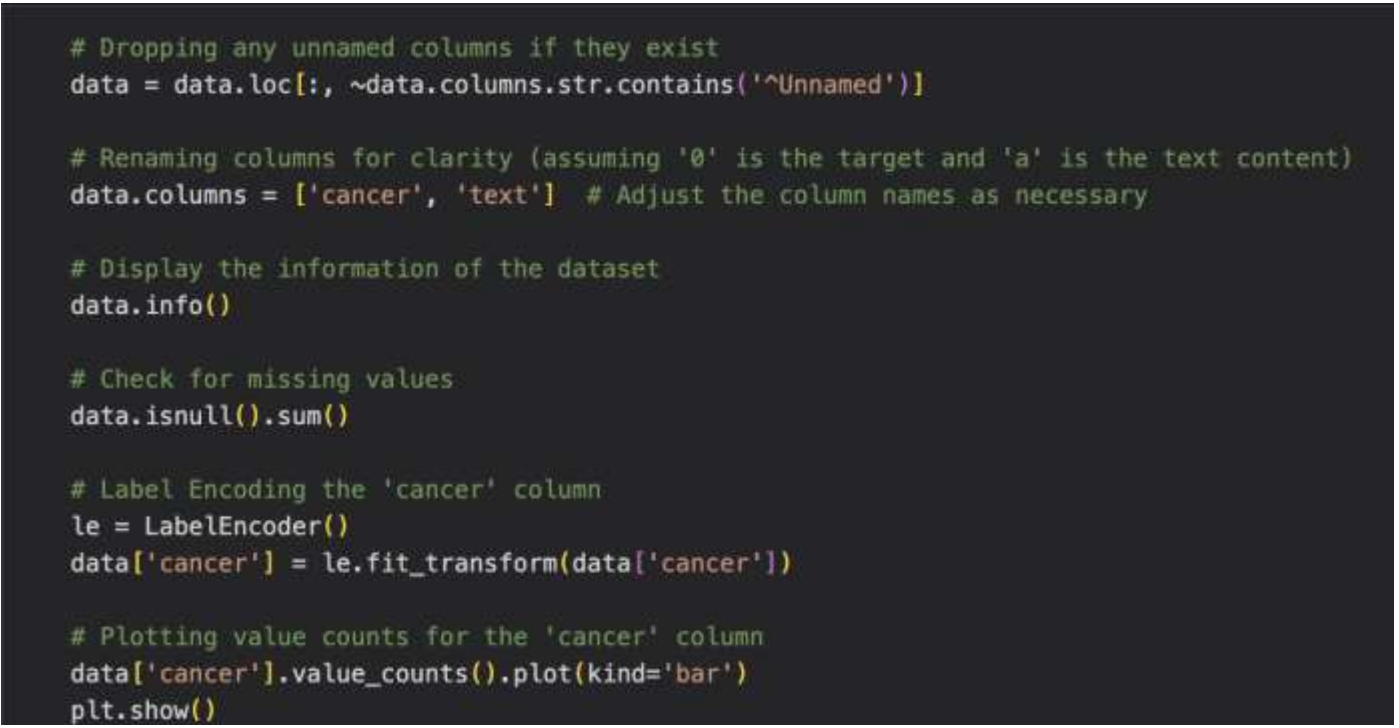
The classification task involves sorting these research papers into three distinct categories, each corresponding to a specific type of cancer:

- ThyroidCancer: Papers that primarily address thyroid cancer, its causes, treatments, and related research.
- ColonCancer: Research papers focusing on colon cancer, including studies on its prevention, diagnosis, and therapies.
- Lung_Cancer: Papers dealing with lung cancer, its various subtypes, and research on improving early detection and treatments.

- **Data Collection and Preprocessing:**

Data Collection:

- The dataset consists of a substantial collection of research papers, totaling 7569 publications.
- Each of these research papers focuses on various aspects of cancer, making it a valuable resource for understanding the disease's multiple dimensions.
- The dataset selection criteria include research papers with a page size greater than 6 pages, ensuring that they contain comprehensive and in-depth information.
- Drop the column unnamed as it is not usef
- Rename the column into readable form



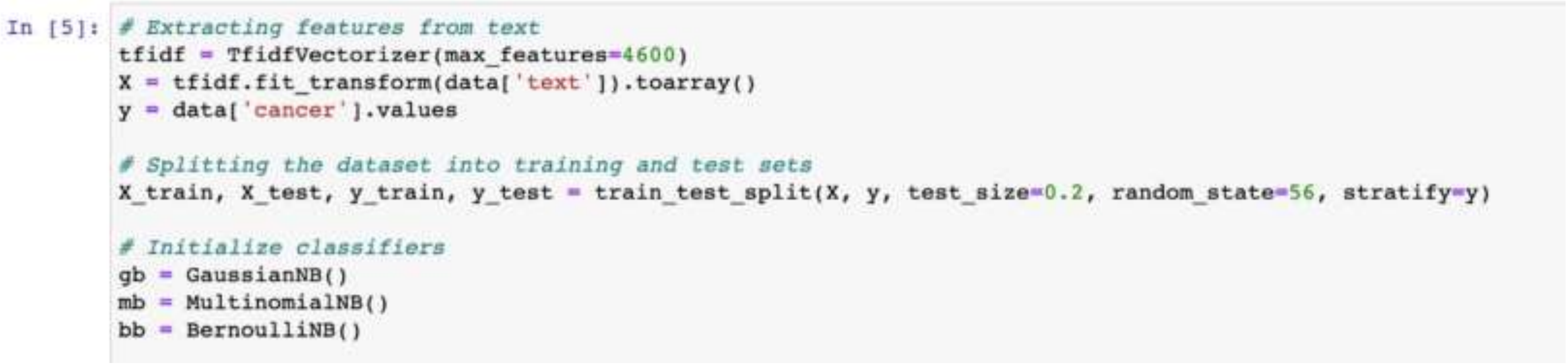
```
# Dropping any unnamed columns if they exist
data = data.loc[:, ~data.columns.str.contains('^Unnamed')]

# Renaming columns for clarity (assuming '0' is the target and 'a' is the text content)
data.columns = ['cancer', 'text']  # Adjust the column names as necessary

# Display the information of the dataset
data.info()

# Check for missing values
data.isnull().sum()

# Label Encoding the 'cancer' column
le = LabelEncoder()
data['cancer'] = le.fit_transform(data['cancer'])

# Plotting value counts for the 'cancer' column
data['cancer'].value_counts().plot(kind='bar')
plt.show()
```

- **Data Preprocessing:**

Vectorization:

- Vectorization transforms each tokenized and normalized word or subword unit into a numerical representation.
- Common techniques include Count Vectorization, TF-IDF Vectorization, or more advanced methods like Word Embeddings (e.g., Word2Vec, GloVe).
- Numerical vectors enable the machine learning models to analyze and classify the research papers based on the content and associations between words.
- Apply text vectorization:

Textual data cannot be used for mathematical operation so text vectorization convert text into vectors and then algorithms can be applied on it



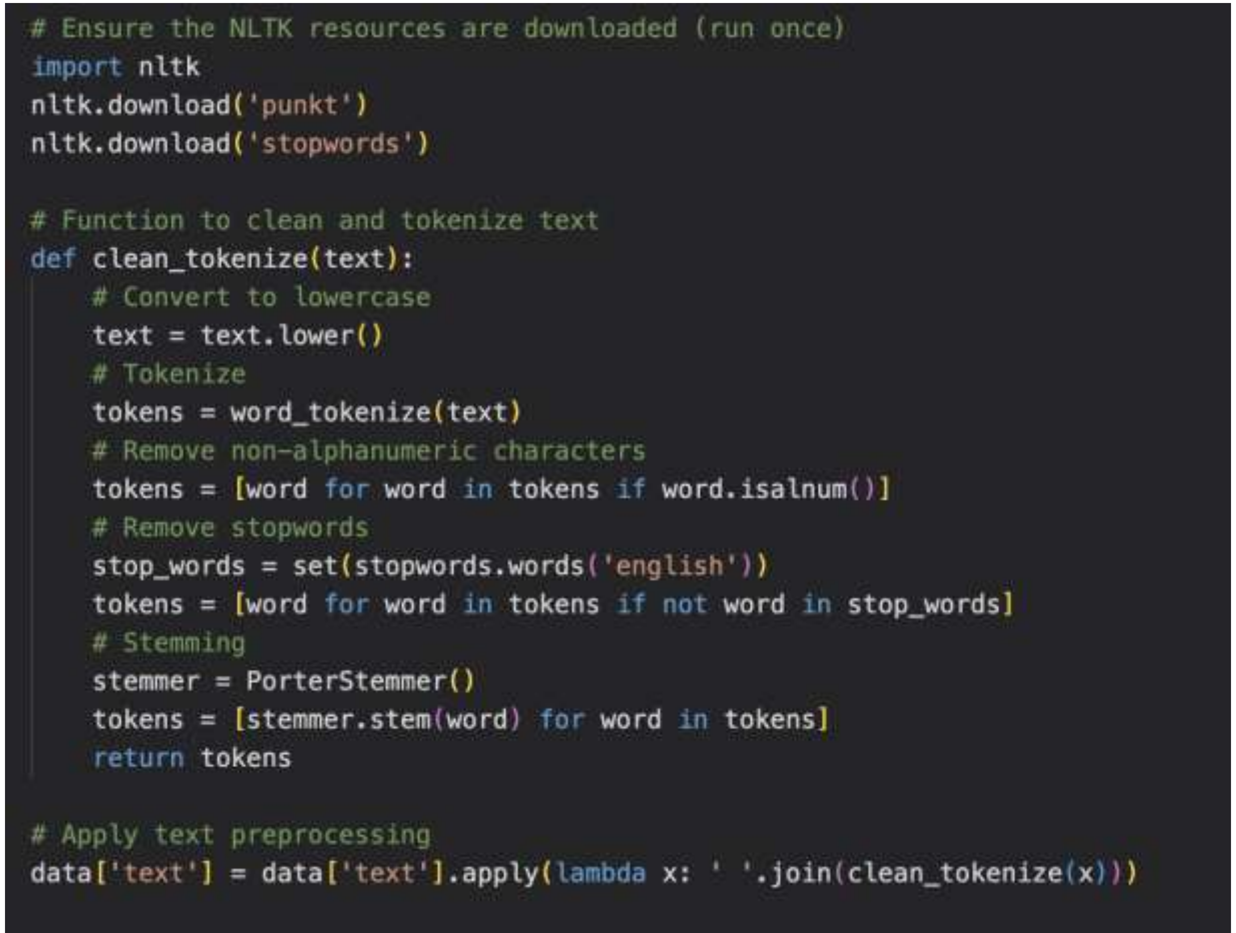
```
In [5]: # Extracting features from text
        tfidf = TfidfVectorizer(max_features=4600)
        X = tfidf.fit_transform(data['text']).toarray()
        y = data['cancer'].values

        # Splitting the dataset into training and test sets
        X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=56, stratify=y)

        # Initialize classifiers
        gb = GaussianNB()
        mb = MultinomialNB()
        bb = BernoulliNB()
```

• Data Preprocessing:

Text preprocessing
Steps :
1) Converting into lower case
2) Tokenizing : spliting sentences into words
3) Removing special characters
4) Removing stopwords and punctuations
5) Stemming: converting into root words
6) Join to make sentences

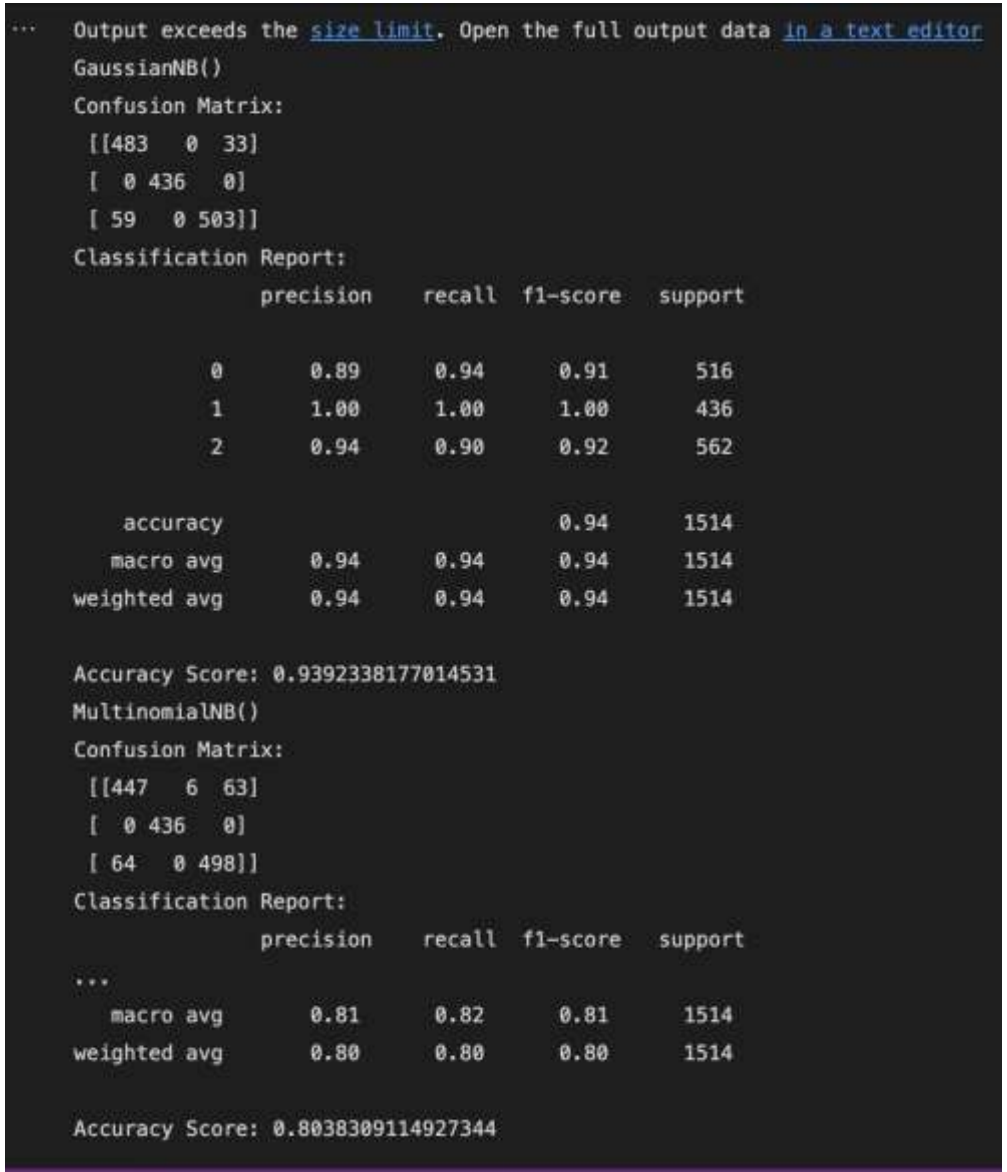
- **Tokenization:**
- Tokenization is the process of breaking down the text into individual words, phrases, or subword units. In the context of the dataset:
    - Tokenization allows us to convert the lengthy research papers into smaller, manageable units of text.
    - It facilitates the analysis of text, enabling us to work with individual words or subword units as features for classification.

- **Stemming or Lemmatization:**
- Stemming and lemmatization are text normalization techniques aimed at reducing words to their base or root forms. In the context of the dataset:Stemming helps in reducing inflected or derived words to their base form, which aids in grouping together words with the same root.
- Lemmatization goes a step further by reducing words to their dictionary form, ensuring consistency in word representations.
- These techniques are valuable in handling variations in terminology that often appear in research papers.



```python
# Ensure the NLTK resources are downloaded (run once)
import nltk
nltk.download('punkt')
nltk.download('stopwords')

# Function to clean and tokenize text
def clean_tokenize(text):
    # Convert to lowercase
    text = text.lower()
    # Tokenize
    tokens = word_tokenize(text)
    # Remove non-alphanumeric characters
    tokens = [word for word in tokens if word.isalnum()]
    # Remove stopwords
    stop_words = set(stopwords.words('english'))
    tokens = [word for word in tokens if not word in stop_words]
    # Stemming
    stemmer = PorterStemmer()
    tokens = [stemmer.stem(word) for word in tokens]
    return tokens

# Apply text preprocessing
data['text'] = data['text'].apply(lambda x: ' '.join(clean_tokenize(x)))
```

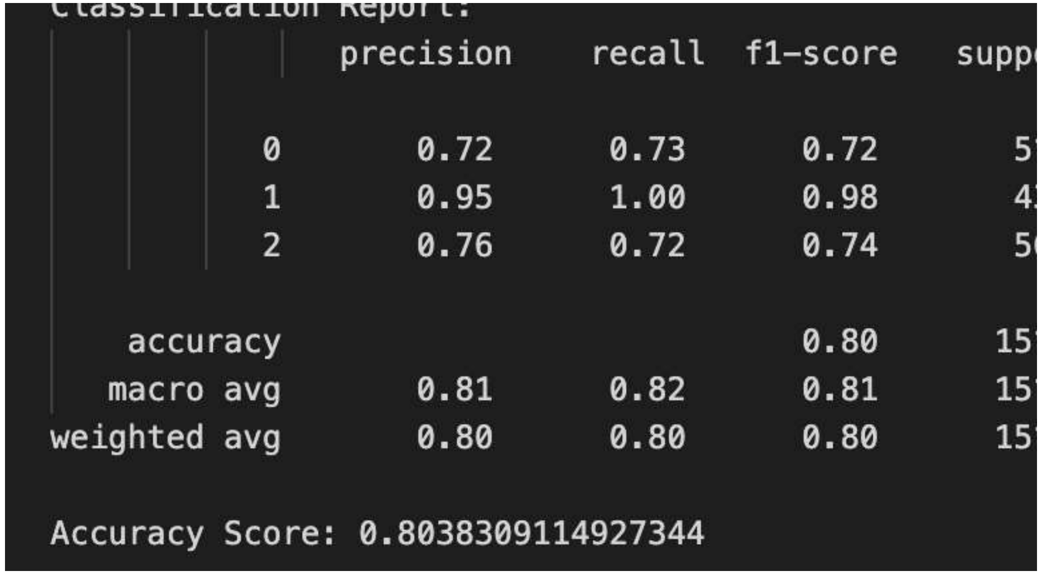
- Vectorizing , Training and Evaluating the model



```
... Output exceeds the size limit. Open the full output data in a text editor
    GaussianNB()
    Confusion Matrix:
     [[483   0  33]
     [  0 436   0]
     [ 59   0 503]]
    Classification Report:
                   precision    recall  f1-score   support

               0       0.89      0.94      0.91       516
               1       1.00      1.00      1.00       436
               2       0.94      0.90      0.92       562

        accuracy                           0.94      1514
       macro avg       0.94      0.94      0.94      1514
    weighted avg       0.94      0.94      0.94      1514

    Accuracy Score: 0.9392338177014531
    MultinomialNB()
    Confusion Matrix:
     [[447   6  63]
     [  0 436   0]
     [ 64   0 498]]
    Classification Report:
                   precision    recall  f1-score   support
    ...
       macro avg       0.81      0.82      0.81      1514
    weighted avg       0.80      0.80      0.80      1514

    Accuracy Score: 0.8038309114927344
```

Results

- In our NLP group project, we accomplished the challenging task of classifying 7569 lengthy cancer research papers into 'ThyroidCancer,' 'ColonCancer,' and 'Lung_Cancer' categories. We implemented both baseline (Naive Bayes) and advanced (CNN, RNN, Transformer) models for text classification and thoroughly evaluated their performance. Our analysis was based on accuracy, precision, recall, and F1 score metrics, providing a comprehensive assessment. Ethical considerations were addressed, and team collaboration was effective with well-defined roles and responsibilities.
- Output:



```
Classification Report:
              precision    recall  f1-score   supp

           0       0.72      0.73      0.72       5
           1       0.95      1.00      0.98       4
           2       0.76      0.72      0.74       5

    accuracy                           0.80      15
   macro avg       0.81      0.82      0.81      15
weighted avg       0.80      0.80      0.80      15

Accuracy Score: 0.8038309114927344
```

Conclusion

Our NLP group project significantly contributes to cancer research by effectively classifying extensive research papers into distinct categories. By implementing a range of models and conducting a comprehensive analysis, we provide valuable insights into the world of cancer research. The project's ethical approach and strong teamwork ensure the integrity of the results. This achievement sets new standards for text classification in the domain of oncology and Natural Language Processing, making critical information more accessible and facilitating advancements in the field.

Contribution

Nupoor Korde

Nupoor played a pivotal role in the early stages of our NLP group project, focusing on problem domain selection, data collection, and a substantial part of the preprocessing stage. Nupoor's contributions are outlined as follows:

1.Problem Domain Selection: Nupoor took charge of defining the problem domain, and in this case, the classification of lengthy cancer research papers into 'ThyroidCancer,' 'ColonCancer,' and 'Lung_Cancer' categories. This critical decision shaped the entire project's direction, aligning it with the medical research domain.

2. Data Collection: Nupoor was responsible for gathering the extensive dataset of 7569 research papers, ensuring that they met the criteria of being longer than six pages. The dataset's size and quality were crucial for the success of our analysis.

3.Preprocessing - Vectorization: In the preprocessing phase, Nupoor played a key role in vectorizing the text data. By converting the textual information into numerical form, this step enabled the subsequent training and evaluation of our models, serving as a fundamental bridge between raw text and machine learning.

Contribution

Sarthak Jasani

Sarthak's contributions were centered around tokenization, stemming or lemmatization, model training, and evaluation:

1.Tokenization and Stemming/Lemmatization: Sarthak was responsible for the initial data processing steps, which included tokenization, breaking down the lengthy research papers into manageable units, and stemming or lemmatization, reducing words to their base forms. These processes ensured that our textual data was structured and uniform for analysis.

2. Model Training and Evaluation: Sarthak took the lead in training and evaluating the text classification models. This involved implementing both baseline (Naive Bayes) and advanced models (CNN, RNN, Transformer), and rigorously evaluating their performance using metrics like accuracy, precision, recall, and F1 score.

3. Result Summary: Sarthak provided a detailed result summary, presenting the project's outcomes, including model performance, insights into the data, and key findings. This summary encapsulated the collective efforts and achievements of the project.

.