CSCE 5290: Natural Language Processing

Project Proposal

Title: A Survey of Text summarization and Classification in different Models using "BBC news Summary" dataset

Group 23 Team Members:

- Kartheek Sure
- Manish Manda
- Rohit Ibrahimpatnam
- Shiva Sayi Shesshank Mamidi

1. Motivation

There is a need for Media organizations and the Readers to efficiently categorize the news information into the respective subjects in this age of information overload. The application of text summarization and classification can be integrated into various mobile and web applications as a feature where end users can set and customize their preferences of News categories , also can read the summarized and condensed text,thereby reducing the manual effort and time .The varied content of the "BBC News Summary" dataset offers a chance of multi-label categorization of data based on the topic categorization despite of the dataset original intent as a summary tool.The main goal of this project is to thoroughly examine and correctly assign the several subject labels of news stories by text summarizing and categorization. This will improve the content structure and make the user to easily navigate through bulk amounts of news contents.

2. Significance

We offer an impartial and thorough assessment of the capabilities and performance of various models by testing them on the same dataset, creating a solid baseline for further studies on news article classification and text summarizing. The results of this project hopefully will provide useful information to help experts choose the best models for practical news processing applications, possibly increasing the effectiveness and precision of content management systems. Our ultimate goal is to actively generate a text classification and summarization based on the News dataset. how different Nlp models are compared and analyze to handle the complexities of Nlp models used.

3. Objectives

Our project aims to provide a reliable, efficient extractive summarization system for BBC news items and classify them . We want to design algorithms that can condense long news articles into brief, interesting summaries while retaining the main idea and facts. Our prior goal is to achieve good accuracy ,fine tune models according to it , and compare the results of different models, study the outputs and also evaluation of models

using F1 and Rouge-1 score, and choose the best model . We also want to use a method named novel sentence importance rating approach that focuses on sentence position, news article and structure factors.

We will provide various visualization techniques and algorithms to our model to ensure the judged automated metrics, summary coherence,accuracy and information evaluation are giving the expected outputs and solutions.

4. Features

Some of the key feature of this project are listed below:

- **Text summarization and Classification:** Summarize in such a way where the key information of the BCC news articles are condensed using various algorithms like TextRank, LSA and BART while doing the evaluation using ROUGE and F1 scores(if possible). Classification of the summaries into different segments like politics, business and sports using different models like SVM, Navie Bayes, BERT by using precision, recall and accuracy metrics.
- **Model Comparison:** Here we will compare the different models results for summarization and classification to take the best outcome of the results. We will also fine-tune the models to improve accuracy, processing time and robustness.
- **Novel Sentence Importance Rating:** Here we are going the implement the unique custom method for the ranking sentence in importance based on the position, content relevance and the news structure to enhance summarization of the accuracy.
- **Data Preprocessing:** Here we will thoroughly clean the dataset and extract text from different folders of summaries and new articles,convert into suitable format ,for any of the HTML tags, special characters, and the missing values. We are also going to tokenize, lemmatize and convert different encodings to unified format.
- **Performance Optimization:** This is another unique feature we are trying to implement other than the NSIR(Novel Sentence Importance Rating) where we are trying to optimize the performance of the processing of the article under quicker time.. This feature will help a lot when we are trying to process a huge amount of articles and also helps to minimize the wait time while maintaining the accuracy of the model.
- **Milestones:** Some of the key milestone of the project include **dataset preprocessing**, **initial model building, model comparison and analyzing the results, concluding based on that results** ,the final report submission with the source code pushed in the GitHub.
- **Deliverables:** Deliverables of the project include preprocessed data of the dataset, up and running model for the summarization and classification, compared results of all the models implemented in the project, app deployment with front end user interface(if possible) and sophisticated report of total project with the working code in the GitHub.

5. Dataset.

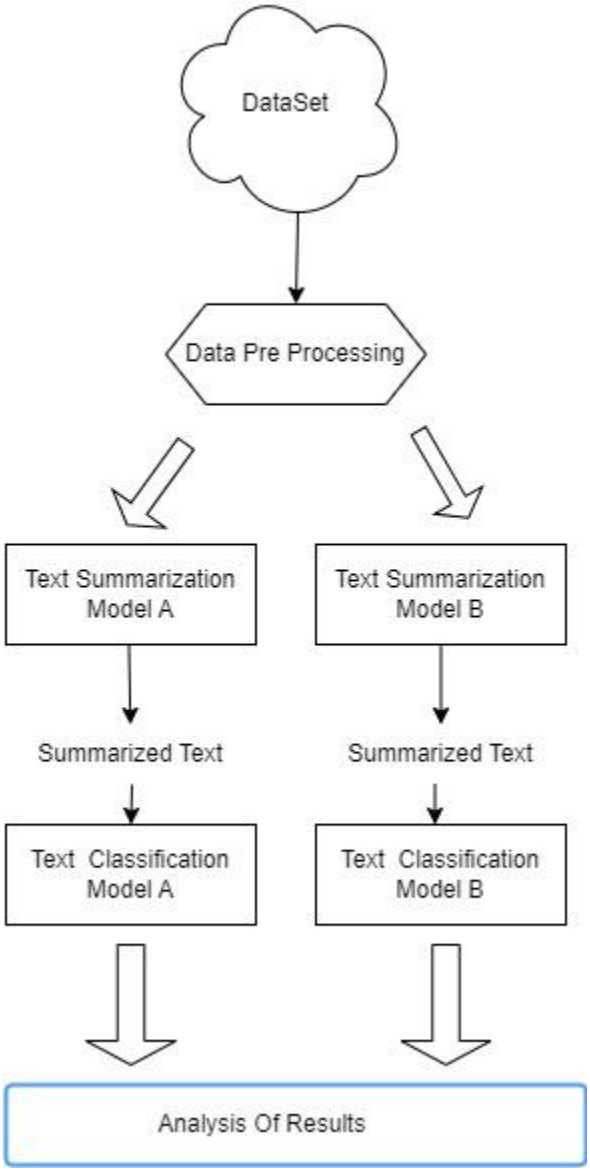
Name- **BBC News Summary**

Source - **https://www.kaggle.com/datasets/pariza/bbc-news-summary**

Size **-9MB**

PreProcessing - Dataset is categorized into different category folders ,it must be extracted from multiple folders with different encodings make sure they are convert into suitable format ,,remove white spaces and handle any missing text summaries ,remove any html tags or special characters present in the text.

6. Visualization



Work Flow

The above image represents comparison of two models in a pipelines approach which includes text summarization and Text Classification

1.Data Preprocessing - Data is preprocessed using different techniques to remove unwanted data,convert it into a suitable format and prepare for the models

2.**Text Summarization Model- News Data** is fed into two different models Model A and Model B for Text Summarization , Models generate Text Summaries as output

3.**Text Classification Model-**The Output of Text summarization Model , Text summary is given as input to classification Model , where summarized text is classified into different categories like Politics, business, Tech sport etc.

4.**Analysis of Results** - The results generated from two Models are compared ,here performance is evaluated ,Models are fine tuned for better performance and Results of Both Models are analyzed

Github Link - https://github.com/manishmanda29/NLP--Project