Ghulam Ishaq Khan Institute of Engineering Sciences and Technology Topi Swabi.

AI407L: PBL for AI Lab

Project: **AI-Powered News Classification and Abstractive Summarization**

Name: Aiman Gohar
Registration No 2020059

Table of Contents

AI-Powered News Classification and Abstractive Summarization

Abstract:

In the past newspapers were only sources of information about what is happening in different parts of the world. Nowadays very few people use newspapers to read news as thousands of online news platforms are present that notify you about an accident or event happening in any part of the world within a few seconds. You can read news through your mobile phone or any device that has internet access, but due to lack of time or a vast amount of news, it becomes very difficult for an individual to read about all the events. This project focuses on this objective to provide a platform to classify and summarize a news article so that a user can read a news article of his interest in less time by reading only the article's main points. While existing solutions often focus on either news classification or summarization, this research uniquely endeavors to optimize both tasks, presenting a comprehensive approach that encompasses a diverse array of machine learning models, including Support Vector Machine, Logistic Regression, Multinomial Naïve Bayes, Random Forest, and Decision Tree Classifier, as well as exploring various word embedding techniques like word2vec, CountVectorizer, TF-IDF, and Fasttext. Abstractive summarization using BART is implemented for the summarization task.
In this project three datasets have been used; for classification, the CNN News Articles dataset collected from 2011 to 2022 [1] , the News Article Category dataset [2], has been used and for summarization BBC News Summary dataset [3] has been used. Through experiments, it can be concluded that SVM with TF-IDF as a word embedding technique achieves remarkable performance as it has achieved precision and recall values of 0.88% on the classification datasets.

Problem Statement:

In the era of digital technology, people prefer to access information through online sources rather than reading paper-based text information. These platforms offer a wealth of information on global events, enabling individuals to stay informed from anywhere. While various online news platforms deliver comprehensive content across different fields such as sports and entertainment, not all readers are interested in every category. This project addresses this challenge by conducting news classification and summarization using a dataset comprising more than 40,000 news articles spanning diverse fields. The project explores a range of machine learning algorithms and word embedding techniques to identify an optimal algorithmic stack for effective news categorization. Additionally, the abstractive summarization technique is utilized to provide concise and categorized news articles, delivering information according to individual reader preferences and ensuring the retention of essential details.

Project Scope:

The project utilizes extensive datasets for news classification and summarization. Various machine learning techniques, along with different word embedding techniques, are employed for training to identify a model and word embedding technique that yields

the best result in the classification task. Additionally, the project aims to generate summaries of news articles using an abstractive summarization technique, employing a transformer model.

Objectives:

To perform news classification and summarization so that readers can acquire knowledge about events happening in different parts of the world in less time, this project focuses on utilizing machine learning and deep learning techniques. Additionally, it employs various web development technologies to develop a user-friendly interface. The project ultimately aims to streamline the process of news consumption, providing a platform where users can upload any article and obtain its class as well as a summary to read it in less time.

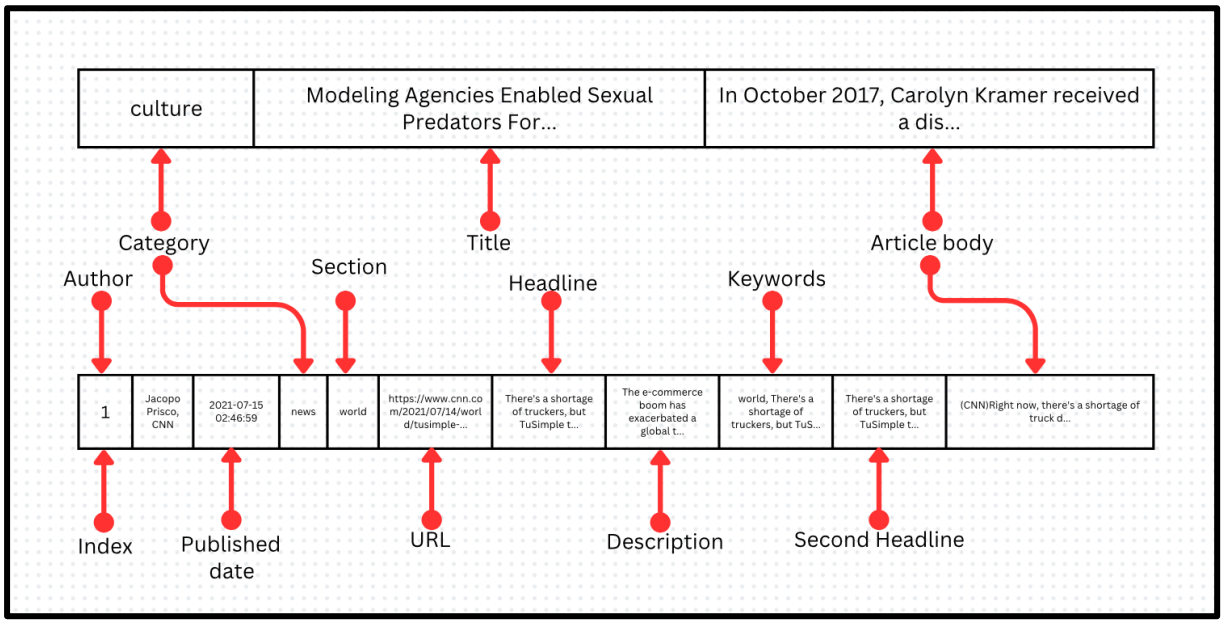
Methodology:

News Classification:

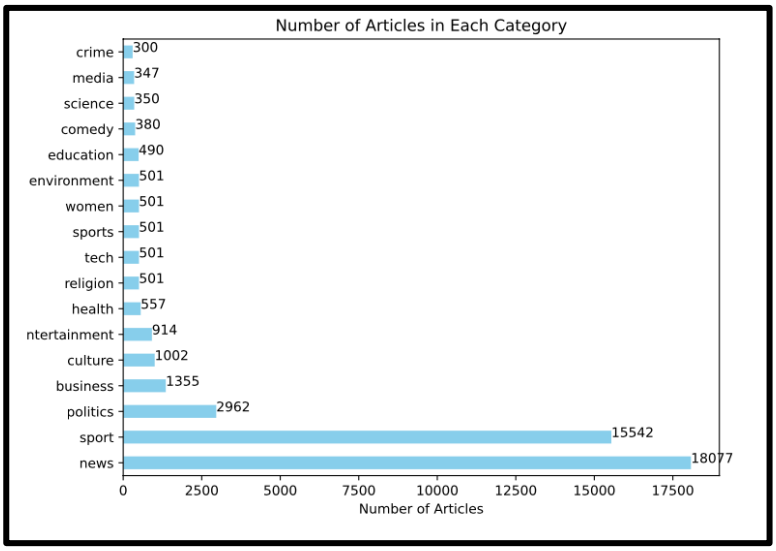
Datasets:

In this project for a news classification task; two datasets have been combined. The CNN news articles dataset was collected for 12 years and the News Article Category Dataset contains news articles belonging to different categories. There are more than 38,000 news articles in the CNN news articles dataset and more than 6000 news articles in the News Article Category dataset. The articles belong to a diverse range of topics such as ARTS & CULTURE, BUSINESS, COMEDY, CRIME, EDUCATION, ENTERTAINMENT, ENVIRONMENT, MEDIA, POLITICS, RELIGION, SCIENCE, SPORTS, TECH, and WOMEN.

The dataset has been divided into 80:20; 80% data is used for training different machine learning models while 20% data is being used for evaluating the performance of various machine learning models.



Sample Datasets Image: The first dataset comprises entries from the News Article Category dataset. The second dataset consists of CNN News Articles spanning the years 2011 to 2022.



CNN News Articles from 2011 to 2022 and News Article Category combined dataset

Preprocessing:

In the data preprocessing phase, two datasets, namely the CNN News Articles dataset from 2011 to 2022, and the News Category Dataset are employed and they are combined into a single dataset for the news classification task. The preprocessing steps involve tokenization into words and sentences, followed by the removal of punctuation, stop words, and other special characters. Subsequently, lemmatization is applied to standardize word forms. After lemmatization, vectors of words are generated, and features are extracted using TF-IDF, GloVe, Fasttext, and Word2Vec word embedding techniques. These embeddings serve as input for machine learning classifiers.

Word Embedding Techniques:

Different word embedding techniques have been used in the project for feature extraction. The word embedding techniques used in this project include TF-IDF, CountVectorizer, FastText, and Word2Vec word embedding techniques. Each word embedding technique is used with each machine learning model to find the best pair of word embedding techniques and machine learning models to achieve the goal of classifying a news article.

Machine Learning Models:

After obtaining word embeddings by implementing different word embedding techniques; different machine learning models including Support Vector Machine (SVM), Decision Tree Classifier (DTC), Random Forest (RF), Multinomial Naïve Bayes (MNB), and Logistic Regression has been paired with different word embedding techniques to find the optimal and best combination of word embedding technique and machine learning model for news classification.

Evaluation Methods:

To evaluate the performance of different combinations of word embedding techniques and machine learning models; precision, recall, f1-score, and accuracy have been used as evaluation metrics.

Experimental Settings:

Initial Model Evaluation:

For news articles classification, different combinations of word embedding techniques and machine learning models are trained and the evaluation metrics showed that among all the trained models; SVC with TF-IDF as a word embedding technique showed superior results. As the news classification dataset is imbalanced, precision and recall evaluation metrics values are studied to evaluate different machine learning models' performance.

Data Balancing:

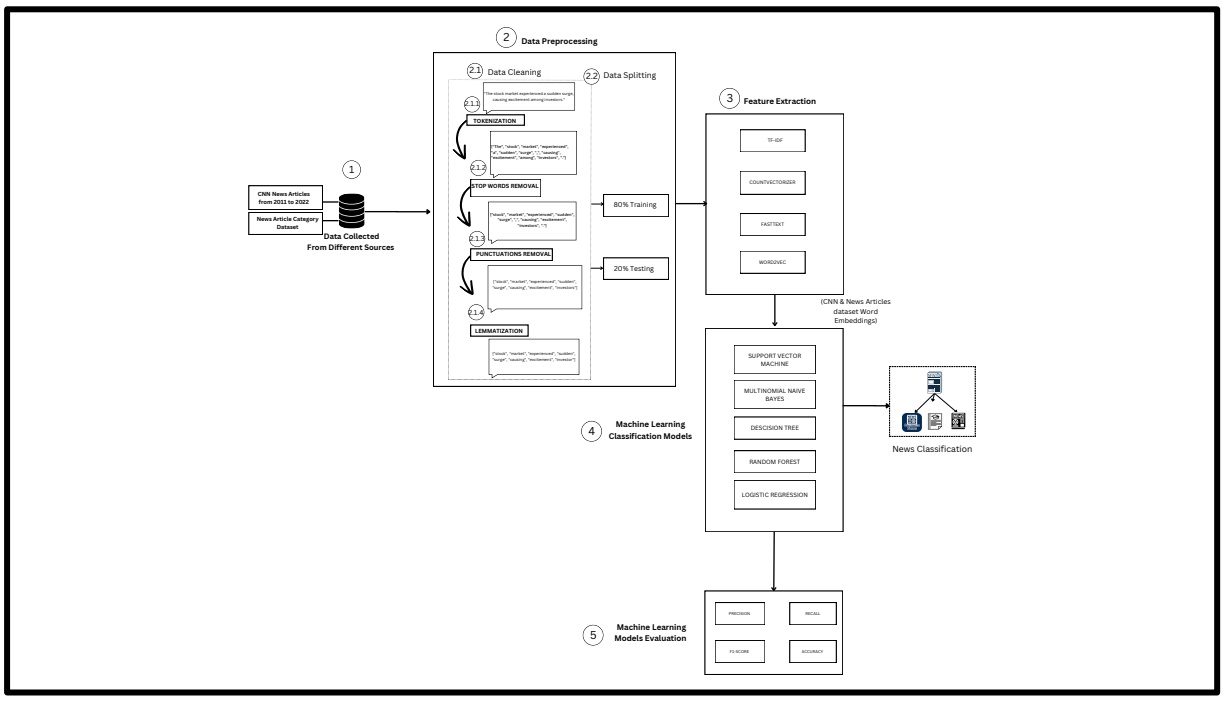
To address the issue of data imbalance, the random under-sampling technique has been used to reduce the imbalance in the dataset, and SVC models with different kernels (such as linear, rbf, etc.) and using TF-IDF as a word embedding technique were trained and evaluated using evaluation metrics. Logistic Regression was also implemented for comparison. Results revealed that SVC with linear kernel consistently outperformed other models.

Hyperparameter Tuning:

Given SVM's consistent performance across experiments, the complete classification dataset was trained using Support Vector Classifier (SVC), which was then hyper-tuned using various hyperparameters. The resulting optimized model was saved and utilized for the classification task. The table below illustrates the different parameters and their corresponding values:

Hyperparameter	Values
C	0.1, 1
gamma	1, 0.1
Kernel	linear, poly, rbf, sigmoid

Workflow Figure:

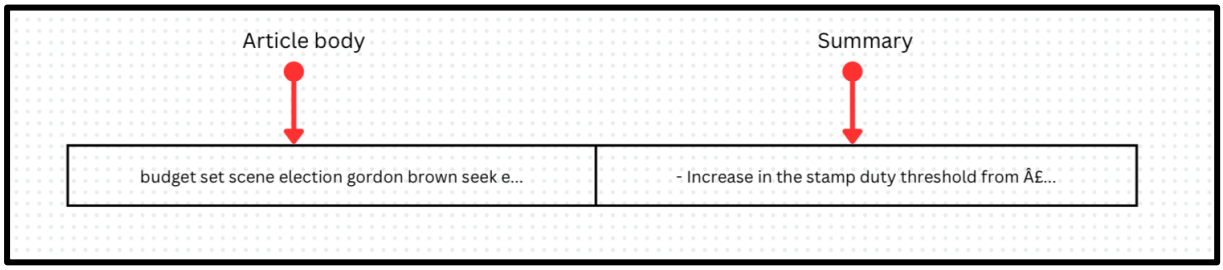


Workflow Figure. (1) Two datasets are utilized in this study. (2.1) The datasets undergo various cleaning processes, including tokenization, stop word removal, punctuation removal, and lemmatization techniques as part of the data preprocessing section. (2.2) After preprocessing, the combined News Article Category and CNN News Article datasets are divided into training and testing data. (3) The training data is then vectorized using different word embedding techniques. (4) The features extracted by applying different word embedding techniques are utilized by various Machine Learning Classification Models to perform the News Classification Task. (5) Different classification models are evaluated using test data, employing accuracy, precision, recall, and F1-score as evaluation metrics.

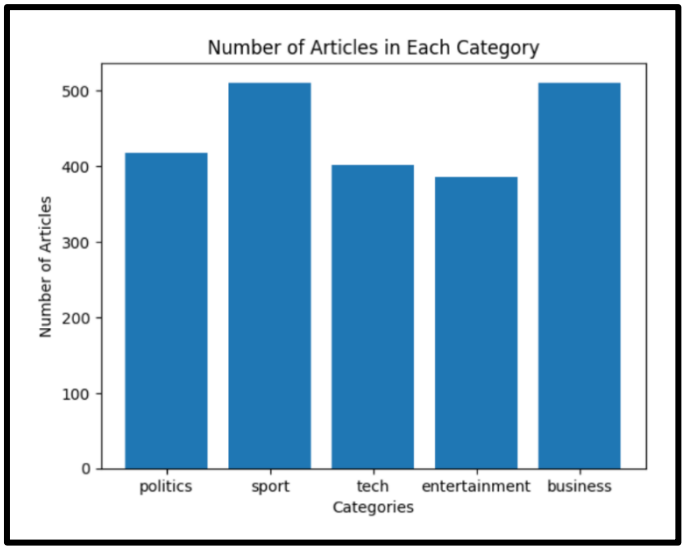
News Summarization:

Dataset:

The dataset [3] utilized in this project for the task of news summarization is the BBC News Summary dataset which was collected from the BBC news website in the years of 2004-2005. The dataset includes news articles belonging to five different domains, including politics, tech, sport, entertainment, and business. The dataset contains news article summaries along with the news articles and this fact makes this dataset perfect for training a machine learning or deep learning model to perform news summarization.



Sample Dataset



Number of Articles in Each Category of BBC News Summary Dataset

Preprocessing:

The dataset is tokenized and split into different tokens using Hugging Face Auto Tokenizer. The tokenizer performs all the necessary data preprocessing steps such as lowercasing, removing special characters, etc. The dataset is divided into 70:30; 70% of the dataset is used for training the BART model, 20% is used for validation purposes and 10% of the dataset is used as testing data for evaluating the performance of the trained model.

Summarization Model:

This project uses the BART (Bidirectional and Auto-Regressive Transformers) model pre-trained on a dataset consisting of 295,174 news articles scrapped from a Mexican Newspaper, along with its summary. For simplicity, the Spanish news articles were translated into English language using a natural language processing model. BART is fine-tuned on the BBC News Summary dataset [3], which consists of articles and their corresponding summaries across various categories such as politics, sports, technology, entertainment, and business.
BART is a sequence-to-sequence model, it consists of an encoder-decoder architecture with multiple layers of self-attention and feed-forward neural networks.

Evaluation Method:

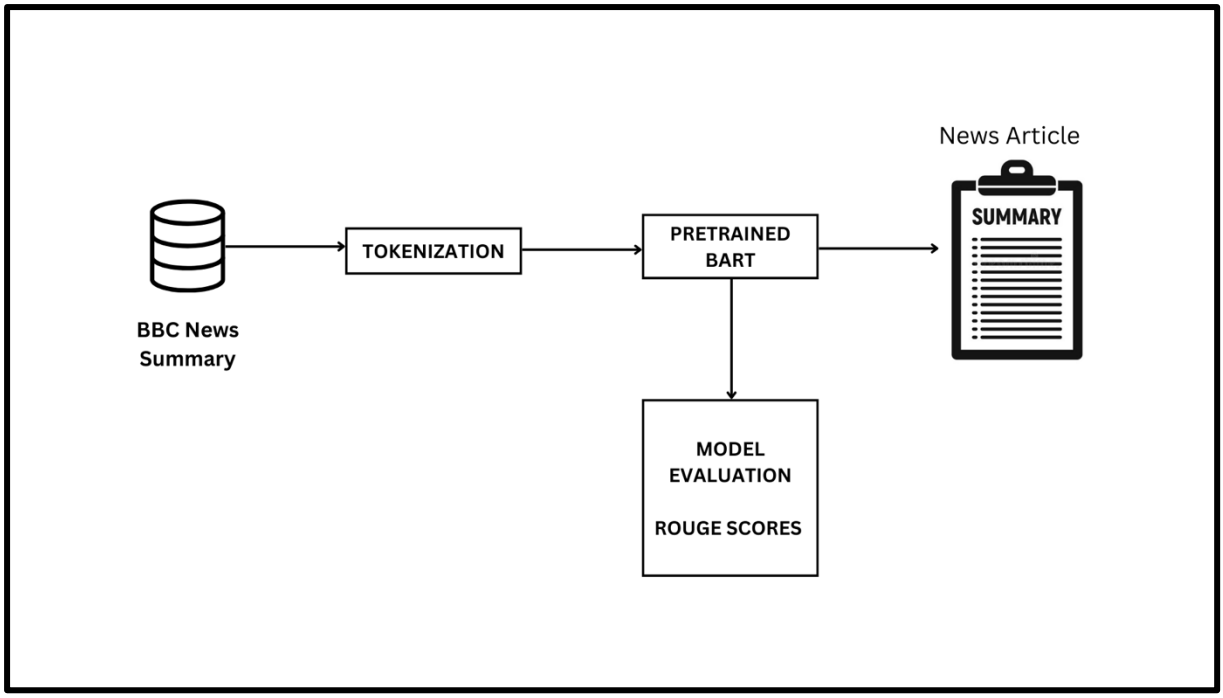
The performance of the fine-tuned BART model is evaluated on the test dataset and ROUGE scores are computed to assess the quality of generated summaries.

Experimental Settings:

Different experimental settings used for fine-tuning Hugging face BART model on the BBC News Summary dataset are shown in the table below.

Setting	Value
Train Batch per device	1
Evaluation Batch per device	1
Epochs	8
Gradient Accumulation Steps	1

Workflow Figure:

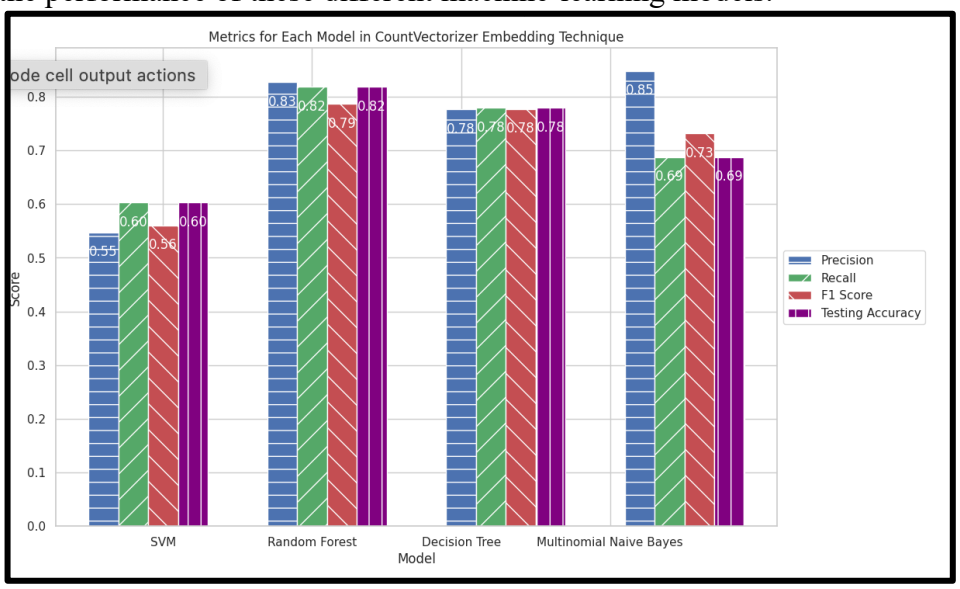


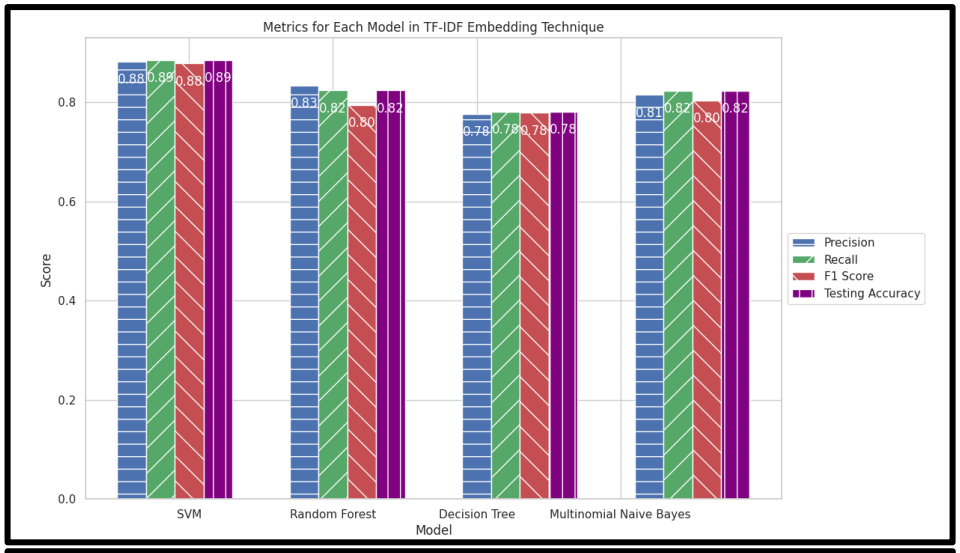
The BBC News Summary Dataset, after preprocessing, is employed for the news summarization task using Absractive Summarization Technique ROUGE is employed as evaluation metrics to assess the performance of the news summarization task.

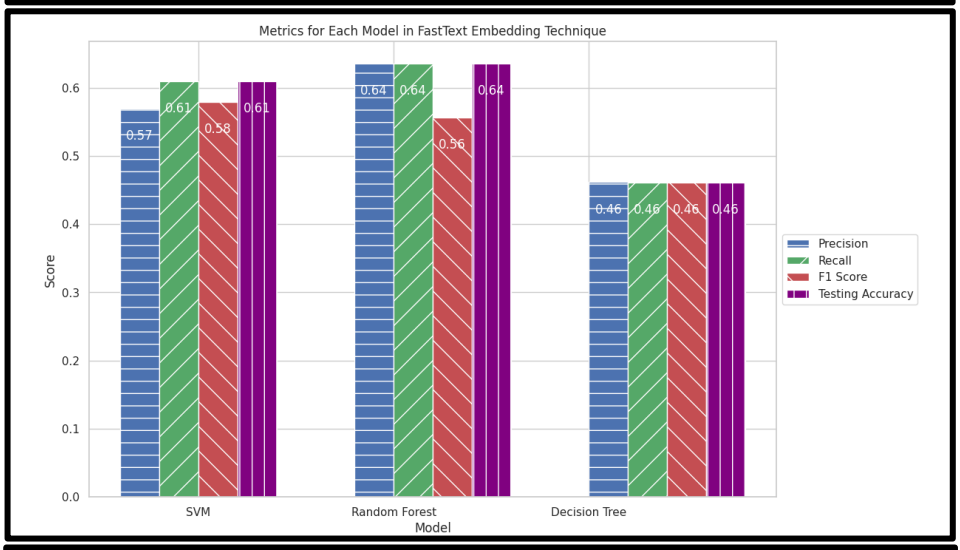
Results:

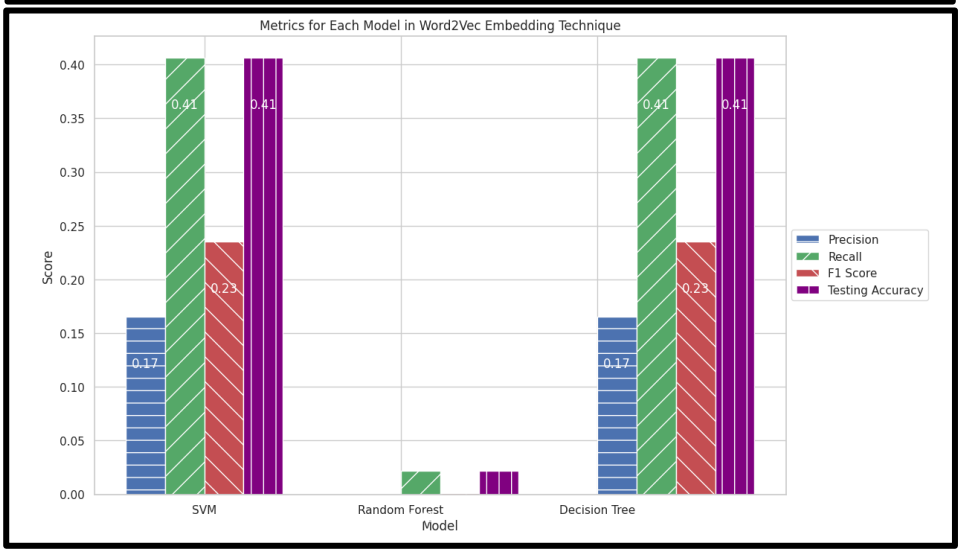
News Classification:

After applying all the necessary preprocessing steps to clean the dataset, various machine learning models were trained using different word embedding techniques. The figures below illustrate the performance of these different machine-learning models.

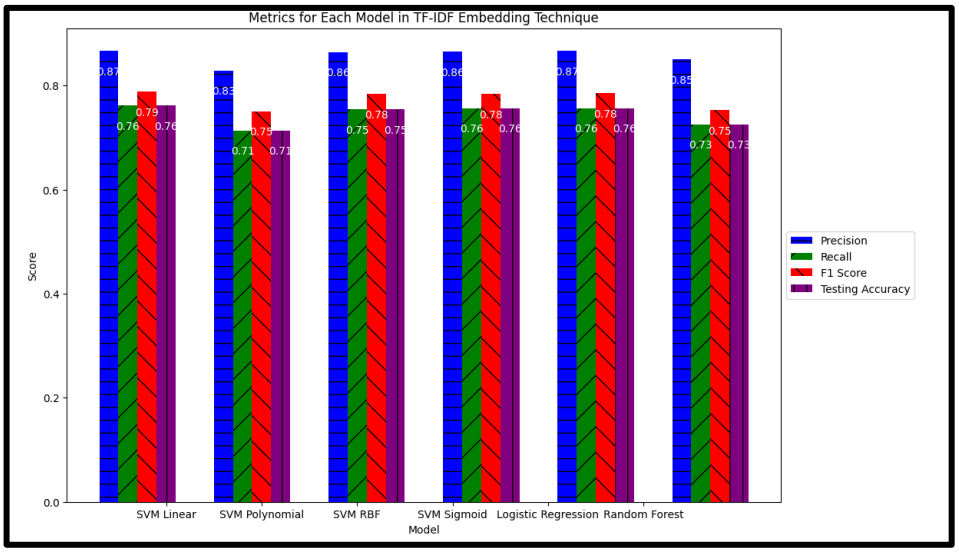








Since the dataset is imbalanced, random under sampling was applied. Various kernels of Support Vector Classification (SVC) were trained using TF-IDF as the word embedding technique. SVC and Random Forest with TF-IDF exhibited superior results. Logistic Regression was also applied to assess its performance on our news classification dataset. The figure below displays the results of this training.



Given that SVC with TF-IDF as the word embedding technique demonstrated excellent results in the news classification task, it was further trained with hyperparameter tuning conducted. The best model was saved for deployment purposes.
Below are the results of SVC trained using liner kernel.

	precision	recall	f1-score	support
0	0.63	0.81	0.71	251
1	0.84	0.75	0.79	76
2	0.47	0.54	0.50	52
3	0.70	0.85	0.77	199
4	0.67	0.86	0.75	97
5	0.72	0.75	0.74	216
6	0.54	0.66	0.59	101
7	0.44	0.65	0.53	104
8	0.73	0.67	0.70	76
9	0.92	0.83	0.87	3567
10	0.74	0.87	0.80	596
11	0.60	0.77	0.67	103
12	0.37	0.61	0.46	74
13	0.98	0.97	0.97	3236
14	0.70	0.67	0.69	111
15	0.47	0.67	0.55	90
accuracy			0.86	8949
macro avg	0.66	0.74	0.69	8949
weighted avg	0.88	0.86	0.87	8949

Testing Accuracy: 0.86

News Summarization:

Initially, I implemented the Facebook pre-trained BART model and tuned it on my training dataset. Its results are shown below.

16	0.079500	0.261793
17	0.058900	0.263951
18	0.076600	0.270139
19	0.069700	0.269640
20	0.049000	0.272139
21	0.065200	0.276096
22	0.054300	0.276363
23	0.053600	0.279557
24	0.054200	0.282173
25	0.051300	0.284094
26	0.044500	0.284857
27	0.037900	0.286117
28	0.041700	0.285915
29	0.044600	0.287049
30	0.040600	0.286929

Epoch	Training Loss	Validation Loss
1	3.298900	2.520117
2	0.436700	0.313270
3	0.305900	0.251232
4	0.227600	0.243481
5	0.231600	0.236206
6	0.197300	0.237339
7	0.148500	0.237993
8	0.169400	0.237195
9	0.131900	0.236959
10	0.106000	0.246740
11	0.108400	0.246171
12	0.112700	0.253294
13	0.093900	0.255863
14	0.105300	0.256172
15	0.104300	0.257465

	rouge1	rouge2	rougeL	rougeLsum
0	(0.758067872900367, 0.7058285250587599, 0.7263...	(0.6644735624466196, 0.6216573293960086, 0.637...	(0.539825093228228, 0.5083721948759978, 0.5195...	(0.5425686140371844, 0.5105755500100964, 0.523...
1	(0.7784152883529202, 0.7332614937163153, 0.747...	(0.6925242634550548, 0.6560170920886388, 0.666...	(0.5684960530936642, 0.5381122733545944, 0.547...	(0.5694086397649943, 0.539755192036399, 0.5484...
2	(0.8015358269667222, 0.762353215118614, 0.7702...	(0.7177137496480084, 0.6867743875439348, 0.692...	(0.5988107127052793, 0.5713892888003798, 0.577...	(0.5971889665782414, 0.5695437069824367, 0.576...

Rouge Values

Later on, I attempted to train the pre-trained BART model from Hugging Face, which was specifically fine-tuned for news article summarization on a large corpus of Mexican articles translated into English. Despite multiple attempts, I encountered persistent memory issues during training. As a result, I was only able to complete training up to 8 epochs before experiencing crashes.

Epoch	Training Loss	Validation Loss
1	0.022200	0.056141
2	0.053200	0.051751
3	0.023300	0.052947
4	0.014700	0.052331
5	0.017700	0.056438
6	0.006500	0.058420
7	0.078800	0.057094
8	0.010900	0.062169

Technical Setup and Deployment:

For b oth the news classification and news summarization tasks, different techniques are employed and tested to get the best models and techniques and models so that we can make a system that can classify any news article text and summarize the article into main points.
The following outlines the implementation details:

News Classification:

Various machine learning models and word embedding techniques are experimented with using our extensive dataset to find the best classification model and word embedding technique. Ultimately SVM along with TF-IDF word embedding technique gives the best results in terms of different evaluation metrics. The model is hyper-tuned, trained, and saved for deployment purposes.

News Summarization:

The pre-trained hugging face BART model is trained on the BBC News Summary dataset and saved for deployment purposes.

Model Deployment:

Flask is used for backend implementation and RESTful API is used for seamless integration of classification and summarization functionalities.

User Interface:

To create a responsive and user-friendly interface, HTML, CSS, and JavaScript are employed.

References

[1] H. Unger, "CNN News Articles from 2011 to 2022," [Online]. Available: https://www.kaggle.com/datasets/hadasu92/cnn-articles-after-basic-cleaning.

[2] B. Timilsina, "News Article Category Dataset," [Online]. Available: https://www.kaggle.com/datasets/timilsinabimal/newsarticlecategories.

[3] P. Sharif, "https://www.kaggle.com/datasets/pariza/bbc-news-summary," [Online]. Available: BBC News Summary.

Appendices:

Web Interface: