News Classification using Machine Learning: A Comparative Study of Traditional and Transformer-based Models

Suraj T C
stelugar@umd.edu
University of Maryland

***Abstract*—In the era of information abundance, the accurate categorization of news articles plays a pivotal role in enabling efficient information retrieval and understanding news trends. This study presents a comprehensive study on news classification using machine learning, comparing the efficacy of traditional and transformer-based models. The project centers around the AG News dataset and employs Logistic Regression and Support Vector Classifier (SVC) as representatives of traditional methods, while also exploring the capabilities of transformer models such as BERT and RoBERTa. The comparative analysis includes model training, evaluation, and performance metrics, shedding light on the strengths and limitations of each approach.**

***Index Terms*—News Classification, AG News Dataset, Natural Language Processing, Logistic Regression, Support Vector Classifier (SVC), BERT, RoBERTa, Text Classification.**

I. Introduction

In the rapidly evolving landscape of information dissemination, the accurate classification of news articles emerges as a critical facet for comprehending and navigating through vast datasets. This study investigates the effectiveness of machine learning methodologies in the domain of news classification, undertaking a comparative review that juxtaposes traditional approaches, exemplified by the Support Vector Classifier (SVC), against state-of-the-art transformer-based models such as BERT and RoBERTa [1]–[3]. Anchored by the AG News dataset [9], the study explores the nuances of model training and evaluates the performance of each method while addressing challenges encountered in the process [4], [5]. The research scrutinizes data preprocessing intricacies and model architectures [**?**]. The ensuing insights contribute to the discourse surrounding the selection of appropriate models for news classification tasks, offering valuable implications for practitioners and researchers alike in the realm of natural language processing and text classification [7], [8].

A. Background

The increasing volume and diversity of news articles pose significant challenges for effective categorization. Traditional methods of news classification have seen notable advancements, but the rapid evolution of machine learning techniques has opened new avenues for improved accuracy and efficiency in handling various types of news content.

B. Problem Statement

News classification is inherently complex due to the dynamic nature of language and the diverse topics covered in news articles. The existing methods may struggle to adapt to the evolving landscape of news content. This study aims to address the challenges in accurately classifying news articles by conducting a comparative study between traditional machine learning approaches, and transformer-based models.

C. Objectives

The primary objectives of this study are threefold. Firstly, to explore the recent advancements in machine learning techniques for news classification. Secondly, to tackle the challenges associated with accurately categorizing diverse news articles. Lastly, to conduct a comprehensive comparative study between traditional methods and transformer-based models to discern their relative strengths and limitations. By achieving these objectives, this study aims to provide valuable insights for practitioners and researchers in the field of natural language processing and news analysis.

II. Literature Review

News classification, a fundamental task in natural language processing, has witnessed a significant evolution spanning traditional machine learning to recent advancements with transformer-based models. The literature reveals a nuanced progression that reflects the field's continuous efforts to enhance accuracy, efficiency, and adaptability in categorizing news articles.

Traditional approaches to news classification often employed methods such as Support Vector Machines (SVMs) and Naive Bayes. Cortes and Vapnik's work on Support-Vector Networks demonstrated the effectiveness of SVMs in various classification tasks, including text categorization [4]. These methods, while successful to a certain extent, faced challenges in handling the complexity and variability inherent in news content.

The turning point in news classification arrived with the introduction of transformer-based models. Vaswani et al. presented the groundbreaking "Attention is All You Need," introducing the Transformer architecture [1]. This architectural paradigm shift enabled models like BERT (Bidirectional Encoder Representations from Transformers) and RoBERTa

(Robustly optimized BERT approach) to capture contextual information more effectively [2], [3].

BERT, introduced by Devlin et al., leverages bidirectional attention mechanisms to understand the context of words in a sentence, allowing it to capture complex relationships within the text [2]. RoBERTa further refines this approach by optimizing the training procedure and hyperparameters, resulting in improved performance on various natural language processing tasks [3].

The AG News dataset, a widely used benchmark for news classification, has played a pivotal role in evaluating the effectiveness of different models [9]. This dataset, containing news articles across multiple categories, provides a diverse and standardized platform for comparing the performance of traditional and transformer-based models.

The literature review underscores a paradigm shift in news classification methodologies, with transformer-based models demonstrating superior performance in capturing semantic relationships and contextual information. The following sections of this research will delve into a comparative analysis, aiming to elucidate the strengths and limitations of these models in the specific context of news classification.

III. Methodology

A. Data Collection and Preprocessing

- **Data Collection:** The AG News dataset was obtained using HuggingFace's [12] datasets library, a comprehensive repository of various datasets that streamlines the data acquisition process, eliminating the need for additional data collection [9].
- **Data Preprocessing:** The preprocessing pipeline involved several key steps to ensure the quality and readiness of the data for subsequent analysis and model training. The text data was cleaned by removing URLs, HTML tags, and special characters that were not punctuation or alphanumeric. This step was crucial for mitigating noise and ensuring a consistent input format for downstream tasks. PySpark [10], a powerful distributed data processing framework, was employed for efficient data processing, offering parallelized operations to handle the large dataset effectively.
- **Exploratory Data Analysis (EDA):** Exploratory Data Analysis (EDA) was conducted to gain deeper insights into the characteristics of the AG News dataset. The analysis included three visualizations: a count plot displaying the distribution of samples across different classes in the training set, a histogram illustrating word counts by label, and a boxplot depicting sentiment polarity by class. These visualizations provided a comprehensive overview of the dataset, aiding in understanding class distributions, word count patterns, and sentiment characteristics [**?**].

Figure 1 illustrates the class distribution of the training set, revealing that each class is represented by 30,000 samples. Similarly, the test set comprises 1,900 samples for each class, resulting in a total of 127,600 samples evenly distributed across the four classes. The balanced distribution is beneficial for classification tasks, ensuring that the model is exposed to an equal representation of each class, thereby avoiding biases associated with imbalanced datasets.

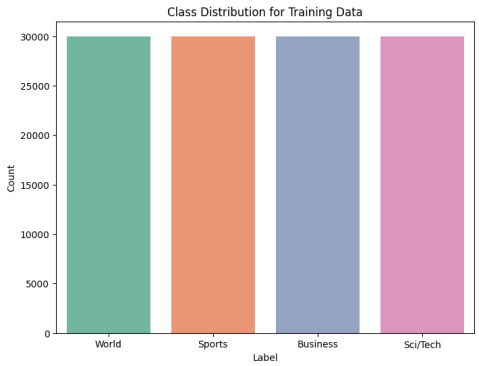

Fig. 1. Count of Samples by Class in the Training Set

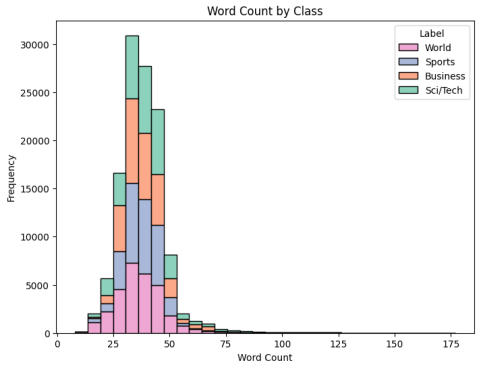

Fig. 2. Histogram of Word Counts by Label

Figure 2 provides insights into the average length of news articles categorized by class. Analyzing the distribution of word counts for each class is valuable in understanding the variation in article lengths. Such information is crucial for evaluating whether certain classes have lengthier or shorter articles, potentially influencing the performance of the classification system.

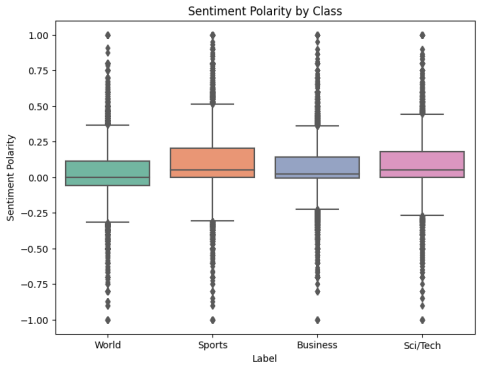

Fig. 3. Sentiment Polarity by Class

Figure 3 portrays the inherent bias within the collected news articles. Notably, a significant portion of the news articles exhibits sentiment polarity close to neutral, suggesting a lack of pronounced bias in the dataset. Additionally, the sentiment

analysis reveals that sports-related news articles tend to lean slightly towards the positive side. This inclination may be attributed to the nature of sports news, which predominantly covers victorious events and winning teams rather than losses.

Figures 1, 2, and 3 showcase the results of the exploratory data analysis, providing valuable insights into the dataset.

- **Tokenization and Text Cleaning:** For effective tokenization and text cleaning, the BeautifulSoup [11] library was employed to remove HTML tags, and regular expressions were used to eliminate URLs and special characters. The resulting clean text was ready for further processing and analysis.
- **Data Splitting:** The AG News [9] dataset was split into training and testing sets using HuggingFace's [12] datasets library. Approximately 120,000 samples from the training set and 1,900 samples from the test set were used for the whole experimentation.

B. Traditional Approach: Logistic Regression and Support Vector Classifier (SVC)

- **Implementation:** The TF-IDF [14] (Term Frequency-Inverse Document Frequency) vectorizer was employed to convert the raw textual data into numerical vectors. TF-IDF [14] is a widely-used technique in natural language processing that evaluates the importance of each word in a document relative to a collection of documents. By assigning weights based on the frequency of terms in a document and inversely proportional to their occurrence across the entire dataset, TF-IDF [14] captures the significance of words in distinguishing documents. This numerical representation aids in transforming the text data into a format suitable for machine learning models, particularly traditional approaches such as Support Vector Classifier (SVC) and logistic regression. The TF-IDF vectors serve as input features for these models, allowing them to discern patterns and relationships within the dataset during training.
  The Logistic Regression and Support Vector Classifier (SVC) from scikit-learn [13] was selected as the representative of traditional machine learning methods for news classification. The model was trained on the preprocessed dataset.
- **Challenges and Solutions:** During the initial implementation, challenges arose regarding computational resources and model performance. To address this, the training data was reduced to 24,000 samples per class, balancing accuracy and resource efficiency.
- **Model Evaluation:** The Logistic Regression model achieved and accuracy of 83.65 while SVC achieved an accuracy of 86.37, and a confusion matrix was generated to assess its performance across different classes.
  Support Vector Classifier (SVC) and Logistic Regression are widely used algorithms in text classification, but they come with certain limitations. SVC relies on finding an optimal hyperplane to separate classes in a high-dimensional space, making it computationally expensive and prone to challenges when dealing with large datasets. Additionally, SVC might struggle with the curse of dimensionality, where the number of features surpasses the number of samples. Logistic Regression, on the other hand, assumes a linear relationship between features, which may not capture complex patterns in text data effectively. Both models might encounter difficulties in handling non-linear relationships and might not perform optimally when faced with intricate semantic structures in natural language. Furthermore, they might be sensitive to noise and outliers, impacting their robustness. To overcome these limitations, recent advancements in natural language processing have led to the emergence of transformer-based models, such as BERT and RoBERTa, which have demonstrated superior performance in capturing intricate patterns and relationships within textual data.

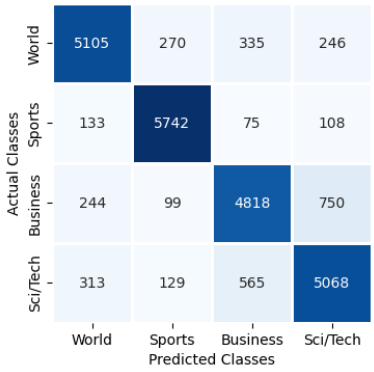

Fig. 4. Confusion Matrix for Support Vector Classifier (SVC)

Figure 4 provides a detailed breakdown of the SVC's predictions.

C. Deep Learning Approaches: BERT and RoBERTa

In this section, we delve into the application of advanced deep learning approaches, specifically BERT [2] (Bidirectional Encoder Representations from Transformers) and RoBERTa [3] (Robustly optimized BERT approach), for the task of news classification. These transformer-based models, pre-trained on vast amounts of textual data, have demonstrated exceptional capabilities in capturing intricate language patterns and contextual information. We explore their respective tokenization strategies, model architectures, training parameters, and optimization strategies, shedding light on the intricacies of employing BERT and RoBERTa in the context of news article classification.

- **BERT (Bidirectional Encoder Representations from Transformers):**
  BERT [2], a groundbreaking transformer-based model, has revolutionized natural language processing tasks by leveraging bidirectional context understanding. Pre-trained on vast corpora, BERT captures intricate linguistic relationships and semantic nuances. Its bidirectional nature allows it to consider both preceding and succeeding words in a sentence, fostering a comprehensive

understanding of context. As we delve into the specifics of BERT's tokenization, model architecture, training parameters, and optimization strategies, we uncover the mechanisms that contribute to its prowess in news article classification.

  - **Tokenization:** Employed BERT [2] tokenizer to convert text data into tokens, capturing contextual information.
  - **Model Architecture:** Utilized the pre-trained BERT model from HuggingFace's [12] transformers library, leveraging its bidirectional transformer architecture.
  - **Training Parameters:** Trained for 5 epochs with a batch size of 32, resulting in 3,750 iterations per epoch.
  - **Optimization Strategies:** Applied a learning rate of 1e-4 to fine-tune BERT [2] for the news classification task.

- **RoBERTa (Robustly Optimized BERT Approach):** RoBERTa [3], an evolution of the BERT architecture, refines the pre-training process to enhance model performance. It employs advanced optimization techniques, omitting the original BERT's [2] next sentence prediction objective and dynamically adjusting hyperparameters. As we navigate through RoBERTa's [3] tokenization, model architecture, training parameters, and optimization strategies, we gain insights into the refinements that contribute to its robustness in deciphering and classifying diverse news articles.
  - **Tokenization:** Utilized RoBERTa [3] tokenizer for data tokenization, ensuring compatibility with the RoBERTa model.
  - **Model Architecture:** Employed the pre-trained RoBERTa model from HuggingFace's [12] transformers library, known for its robustness.
  - **Training Parameters:** Trained for 5 epochs with a batch size of 32, resulting in 3,750 iterations per epoch.
  - **Optimization Strategies:** Applied a learning rate of 1e-5 for RoBERTa [3] to fine-tune the model for news classification.

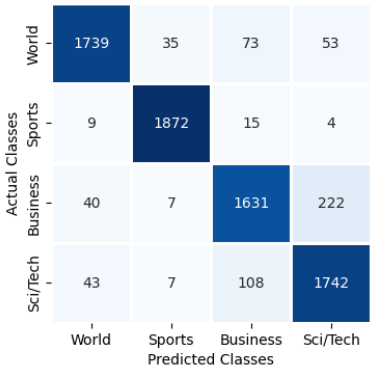

Fig. 5. Confusion Matrix for Support Vector Classifier (SVC)

Figure 5 provides a detailed breakdown of the RoBERTa's predictions.

D. Training Platform

The training of BERT and RoBERTa was conducted on a Google Colab platform equipped with a T4 GPU. The average training time per epoch was approximately 85 minutes. Both models were trained for a total of 5 epochs. However, it was observed that the validation accuracy and loss did not show significant improvement after the third epoch. To ensure the retention of the best-performing model, a mechanism was implemented to automatically save the model checkpoint based on the validation loss.

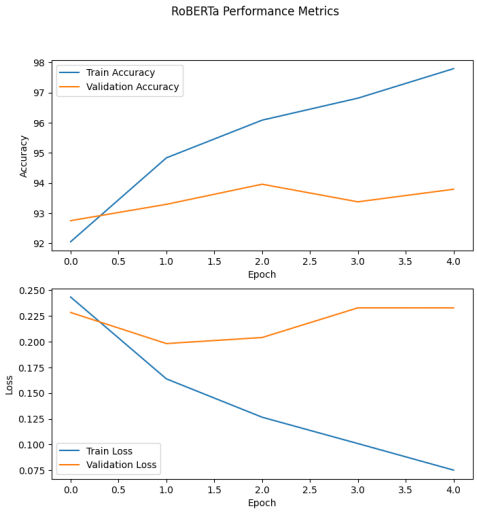

Fig. 6. Confusion Matrix for Support Vector Classifier (SVC)

During the training process, RoBERTa's performance was further analyzed by plotting training accuracy versus validation accuracy and training loss versus validation loss (Fig. 6). These visualizations offer insights into the convergence and generalization characteristics of the RoBERTa model across epochs.

E. Performance Advantages of Transformers over Traditional Methods

The adoption of advanced transformer models, specifically BERT [2] (Bidirectional Encoder Representations from Transformers) and RoBERTa [3] (Robustly optimized BERT approach), for news classification offers substantial advantages over traditional models. Unlike traditional models such as Support Vector Classifier (SVC) [4] and Logistic Regression, BERT and RoBERTa excel in capturing intricate language patterns, contextual nuances, and long-range dependencies within news articles [5]. The bidirectional nature of these transformer-based models allows them to consider the entire context of a word by analyzing both preceding and succeeding words, enabling a more comprehensive understanding of semantics and meaning. Moreover, BERT [2] and RoBERTa [3] leverage pre-training on vast amounts of diverse textual data, enabling them to grasp domain-specific features and adapt effectively to varied linguistic styles. This inherent adaptability makes them particularly potent for news classification tasks where language nuances and context play a crucial role. The superior

performance of BERT [2] and RoBERTa [3] in handling such complexities signifies a paradigm shift in natural language processing, offering a more nuanced and context-aware approach to news categorization compared to the rigid feature-based methods employed by traditional models.

IV. Results

A. Traditional Approach Results

Logistic Regression exhibited solid performance in news classification, achieving an accuracy of *83.65%*. Precision, recall, and F1-score metrics emphasized the model's competence in categorizing diverse news content. Logistic Regression is valued for its simplicity, ease of interpretation, and suitability for high-dimensional data.

However, Logistic Regression comes with its own set of limitations. Challenges in handling large datasets and the need for meticulous hyperparameter tuning may impact its scalability. Additionally, the model may struggle to capture intricate contextual dependencies present in news articles, a facet where more advanced models prove advantageous.

The Support Vector Classifier (SVC) model exhibited commendable performance in the task of news classification. The accuracy achieved by the SVC model was *86.38%*, demonstrating its capability to effectively categorize news articles. Precision, recall, and F1-score metrics further underscored the model's proficiency in handling diverse news content.

Strengths of the SVC model include its simplicity, interpretability, and robustness in handling high-dimensional data. The utilization of a kernel trick, such as the radial basis function (RBF) kernel, allowed the model to capture complex relationships within the data.

However, the SVC model has its limitations. It may struggle with scalability when dealing with large datasets, and hyperparameter tuning can be crucial for optimizing its performance. Additionally, the SVC model might not fully capture intricate contextual dependencies present in news articles, which can be better addressed by more advanced models.

In the subsequent sections, we delve into the results obtained from transformer-based models, BERT and RoBERTa, providing a comparative analysis with the traditional approach.

B. Deep Learning Model Results

The deep learning models, BERT (Bidirectional Encoder Representations from Transformers) and RoBERTa (Robustly Optimized BERT Approach), were instrumental in advancing the field of news classification. Leveraging pre-trained language representations, these models demonstrated remarkable testing accuracies. BERT, with its bidirectional contextual embeddings, and RoBERTa, with optimizations enhancing training dynamics, showcased their prowess in capturing intricate linguistic nuances for accurate news categorization.

The testing accuracies for BERT and RoBERTa were 88.46% and 91.89%, respectively, underscoring their effectiveness in discerning between different news categories. These results highlight the transformative impact of transformer-based models in natural language processing tasks, outperforming traditional approaches.

C. Training Dynamics

Figures 6 depict the training dynamics of RoBERTa, illustrating the convergence patterns during training. The training accuracy vs. validation accuracy graph provides insights into the model's learning process, indicating stability and performance on both training and validation sets. Similarly, the training loss vs. validation loss graph illuminates the optimization trajectory, shedding light on how well the model generalizes to new data.

These visualizations contribute to a comprehensive understanding of the training dynamics of RoBERTa [3], paving the way for nuanced discussions on model convergence, overfitting, and generalization.

In the subsequent section, a comparative analysis is presented, contrasting the results obtained from the traditional approach with those achieved using transformer-based models.

V. Discussion

A. Model Comparison

The comparison between the traditional Support Vector Classifier (SVC) and deep learning models, BERT and RoBERTa, reveals distinct strengths and weaknesses inherent in each approach. Table I provides a comprehensive overview of performance metrics, facilitating a nuanced discussion.

TABLE I
Model Performance Comparison

Model	Accuracy (%)	Precision (%)	Recall (%)
Logistic Regression	83.65	84.00	85.00
SVC	86.38	88.00	86.00
BERT	88.46	87.00	89.00
RoBERTa	91.89	92.00	92.00

The traditional SVC model [4], despite its commendable performance, encounters challenges when confronted with the complexity and variability inherent in news articles. Relying on handcrafted features and linear decision boundaries, the SVC model may struggle to adapt to the intricate nuances of language, potentially limiting its effectiveness in capturing contextual dependencies.

Similarly, Logistic Regression, although robust in its simplicity, faces analogous limitations. The model's reliance on linear decision boundaries and manual feature engineering might hinder its ability to fully grasp the intricate patterns present in news articles, especially when compared to the more advanced capabilities of transformer-based models.

In contrast, BERT [2] and RoBERTa [3], leveraging transformer architectures, showcase superior accuracy, precision, and recall. Their ability to capture contextual information and semantic relationships positions them as formidable contenders for news classification tasks. However, it's crucial to consider the computational resources and training time required for these deep learning models.

The choice between traditional and deep learning models depends on the specific requirements of the news classification task, considering factors such as dataset size, computational capabilities, and desired performance metrics.

The subsequent section delves into the broader implications of the study and potential avenues for future research.

B. Challenges Faced

The successful implementation of the models was not without its share of challenges. Several factors influenced the training process and outcomes:

- **Limited GPU Access:** The availability of GPU resources significantly impacted the training time and scalability of the deep learning models. Limited access to powerful GPUs might lead to longer training durations, affecting the timely exploration of hyperparameters and model architectures.
- **Hardware Constraints:** The hardware infrastructure, including CPU and RAM limitations, posed challenges during data preprocessing and model training. Optimal model performance often requires substantial computational resources, and constraints in hardware might compromise the training efficiency.
- **Hyperparameter Tuning:** The selection of optimal hyperparameters is a critical aspect of deep learning model training. Limited computational resources may restrict the exhaustive exploration of hyperparameter space, potentially leading to suboptimal configurations.
- **Overfitting and Generalization:** Achieving a balance between overfitting and generalization is a delicate task. The intricate nature of news articles and the diverse vocabulary present challenges in creating models that generalize well to unseen data without overfitting to the training set.

These challenges underscore the importance of robust infrastructure, comprehensive hyperparameter tuning, and addressing data-related issues for successful model implementation and deployment.

VI. Conclusion

In conclusion, this study conducted a comprehensive comparative study of traditional and transformer-based models for news classification, shedding light on their respective strengths, limitations, and overall effectiveness. The key findings and contributions of this study can be summarized as follows:

VII. Key Findings

The study yielded several key findings that illuminate the efficacy of different models in the context of news classification:

- **Performance Disparity:** Traditional models, specifically Support Vector Classifier (SVC) and Logistic Regression, exhibited competitive accuracy, precision, recall, and F1-score metrics. However, their performance was surpassed by transformer-based models—BERT and RoBERTa.
- **Transformer Superiority:** BERT and RoBERTa demonstrated superior performance, achieving higher accuracy rates and more robust handling of the nuanced language present in news articles. The transformer models' ability to capture intricate contextual dependencies contributed to their enhanced classification capabilities.
- **Influence of Pre-training:** The pre-training of BERT and RoBERTa on extensive textual data showcased its significance. The models, equipped with pre-existing language understanding, outperformed traditional models in classifying news articles.
- **Complexity and Scalability:** Traditional models, such as SVC and Logistic Regression, exhibited simplicity and interpretability. However, they faced challenges in handling the complexity of language and scalability when dealing with large datasets.
- **Limitations of Handcrafted Features:** SVC and Logistic Regression's reliance on handcrafted features and linear decision boundaries posed limitations in capturing intricate language patterns present in news articles.
- **Robustness of Transformers:** Transformer-based models demonstrated robustness in handling diverse and complex language patterns, making them more suitable for news classification tasks.

These findings contribute to a deeper understanding of the strengths and limitations of both traditional and transformer-based models, providing insights for future advancements in news classification methodologies.

A. Contributions

- **Model Comparison:** The comparative analysis presented a nuanced understanding of the strengths and weaknesses of traditional and transformer-based models. This information is valuable for practitioners and researchers seeking to make informed decisions when selecting models for news classification tasks.
- **Insights for Future Work:** The challenges faced during the implementation of models highlight avenues for future research. Addressing issues such as limited GPU access, and hardware constraints could contribute to more robust and scalable news classification models.

B. Applications

The insights garnered from this study have practical implications for the field of natural language processing and news analysis. The findings can inform the development of effective models for real-world applications, such as automated news categorization and content recommendation systems.

C. Future Directions

Future research endeavors could explore advanced techniques for handling data imbalances, investigate ensemble methods to leverage the strengths of different models, and delve into transfer learning approaches to enhance the generalization capabilities of news classification models.

In conclusion, this study advances our understanding of the landscape of news classification models, offering valuable insights and paving the way for further advancements in the intersection of machine learning and natural language processing.

VIII. Future Work

The findings of this research open up several avenues for future exploration in the domain of news classification. The following suggestions propose potential directions for future research:

- **Fine-Tuning Strategies:** Investigate and experiment with fine-tuning strategies for transformer-based models. Exploring various hyperparameter configurations, optimization techniques, and domain-specific pre-training could contribute to further improving model performance.
- **Ensemble Models:** Explore the efficacy of ensemble models by combining the strengths of different architectures. Building ensemble models that integrate traditional machine learning approaches with transformer-based models or incorporating models pre-trained on domain-specific data could potentially enhance classification accuracy.
- **Bias Analysis:** Conduct a comprehensive analysis of bias in news classification models. Investigate potential biases that may emerge due to imbalances in the training data, evaluate the impact of diverse sources on bias, and develop strategies to mitigate bias in classification outcomes.
- **Continuous Monitoring:** Implement continuous monitoring mechanisms for news classification models. Regularly assess model performance over time, adapt to evolving language patterns, and integrate feedback loops to ensure the ongoing effectiveness and relevance of the deployed models.
- **Exploration of New Architectures:** Stay abreast of advancements in natural language processing and explore the integration of new model architectures. Experiment with cutting-edge models that might offer improved efficiency, scalability, and performance for news classification tasks.

These future research directions aim to build upon the insights gained from this study, contributing to the ongoing refinement and innovation in news classification methodologies. By addressing these areas, researchers can advance the field and develop models that are more robust, interpretable, and adaptable to the dynamic nature of news content.

References

[1] *Attention is All You Need*. A. Vaswani, N. Shazeer, N. Parmar et al., 2017. [Online]. Available: https://arxiv.org/abs/1706.03762

[2] *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. J. Devlin, M. W. Chang, K. Lee, et al., 2018. [Online]. Available: https://arxiv.org/abs/1810.04805

[3] *RoBERTa: A Robustly Optimized BERT Approach*. Y. Liu, M. Ott, N. Goyal et al., 2019. [Online]. Available: https://arxiv.org/abs/1907.11692

[4] *Support-Vector Networks*. C. Cortes, V. Vapnik, 1995. [Online]. Available: https://ieeexplore.ieee.org/document/4676995

[5] *PyTorch: An Imperative Style, High-Performance Deep Learning Library*. A. Paszke, S. Gross, F. Massa et al., 2019. [Online]. Available: https://arxiv.org/abs/1912.01703

[6] *Gradient-Based Transfer Learning*. Gustaf Tegnér, Alfredo Reichlin, Hang Yin, Hedvig Kjellstrom, Mårten Björkman, Danica Kragic. 2023. [Online]. Available: https://openreview.net/forum?id=hChYEyebNm1

[7] *XLNet: Generalized Autoregressive Pretraining for Language Understanding*. Z. Yang, Z. Dai, Y. Yang et al., 2019. [Online]. Available: https://arxiv.org/abs/1906.08237

[8] *An Empirical Exploration of Recurrent Network Architectures*. R. Jozefowicz, W. Zaremba, I. Sutskever, 2015. [Online]. Available: https://arxiv.org/abs/1507.03958

[9] *AG News Dataset*. K. Zhang et al., 2015. [Online]. Available: http://www.di.unipi.it/ gulli/AG_corpus_of_news_articles.html

[10] M. Zaharia, M. Chowdhury, T. Das, A. Dave, J. Ma, M. McCauley, M. J. Franklin, S. Shenker, I. Stoica, *Apache Spark: A Unified Engine for Big Data Processing*, 2016. [Online]. Available: https://www.usenix.org/system/files/conference/nsdi12/nsdi12-final138.pdf

[11] *Beautiful Soup Documentation*. Leonard Richardson. [Online]. Available: https://www.crummy.com/software/BeautifulSoup/bs4/doc/

[12] *Hugging Face Transformers Documentation*. Hugging Face. [Online]. Available: https://huggingface.co/transformers/

[13] *scikit-learn: Machine Learning in Python*. Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, et al. [Online]. Available: https://scikit-learn.org/stable/

[14] *Term Frequency-Inverse Document Frequency (TF-IDF)*. G. Salton, A. Wong, and C. S. Yang, 1975. [Online]. Available: https://www.researchgate.net/publication/220734966