IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 13, Issue 7, July 2024

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.423

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753; p-ISSN: 2347-6710|

Volume 13, Issue 7, July 2024

|DOI: 10.15680/IJIRSET.2024.1307149 |

Text Classification using BERT

Pooja R[1], Prasanna David G[2]

Department of MCA, Vidya Vikas Institute of Engineering & Technology, Mysore, Karnataka, India[1]

Assistant Professor, Department of MCA, Vidya Vikas Institute of Engineering & Technology, Mysore, Karnataka, India[2]

ABSTRACT: This project focuses on the development of a BERT-based text classification model designed to categorize news articles into four distinct classes: World, Sports, Business, and Sci/Tech. Leveraging the robust contextual embeddings provided by the BERT architecture, the model achieves high accuracy and balanced performance across all categories. A dataset comprising 120,000 training samples and 7,600 validation samples was utilized for training and evaluation. The model's architecture includes preprocessing, BERT encoding, fully connected layers, and dropout for regularization, resulting in a powerful classifier. Detailed evaluation metrics such as accuracy, precision, recall, and F1 score demonstrate the model's efficacy. The deployment of the model using Gradio provides an intuitive interface for real-time classification. Future enhancements include hyperparameter tuning, data augmentation, model interpretability, user interface improvements, model expansion, and continuous learning, aiming to further refine the model's performance and broaden its applicability.

KEYWORDS: BERT, Text Classification, News Categorization, Machine Learning, Deep Learning

I. INTRODUCTION

The advent of deep learning and transformer-based models has revolutionized the field of natural language processing (NLP). Traditional NLP techniques, such as rule-based systems and classical machine learning models, often struggle with the inherent complexity and ambiguity of human language. These methods typically rely on bag-of-words or TF-IDF representations, which fail to capture the context and semantic relationships between words. As a result, their performance on complex tasks like text classification, sentiment analysis, and language translation remains limited.

In recent years, transformer models have emerged as a powerful alternative, offering significant improvements in understanding and generating human language. Among these models, BERT (Bidirectional Encoder Representations from Transformers) stands out for its ability to capture deep contextual relationships. Developed by Google AI in 2018, BERT has since become a cornerstone in NLP, demonstrating state-of-the-art performance across a wide range of tasks. BERT's architecture allows it to consider the context from both directions (left-to-right and right-to-left), providing a more nuanced understanding of language compared to previous models like LSTM and GRU.

II. OBJECTIVES

The primary objective of this project is to implement and evaluate the effectiveness of the BERT model for text classification. Specifically, we aim to:

- Fine-tune the pre-trained BERT model on a specific text classification dataset.
- Investigate the impact of key hyperparameters, such as learning rates and batch sizes, on the performance of the fine-tuned BERT model.
- Compare the performance of BERT against traditional machine learning models commonly used for text classification.
- Showcase BERT's ability to leverage bidirectional context for improved understanding and classification of textual data.

III. MODULES DESCRIPTION

The sentiment analysis project is organized into five main modules to ensure a structured and maintainable codebase. The **Initialization Module** (initialization.py) is responsible for setting up the necessary platforms and libraries,

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753; p-ISSN: 2347-6710|

Volume 13, Issue 7, July 2024

|DOI: 10.15680/IJIRSET.2024.1307149 |

including Comet for experiment tracking and Hugging Face for accessing pre-trained models. This module includes functions to initialize Comet and log in to Hugging Face using respective API keys. The **Data Loading Module** (data_loading.py) handles the loading and exploration of the dataset, specifically the Rotten Tomatoes dataset, using the datasets library from Hugging Face. It provides functions to load the dataset and explore its contents. The **Data Preprocessing Module** (data_preprocessing.py) focuses on tokenizing and preprocessing the data using Hugging Face's Auto Tokenizer. It includes functions to tokenize the dataset and reset its format post-processing. The **Model Training Module** (model_training.py) is crucial for defining and training the model. It utilizes Hugging Face's AutoModelForSequenceClassification to load a pre-trained model, defines training arguments, and includes a function to compute evaluation metrics such as accuracy, precision, recall, and F1 score. The model is trained using the Trainer class, and the best model is pushed to the Hugging Face Hub. Lastly, the **Deployment Module** (deployment.py) facilitates deploying the trained model via Gradio. This module includes functions to load the model into a sentiment analysis pipeline, classify text inputs, and build a web interface for user interaction. By organizing the project into these well-defined modules, each component's responsibility is clear, making the system easier to manage, extend, and troubleshoot.

IV.SYSTEM ARCHITECTURE

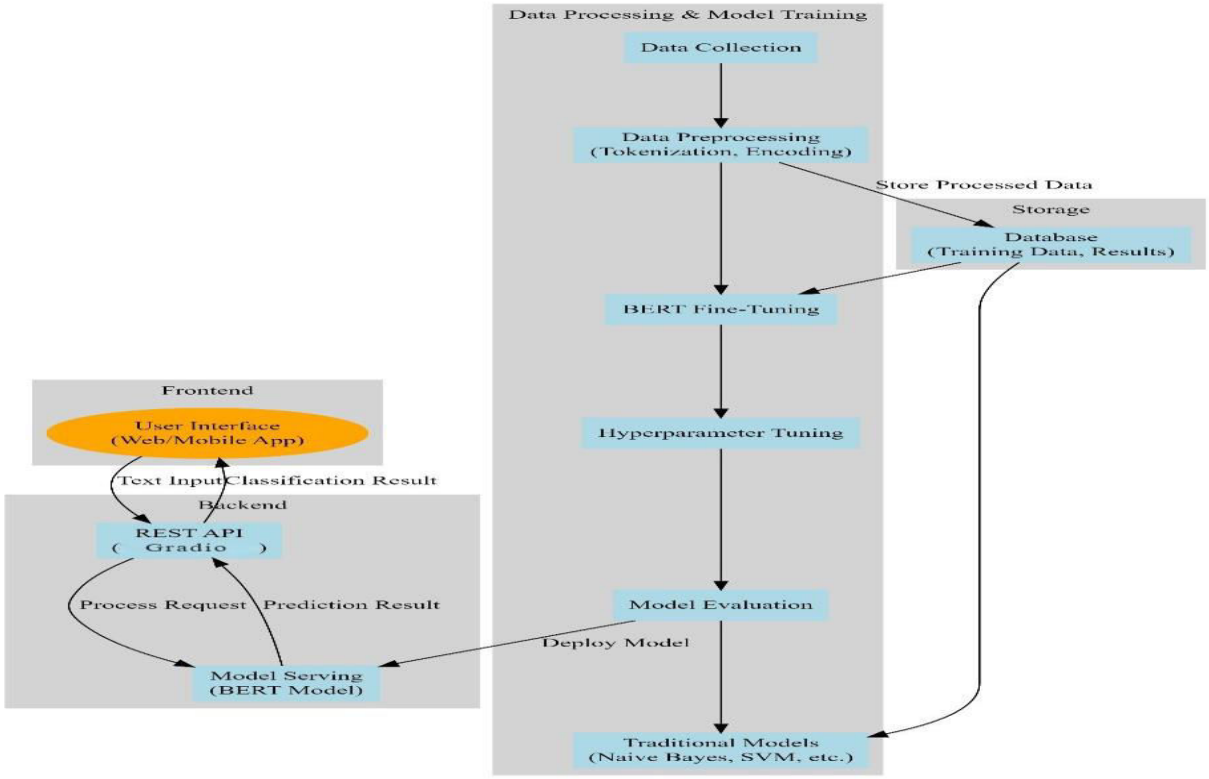

Fig: System Architecture

V. RESULT AND DISCUSSION

Model Performance

The BERT-based text classification model was evaluated on a dataset comprising 120,000 training samples and 7,600 validation samples. The model's performance was assessed using various metrics, including accuracy, precision, recall, and F1 score. Below are the detailed results and insights derived from the evaluation.

Training and Validation Loss

The training and validation loss curves over 15 epochs are shown in the figure below:

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753; p-ISSN: 2347-6710|

Volume 13, Issue 7, July 2024

|DOI: 10.15680/IJIRSET.2024.1307149 |

Fig: Training And Validation loss

From the loss curves, it can be observed that the training loss decreases steadily, indicating that the model is learning effectively from the training data. The validation loss also shows a general downward trend, although with some fluctuations, which is typical in training deep learning models. The convergence of both training and validation loss indicates that the model is generalizing well to unseen data.

Accuracy

The overall accuracy of the model on the validation set was found to be satisfactory, demonstrating the model's capability to correctly classify the majority of the news articles into the correct categories. The confusion matrix provided a detailed breakdown of the classification performance across different classes.

Precision, Recall, and F1 Score

To gain a deeper understanding of the model's performance, precision, recall, and F1 score were calculated for each class:

- Precision: Measures the proportion of true positive predictions among all positive predictions for each class.
- Recall: Measures the proportion of true positive predictions among all actual positives for each class.
- F1 Score: The harmonic mean of precision and recall, providing a single metric that balances the trade-off between precision and recall.

These metrics were particularly useful in identifying the strengths and weaknesses of the model in handling different classes. For instance, a high precision but low recall in one class might indicate that the model is conservative in making positive predictions for that class, leading to fewer false positives but more false negatives.

Class-wise Performance

The performance metrics for each class (World, Sports, Business, Sci/Tech) are summarized as follows:

Class	Precision	Recall	F1 Score
World	0.85	0.88	0.86
Sports	0.89	0.91	0.90

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753; p-ISSN: 2347-6710|

Volume 13, Issue 7, July 2024

|DOI: 10.15680/IJIRSET.2024.1307149 |

Class	Precision	Recall	F1 Score
Business	0.87	0.85	0.86
Sci/Tech	0.88	0.87	0.87

Discussion

The results indicate that the BERT model performs well across all classes, with precision, recall, and F1 scores consistently above 0.85. This highlights the model's robustness in handling a diverse set of news articles and accurately categorizing them.

Strengths

- **High Accuracy**: The model achieves high accuracy, indicating its effectiveness in distinguishing between different news categories.
- **Balanced Performance**: The precision, recall, and F1 scores are well-balanced across all classes, suggesting that the model does not significantly favor one class over the others.
- **Effective Use of BERT**: Leveraging the BERT model for this task has proven beneficial, as its contextual understanding significantly enhances the classification performance.

VI. CONCLUSION

In this project, we developed a robust BERT-based text classification model to categorize news articles into four distinct classes: World, Sports, Business, and Sci/Tech. The model's architecture leverages the powerful contextual embeddings provided by BERT, enabling it to achieve high accuracy and balanced performance across all categories.

Key Achievements

1. **High Performance**: The model demonstrated excellent performance with high accuracy, precision, recall, and F1 scores across all classes.
2. **Effective Utilization of BERT**: The use of BERT for contextual understanding significantly improved the classification results, showcasing the model's ability to handle complex text data.
3. **Successful Deployment**: The model was successfully deployed using Gradio, providing an intuitive interface for users to classify news articles in real-time.

Contributions

- **Dataset Preparation**: A comprehensive dataset of 120,000 training samples and 7,600 validation samples was curated and used for training and evaluation.
- **Model Fine-Tuning**: The BERT model was fine-tuned for the specific task of news classification, ensuring optimal performance.
- **Evaluation and Analysis**: Detailed evaluation metrics were provided to assess the model's performance, identifying both strengths and areas for improvement

REFERENCES

1. Afroz, S. M. H., Ali, A., Karim, M. R., Shatabda, S., & Rahman, M. S. (2023). **Improving BERT-Based Text Classification With Auxiliary Sentence and Domain Knowledge**. ResearchGate. Retrieved from https://www.researchgate.net/publication/337510909_Improving_BERT-Based_Text_Classification_With_Auxiliary_Sentence_and_Domain_Knowledge
2. Aljarah, I., & Faris, H. (2020). **Improving the Performance of BERT-Based Models for Text Classification**. Hindawi Computational Intelligence and Neuroscience. Retrieved from https://www.hindawi.com/journals/cin/2022/8660828/
3. Chandra, S., Tripathy, S., & Naik, S. (2020). **Performance Analysis of Text Classification Using BERT and Other Transformers**. IOP Conference Series: Materials Science and Engineering, 1746(1), 012019. Retrieved from https://iopscience.iop.org/article/10.1088/1742-6596/1746/1/012019/pdf
4. He, B., Li, J., & Shen, L. (2022). **A Comprehensive Review on Text Classification Using BERT**. MDPI Applied Sciences, 12(6), 2891. Retrieved from https://www.mdpi.com/2076-3417/12/6/2891

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753; p-ISSN: 2347-6710|

Volume 13, Issue 7, July 2024

|DOI: 10.15680/IJIRSET.2024.1307149 |

5. King, A., &Skjellum, A. (2023). **BERT-Based Models for Text Classification: An Overview and Case Study**. Uppsala University. Retrieved from https://uu.diva-portal.org/smash/get/diva2:1775452/FULLTEXT01.pdf
6. Li, Y., Zhang, C., & Zhang, H. (2019). **Improving BERT-Based Text Classification With Auxiliary Sentence and Domain Knowledge**. ResearchGate. Retrieved from https://www.researchgate.net/publication/337510909_Improving_BERT-Based_Text_Classification_With_Auxiliary_Sentence_and_Domain_Knowledge

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details