Opinion Analysis of Media Bias

CS5344 Big Data Analytics - Final Report

Wu Tong
School of Computing
National University of Singapore
A0255954R

Yang Haohong
School of Computing
National University of Singapore
A0286177N

***Abstract*—This project presents a comprehensive framework for sentiment analysis applied to media articles related to Singapore, aiming to uncover potential biases and consistent trends in international reporting. Leveraging the Global Geographic Graph Singapore (GGGSG) dataset, which includes over 9 million English-language news articles from April 2017 to July 2024, the study fine-tuned a RoBERTa model for feature extraction and employed a Random Forest classifier for sentiment classification. The proposed approach achieved a significant accuracy improvement of 77.9%, surpassing the 64% baseline obtained using traditional TF-IDF and logistic regression models. The model's performance in cross-domain validation, however, highlighted the challenges of generalizing across datasets, as accuracy dropped to 63.18% on an IMDB Review Dataset due to differing linguistic structures and expressions.**

Key strategies included layer freezing during training, mixed precision computation for efficient memory usage, and custom data generators for handling large-scale data. The findings underscored the complexities of sentiment detection, particularly for negative and indirect expressions, and demonstrated the need for further refinement in domain adaptability and nuanced language interpretation. The research contributes to advancing NLP methodologies for media bias analysis and sets the stage for future work on cross-domain generalization and real-time sentiment analysis tools.

I. Introduction

A. Project Background

Sentiment analysis has become an essential tool for understanding public opinion and biases expressed in the media. In the context of geopolitics and international relations, analyzing media sentiment towards specific countries or regions can reveal trends, biases, and the overall narrative being portrayed. This project aims to conduct a comprehensive sentiment analysis on media articles mentioning Singapore, focusing on detecting potential biases and consistent opinions across different media outlets. By leveraging advanced natural language processing (NLP) models, we seek to capture sentiment patterns that may influence public perception and policy decisions.

Sentiment analysis is increasingly used in media studies and political analysis to dissect narratives and uncover hidden biases. Studies have shown that media bias can shape public perception and influence policy decisionsent of NLP technologies, especially transformer-based models such as BERT and RoBERTa, has enhanced the ability to conduct nuanced sentiment analysis, allowing researchers to extract meaningful insights from large-scale text data.

B. Motivation

Understanding media sentiment is crucial for governments, policymakers, and international relations experts who monitor the portrayal of their country in global media. Singapore, known for its strategic economic and political position in Southeast Asia, often finds itself at the center of various geopolitical narratives. Identifying how different countries' media outlets report on Singapore helps uncover potential biases that may influence international opinions and relationships. This project is motivated by the need to:

Assess media biases: Determine whether certain media sources report more positively or negatively about Singapore and identify trends over time. Media biases can affect public opinion, with implications for international relations and policy .

1) **Inform Policy and Communication Strategies:** Equip policymakers and public relations specialists with data-driven insights to craft strategies that address negative portrayals or reinforce positive ones.
2) **Advance NLP Methodologies:** Show the effectiveness of using fine-tuned transformer models like RoBERTa for sentiment analysis, combined with traditional machine learning approaches such as Random Forest for robust prediction and interpretation.
3) **Combined State-of-the-Art Techniques:** By combining state-of-the-art NLP techniques with data processing and analysis, this project seeks to not only uncover sentiment trends but also provide a blueprint for similar studies in other geopolitical contexts. The outcome aims to demonstrate the importance of computational tools in understanding the complex landscape of media narratives.

C. Project Target

The primary target of this project is to develop a sentiment analysis framework capable of accurately detecting and classifying opinions expressed in media articles about Singapore. The framework aims to:

1) **Build and optimize a sentiment analysis model** that incorporates transformer-based NLP models for feature extraction, enhanced by machine learning models for classification.
2) **Identify and interpret sentiment trends** in media coverage over a period spanning several years, focusing on shifts in sentiment that correlate with significant events such as the COVID-19 pandemic .
3) **Compare sentiment across different media sources,** whether they are region-specific or globally prevalent.
4) **Provide actionable insights** for policymakers and analysts to understand the portrayal of Singapore in international media and address any unfavorable biases effectively.

In addition to analyzing media sentiment and identifying biases, an essential goal of this project was to outperform the baseline model's performance metrics. The baseline used a combination of TF-IDF and logistic regression, achieving the following classification performance:

- Precision: 0.63 (macro and weighted average)
- Recall: 0.63 (macro and weighted average)
- F1-score: 0.58 (macro and weighted average)
- Accuracy: 63%

```
print(classification_report(y_test, y_pred))

              precision    recall  f1-score   support

           0       0.64      0.83      0.72    621816
           1       0.63      0.19      0.29    611344
           2       0.62      0.87      0.72    603846

    accuracy                           0.63   1837006
   macro avg       0.63      0.63      0.58   1837006
weighted avg       0.63      0.63      0.58   1837006
```

Fig. 1. TF-IDF LR Baseline

This project seeks to set a benchmark for sentiment analysis projects that deal with media bias and geopolitical narratives, demonstrating that sophisticated NLP models can be applied effectively to large, diverse datasets.

II. Dataset

A. Overview of the Dataset

The dataset used in this project is the Global Geographic Graph Singapore (GGGSG) dataset, sourced from the GDELT Project, which is known for its comprehensive and global coverage of news articles. This dataset encompasses English-language news articles that mention Singapore, spanning from April 4, 2017, to July 19, 2024. In total, it includes 9,185,305 rows, with each row representing a mention of a Singapore-related entity in a news article. The dataset provides a rich source for sentiment analysis, enabling the examination of international perspectives and potential biases in media coverage.

Key columns in the dataset include:

- **DateTime:** The date and time when the article was published.
- **URL:** The link to the full article.
- **Title:** The title of the article.
- **DocTone:** A machine-generated sentiment score, represented as a float value.
- **DomainCountryCode:** The country code of the article's origin.
- **ContextualText:** A 600-character snippet surrounding the mention of Singapore.
- **CountryCode:** Always "SN" for this dataset, indicating references to Singapore.

This dataset is valuable for sentiment analysis due to its size and diversity, capturing global news from a range of countries and perspectives. However, it comes with challenges such as potential noise in machine-annotated sentiment labels and varied writing styles across sources.

B. Data Preprocessing and Cleaning

Effective data preprocessing is crucial for ensuring the quality and reliability of sentiment analysis models. The preprocessing steps implemented in this project included:

1) **Data Loading:** Given the substantial size of the dataset (over 9 million rows and 8GB in storage), efficient data loading techniques such as stream processing were employed to handle memory constraints. This involved reading the data in chunks and using on-demand data generators to reduce memory load.
2) **Data Cleaning:**
   - **Removal of Irrelevant Columns:** Columns not essential for the analysis, such as image URLs and geographic coordinates, were dropped.
   - **Handling Missing Values:** Rows with critical missing data, such as empty sentiment scores and news content, were filtered out.
   - **Fix Datatype:** For DocTone column, fix datatype as float to avoid unexpected values.
   - **Text Normalization:** The 'ContextualText' was converted to lowercase, and punctuation was removed to standardize the input.
   - **Duplicate Removal:** Duplicate rows, especially those corresponding to the same article, were removed to avoid skewed results.
   - **Quantile-Based Sentiment Labeling:** To label the sentiment of articles, quantiles were computed from the DocTone column to divide scores into segments, facilitating a more granular classification. The quantiles were computed at the 20th, 40th, 60th, and 80th percentiles. These thresholds (Q1, Q2, Q3, Q4) were used to segment the DocTone values into five categories: 'Strongly Negative', 'Negative', 'Neutral', 'Positive', and 'Strongly Positive'.

After that, the sentiment labels were encoded numerically for use in machine learning models using LabelEncoder.

C. Data Challenges and Considerations

1) **Large Dataset Size:** The dataset's size posed challenges for direct in-memory loading. To address this, batch processing and streaming methods were implemented to ensure efficient handling without overloading system memory.
2) **Class Imbalance:** Initial analysis revealed that the distribution of sentiment classes was imbalanced, with neutral sentiment being the most common. Class weights were adjusted during model training to handle this imbalance.
3) **Feature Extraction Efficiency:** To enhance feature extraction efficiency, custom Python generators were used to fetch data in manageable batches, ensuring optimal use of memory and processing power.

D. Visual Analysis

To better understand the dataset's distribution and sentiment trends, visual analyses were conducted:

- **Country-Based Sentiment Analysis:** A bar chart visualizing the average sentiment scores from top reporting countries.

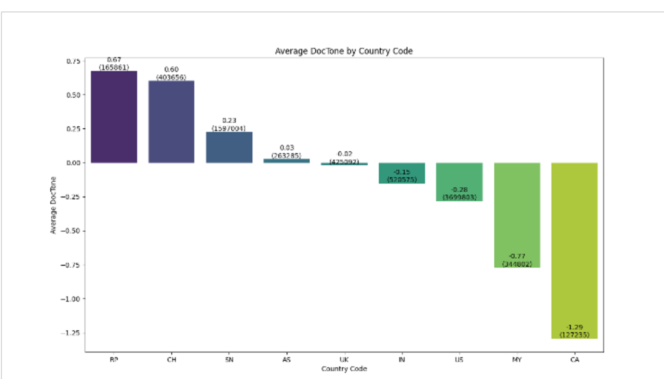


Fig. 2. Average sentiment scores from countries

- **Trend Analysis:** A line graph depicting sentiment changes over significant time periods, with marked events like the COVID-19 pandemic.

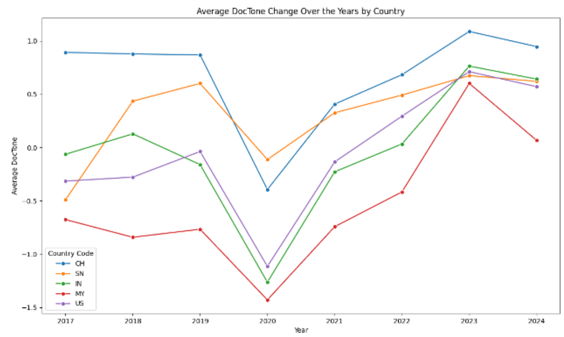


Fig. 3. Sentiment changes over time periods by countries

E. Other Datasets

In addition to the primary Global Geographic Graph Singapore (GGGSG) dataset, which provided a comprehensive source of English news articles related to Singapore, a supplementary dataset was included for evaluating the model's adaptability: the IMDB Review Dataset.

The IMDB Review Dataset is a well-known benchmark used in sentiment analysis, containing movie reviews annotated for sentiment classification. This dataset is characterized by:

- Linguistic Variety: The reviews in the dataset showcase a diverse range of language, including casual, formal, and nuanced expressions, making it a valuable resource for assessing the robustness of sentiment analysis models.
- Complex Emotional Content: The dataset includes reviews that may feature sarcasm, puns, and ambiguity, presenting challenges for models not specifically trained on such expressions.

The purpose of including the IMDB dataset:

- Cross-Domain Evaluation: By incorporating the IMDB Review Dataset, the project aimed to explore the model's generalization capabilities. This allowed for the assessment of how well the model, trained on news content, could adapt to a different domain with distinct language patterns and emotional expressions.
- Understanding Model Limitations: The use of this dataset provided insights into the limitations of the model in recognizing complex, indirect language and differing emotional tones found in movie reviews.

The IMDB dataset's unique qualities make it an excellent testbed for evaluating the flexibility and robustness of sentiment analysis models across domains. The comparison between results from this dataset and the GGGSG dataset helped highlight the necessity for domain-specific adjustments and adaptations in model training.

III. Approaches

A. Overall Strategy

The approach for this project integrates advanced deep learning and traditional machine learning techniques to achieve robust sentiment analysis. The main steps include:

1) **Feature Extraction using Pre-trained Transformer Models:** Utilize RoBERTa to extract rich, context-aware text embeddings.
2) **Model Training and Fusion:** Use these embeddings to train a Random Forest classifier, combining the strengths of both deep learning and ensemble learning for effective sentiment classification.
3) **Hyperparameter Tuning:** Optimize the training process using Optuna for better performance.

This hybrid strategy leverages RoBERTa's superior language understanding for feature extraction and Random Forest's ability to handle non-linear relationships and feature interactions.

B. Data Tokenization

Before feature extraction, the dataset was tokenized to convert raw text into a format compatible with transformer models:

- **Tokenizer:** The tokenizer was used to preprocess the ContextualText column, converting text into input_ids and attention_mask for model input.
- **Benefits:** Tokenization preserves the structure and semantics of the text, enabling the model to better capture contextual meaning.

```
tokens = tokenizer(
    df_filtered['ContextualText'].tolist(),
    max_length=256,
    padding=True,
    truncation=True,
    return_tensors='pt'
)
```

C. Introduction to RoBERTa and Motivation

RoBERTa (Robustly Optimized BERT Approach) is an improvement over BERT that optimizes training strategies, including larger batch sizes, removing next-sentence prediction, and using more extensive training data [2]. This model is well-suited for tasks requiring a deep understanding of context, making it ideal for sentiment analysis where the tone and sentiment nuances are important.

RoBERTa model has been chosen for several reasons:

- **Strong Contextual Understanding:** RoBERTa is known for its robust representation of complex language structures, providing embeddings that capture deep contextual relationships.
- **Proven Performance:** Research shows that RoBERTa outperforms BERT and other baseline models on multiple NLP benchmarks [2].
- **Suitability for Fine-Tuning:** The model's architecture makes it adaptable for domain-specific tasks, enhancing accuracy when fine-tuned on specialized datasets.

D. Fine-Tuning RoBERTa for Feature Extraction

1) Motivation: The pre-trained RoBERTa (Robustly optimized BERT approach) model has demonstrated significant success in understanding complex language structures across various NLP tasks due to its deep, bidirectional, and unsupervised pre-training strategy [2]. However, applying RoBERTa directly without domain-specific adjustments can limit its effectiveness, particularly in specialized areas such as sentiment analysis of geopolitical news. Fine-tuning RoBERTa on a custom dataset, like news articles mentioning Singapore, refines its parameters to better capture the unique language, phrasing, and sentiment variations in this domain. This step is essential for adapting the model to specific data characteristics that pre-trained general models might not adequately capture.

2) Rationale and Supporting Evidence: Research has consistently shown that fine-tuning pre-trained transformer models yields significant performance gains across many NLP tasks. Howard and Ruder (2018) highlighted that the Universal Language Model Fine-Tuning (ULMFiT) approach, which involves pre-training followed by domain-specific fine-tuning, significantly improves task-specific results by adapting the model to particular datasets [1]. RoBERTa extends this paradigm by refining BERT's training regime, showing robust results when fine-tuned on domain-specific data, such as specialized news articles [2].

3) Benefits:

- **Customized Language Representation:** By fine-tuning RoBERTa on domain-specific content, the model can better encode the semantics of specialized language, leading to more precise and context-aware embeddings.
- **Increased Model Performance:** Numerous studies have demonstrated that fine-tuned models surpass their pre-trained-only counterparts in performance metrics across tasks, including sentiment analysis [1] [2].

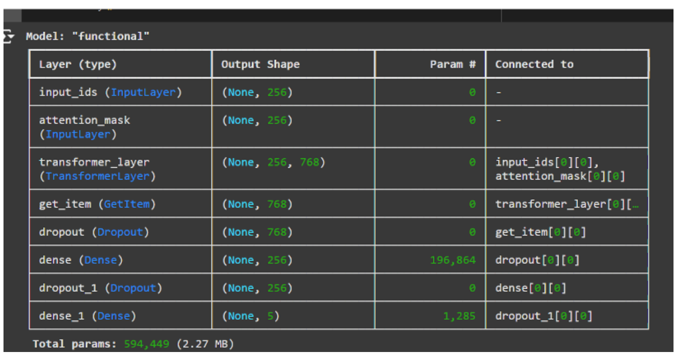
4) Training and Strategy:: To maximize the model's effectiveness, the fine-tuning phase was carefully configured, involving multiple considerations for hyperparameter tuning and resource management:

- **Hyperparameter Tuning with Optuna:** Optuna is a powerful framework for automatic hyperparameter optimization. In this project, it was used to search for the optimal combination of key hyperparameters such as:
  - **Learning Rate:** Fine-tuned within a range of 1×10^{-6} to 5×10^{-6} using logarithmic sampling for a more granular search.
  - **Batch Size:** Experimented with 8, 16, and 32 to balance computational load and training stability.
  - **Epochs:** Varied between 2 and 5 to identify the point at which the model achieves the highest validation performance without overfitting.

```
def objective(trial):
    learning_rate = trial.suggest_float('
        learning_rate', 1e-6, 5e-5, log=True)
    batch_size = trial.suggest_categorical('
        batch_size', [8, 16, 32])
    num_epochs = trial.suggest_int('num_epochs',
         2, 5)
    # Training logic...
    return validation_score
```

- **Training Strategy:**

- **Layer Freezing:** Lower layers of the RoBERTa model were frozen to maintain general language representations while focusing training on the top layers. This approach reduced training time and computational requirements without sacrificing model performance.
- **AdamW Optimizer:** Used for its effective handling of weight decay during gradient updates, enhancing training efficiency and convergence.
- **Early Stopping and Learning Rate Scheduling:** Early stopping monitored the validation loss and stopped training when the model's performance plateaued. Learning rate scheduling reduced the learning rate when a plateau was detected to fine-tune learning.
- **Data Prefetching and Shuffling:** 'tf.data.Dataset' was configured with data prefetching and shuffling to optimize the training efficiency by overlapping data preprocessing and model execution reducing bottlenecks in data loading.



Model: "functional"

Layer (type)	Output Shape	Param #	Connected to
input_ids (InputLayer)	(None, 256)	0	-
attention_mask (InputLayer)	(None, 256)	0	-
transformer_layer (TransformerLayer)	(None, 256, 768)	0	input_ids[0][0], attention_mask[0][0]
get_item (GetItem)	(None, 768)	0	transformer_layer[0][…
dropout (Dropout)	(None, 768)	0	get_item[0][0]
dense (Dense)	(None, 256)	196,864	dropout[0][0]
dropout_1 (Dropout)	(None, 256)	0	dense[0][0]
dense_1 (Dense)	(None, 5)	1,285	dropout_1[0][0]

Total params: 594,449 (2.27 MB)

Fig. 4. Model Detail for Fine-tune RoBERTa

5) Good Practices and Observations: During fine-tuning, it was noted that:

- **Gradual Unfreezing:** Unfreezing layers progressively helped maintain lower-level linguistic features while allowing higher-level layers to adapt to the specific sentiment nuances.
- **Checkpointing and Early Stopping:** These strategies prevented overfitting by monitoring the validation loss and halting training when performance plateaued.

Overall, fine-tuning RoBERTa with domain-specific data and optimizing hyperparameters led to the extraction of embeddings that were contextually enriched and well-suited for subsequent classification tasks. This approach bridges the gap between general-purpose language models and tailored, task-specific applications, affirming its utility in academic and industry settings for sentiment analysis and related NLP challenges.

E. Model Fusion with Random Forest

1) Model Chosen: Here are some reasons that Random Forest model has been chosen for fusion:

- **Robustness:** Random Forest is a powerful ensemble learning method that combines multiple decision trees, offering robust performance and better generalization compared to single models.
- **Non-Linear Capability:** Unlike linear models, Random Forest can handle non-linear interactions between features, making it well-suited for embedding-based classification.
- **Literature Support:** Previous studies have shown that combining deep learning features with traditional machine learning models can enhance predictive performance, as they leverage both the power of contextual embeddings and the simplicity of non-deep classifiers.

2) Benefits:

- **Feature Interaction:** The classifier can effectively utilize complex interactions between the embeddings generated by RoBERTa.
- **Efficient Training:** The model is less resource-intensive to train compared to deep learning models, making it a practical choice for combining with pre-extracted features.

3) Training: The Random Forest classifier was trained on the embeddings extracted from the fine-tuned RoBERTa model. The hyperparameters for the classifier were optimized using Optuna, focusing on parameters like n_estimators and max_depth.

```
from sklearn.ensemble import RandomForestClassifier

rf_model = RandomForestClassifier(**best_params)
rf_model.fit(X_train, y_train)
```

This multi-step approach effectively integrates the power of RoBERTa's deep contextual embeddings with the classification capability of Random Forest, providing a comprehensive solution for sentiment analysis in media.

F. Optimization

To achieve robust performance, accelerate training, reduce memory usage, and customize functionalities beyond hyperparameter tuning and optimizer selection, various strategies were implemented. These optimizations ensured that the model was efficient, scalable, and capable of handling the complexities of the large-scale dataset used in this project.

1) Performance and Memory Enhancements:

- **Mixed Precision Training:** Mixed precision training combines 16-bit and 32-bit floating-point operations, allowing for faster computation and reduced memory usage without compromising model accuracy.
  It allows for significant training speed-up and reduces memory consumption, enabling the model to train with larger batch sizes and more complex data.

```
from tensorflow.keras import mixed_precision
mixed_precision.set_global_policy('mixed_float16
    ')
```

- **Class Weighting for Imbalanced Data:** Class weights were computed to balance the dataset during training, ensuring that the model did not favor the majority class. Class weight helps in mitigating class imbalance issues, leading to better generalization and fairness in predictions across all classes.

```
from sklearn.utils.class_weight import
    compute_class_weight
class_weights = compute_class_weight('balanced',
     classes=np.unique(labels), y=labels)
class_weight_dict = dict(enumerate(class_weights
    ))
```

2) Custom Functionalities for Improved Flexibility:

- **Custom Transformer Layer:** A custom layer was created to integrate the transformer model directly into the Keras framework, allowing for more flexible model architectures. This custom layer provides better control over the usage of transformer outputs, enabling the model to be adapted for various downstream tasks.

```
class TransformerLayer(tf.keras.layers.Layer):
    def __init__(self, transformer_model, **
        kwargs):
        super(TransformerLayer, self).__init__
            (**kwargs)
        self.transformer = transformer_model

    def call(self, inputs):
        input_ids, attention_mask = inputs
        outputs = self.transformer(input_ids,
            attention_mask=attention_mask)
        return outputs.last_hidden_state
```

- **Custom Data Generators:** Custom Python data generators were developed to load training data in manageable batches without shuffling indices, preserving data order and minimizing memory usage. These generators allowed data to be processed in chunks, reducing the need to load the entire dataset into memory and improving training efficiency.

```
def data_generator(ids_dataset, masks_dataset,
    labels_dataset, batch_size):
    dataset_size = labels_dataset.shape[0]
    for start_idx in range(0, dataset_size,
        batch_size):
        end_idx = min(start_idx + batch_size,
            dataset_size)
        yield ({'input_ids': ids_dataset[
            start_idx:end_idx], 'attention_mask'
            : masks_dataset[start_idx:end_idx]},
             labels_dataset[start_idx:end_idx])
```

IV. EXPERIMENTS

A. Environmental Setup

The experiments were conducted to assess the performance of the fine-tuned RoBERTa model integrated with a Random Forest classifier. The main objectives were to evaluate how well the model could classify sentiment into five levels and determine its generalization capability across different datasets. The fine-tuned model was trained and validated using a structured pipeline that incorporated preprocessing, feature extraction, and model optimization steps.

1) Hardware and Environment:

- Training was conducted on a single NVIDIA A100 GPU to leverage its high computational power for handling large-scale models like RoBERTa.
- The dataset, comprising English news articles mentioning Singapore, was processed with a maximum token length of 256, resulting in approximately 13GB of data. The training process took over 40 hours.

2) Evaluation Metrics: Accuracy was used as the primary metric to measure the overall performance of the model.

B. Experiment Results

The fine-tuned RoBERTa + Random Forest model demonstrated significant improvements over the baseline:

- Final Model Performance
  - **Accuracy:** 77.9% on the primary Singapore news dataset. This marked an increase of 15% compared to the baseline approach using TF-IDF with logistic regression, which achieved an accuracy of 64%.
  - **Performance on Positive vs. Negative Sentiments:** The model exhibited higher accuracy for detecting positive sentiments compared to negative ones, indicating that while the model effectively captures positive cues, handling subtle or indirect negative sentiments remained more challenging.

Class	Precision	Recall	F1-Score	Support
0 (Strong Negative)	0.77	0.75	0.76	367429
1 (Negative)	0.76	0.69	0.73	367394
2 (Neutral)	0.72	0.71	0.71	367434
3 (Positive)	0.86	0.88	0.87	367421
4 (Strong Positive)	0.78	0.87	0.82	367383
Accuracy			0.78	1837061
Macro Avg	0.78	0.78	0.78	1837061
Weighted Avg	0.78	0.78	0.78	1837061

Fig. 5. Result for Fine-tuned RoBERTa and Random Forest Fustion Model

- Cross-Domain Performance
  - When tested on the IMDB Review Dataset, which features different linguistic aspects such as sarcasm and metaphor, the model's accuracy dropped to 63.18%. This decrease highlights the challenges of domain adaptation, as different text types have unique expressions and linguistic complexities.

Class	Precision	Recall	F1-Score	Support
0(Negative)	0.62	0.70	0.65	12500
2(Positive)	0.65	0.57	0.61	12500
Accuracy			0.63	25000
Macro Avg	0.63	0.63	0.63	25000
Weighted Avg	0.63	0.63	0.63	25000

Fig. 6. Cross Validation Result

V. Inspiration

A. Discussion of Results

The experimental results highlighted the strengths and limitations of the proposed approach, offering key insights into the effectiveness of combining transformer models with traditional classifiers.

- **Performance Insights:** The fine-tuned RoBERTa + Random Forest model achieved an accuracy of 77.9% on the primary Singapore news dataset, marking a significant improvement over the baseline of 64% achieved with TF-IDF and logistic regression. This validated the hypothesis that using a transformer-based model for feature extraction, followed by a robust classifier, could yield higher accuracy and better handling of complex sentiment expressions.
- **Cross-Domain Challenges:** Despite the promising results, the model's performance dropped to 63.18% accuracy when applied to the IMDB Review Dataset. This revealed challenges in domain adaptation, emphasizing that different types of texts, such as movie reviews with sarcasm and complex linguistic nuances, require further tuning or specialized models.
- **Sentiment Analysis Complexity:** The project underscored the complexity of accurately detecting negative sentiments, which can often be expressed using indirect or euphemistic language to maintain objectivity. The model's difficulty in capturing such subtleties highlights the importance of refining feature extraction and model design to handle nuanced expressions

B. Reflections and Key Takeaways

The lessons learned from this project are not only valuable for sentiment analysis in the context of media bias but also contribute to the broader field of NLP applications:

- **Large-Scale Data Processing:** Handling extensive data sets (e.g., over 9 million rows of the Singapore dataset) highlighted the need for efficient data processing methods. Implementing stream processing tools and data generators helped manage memory and computational constraints effectively.
- **Model and Data Adaptation:** The cross-validation results demonstrated that models trained on one dataset might not perform optimally on others with different characteristics. For instance, the IMDB Review Dataset showed how emotional expressions varied, requiring specialized attention. Movie reviews often include sarcasm, puns, or indirect language, making them harder for models trained on straightforward journalistic language to interpret.
- **Handling Negative Sentiments:** The study reinforced the understanding that negative emotions can be conveyed subtly through techniques like euphemism, parody, and complex linguistic structures. The ability to capture these expressions effectively remains a challenge and a research area for developing more adaptive models.

C. Broader Impact

The insights gained from this project can inspire future work in several directions:

- **Cross-Domain Generalization:** Developing domain-agnostic models or fine-tuning strategies that enable better performance across various datasets.
- **Improving Sentiment Analysis:** Enhancing models to better capture indirect language, such as sarcasm and metaphors, using advanced NLP techniques like context-aware transformers or ensemble learning.

The project served as an illustration of how sentiment analysis models can be applied to media data to uncover biases and trends, while also shedding light on the importance of model adaptability to different linguistic contexts.

VI. Conclusion

The project aimed to develop a robust sentiment analysis framework that could accurately classify media content related to Singapore, uncovering biases and trends in international reporting. The approach combined the powerful contextual understanding of a fine-tuned RoBERTa model with the classification capabilities of a Random Forest classifier, resulting in significant improvements over baseline methods.

A. Key Achievements

- **Improved Model Performance:** The integration of RoBERTa and Random Forest achieved an accuracy of 77.9%, a marked increase compared to traditional TF-IDF with logistic regression, which achieved 64%. This confirmed the value of leveraging deep learning for feature extraction and machine learning for robust classification.
- **Domain-Specific Insights:** The analysis revealed that sentiment expressions in news articles often include positive tones that are more easily detected, while negative sentiments can be subtler and harder to capture. The model's ability to handle these challenges highlighted its potential for applications in media bias detection.
- **Scalable Data Handling:** The use of techniques such as data generators and mixed precision training allowed for efficient processing of a large-scale dataset, overcoming significant computational challenges.

B. Challenges and Limitations

Despite the success, several challenges were identified:

- **Domain Adaptation:** The drop in accuracy to 63.18% when applied to a different dataset (e.g., IMDB Review Dataset) underscored the limitations in cross-domain generalization. This indicates the need for further exploration into domain adaptation techniques to improve model flexibility.
- **Complex Sentiment Detection:** The project found that indirect language and complex expressions of negative sentiment remain difficult for models to interpret, pointing to an opportunity for further refinement in handling nuanced emotional language.

C. Future Work

Based on the findings and challenges, several avenues for future research and development are proposed:

- **Domain Adaptation Techniques:** Investigating methods such as domain-specific pre-training or multi-task learning to improve the model's performance across varied datasets.
- **Enhanced Sentiment Models:** Incorporating advanced architectures or ensemble approaches to better handle sarcasm, metaphors, and other indirect expressions of sentiment.
- **Real-Time Processing and Deployment:** Optimizing the model for real-time analysis and deployment in practical applications, such as media monitoring and public sentiment tracking for policy-making.

D. Final Thoughts

The project demonstrated the potential of combining state-of-the-art transformer models with traditional classifiers to address complex NLP tasks. It reinforced the importance of understanding the nuances of sentiment analysis in different contexts and highlighted the ongoing need for adaptive models that can generalize effectively across domains. This work contributes to the broader field of NLP and serves as a foundation for future innovations in media sentiment analysis and bias detection.

References

[1] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: (2018). arXiv: 1801.06146 [cs.CL]. URL: https://arxiv.org/abs/1801.06146.

[2] Y. Liu et al. "RoBERTa: A Robustly Optimized BERT Pretraining Approach". In: (2019). arXiv: 1907.11692 [cs.CL]. URL: https://arxiv.org/abs/1907.11692.