frontiers | Frontiers in Physics

TYPE Original Research
PUBLISHED 05 December 2024
DOI 10.3389/fphy.2024.1477714

OPEN ACCESS

EDITED BY
Salvatore Micciche',
University of Palermo, Italy

REVIEWED BY
Davide Taibi,
National Research Council (CNR), Italy
Giosue Lo Bosco,
University of Palermo, Italy

*CORRESPONDENCE
Luoyao He,
zcbelhe@ucl.ac.uk

RECEIVED 08 August 2024
ACCEPTED 21 November 2024
PUBLISHED 05 December 2024

CITATION
He L (2024) Enhanced twitter sentiment analysis with dual joint classifier integrating RoBERTa and BERT architectures.
Front. Phys. 12:1477714.
doi: 10.3389/fphy.2024.1477714

Enhanced twitter sentiment analysis with dual joint classifier integrating RoBERTa and BERT architectures

Luoyao He*

Biochemical Engineering Department, University College London (UCL), London, United Kingdom

Sentiment analysis, a crucial aspect of Natural Language Processing (NLP), aims to extract subjective information from textual data. With the proliferation of social media platforms like Twitter, accurately determining public sentiment has become increasingly important for businesses, policymakers, and researchers. This study introduces the Dual Joint Classifier (DJC), which integrates the strengths of RoBERTa and BERT architectures. The DJC model leverages Bidirectional Gated Recurrent Units (BiGRU) and Bidirectional Long Short-Term Memory (BiLSTM) layers to capture complex sequential dependencies and nuanced sentiment expressions. Advanced training techniques such as Focal Loss and Hard Sample Mining address class imbalance and improve model robustness. To further validate the DJC model's robustness, the larger TweetEval Sentiment dataset was also included, on which DJC outperformed conventional models despite increased training time. Evaluations were conducted on the Twitter US Airlines and Apple Twitter Sentiment datasets to verify experiments. The DJC model achieved 87.22% and 93.87% accuracies, respectively, and demonstrated improvement over other models like RoBERTa-GLG, BiLSTM(P), and SVM. These results highlight the DJC model's effectiveness in handling diverse sentiment analysis tasks and its potential for real-world applications.

KEYWORDS

sentiment analysis, natural language processing, BERT, RoBERTa, Bi-GRU, Bi-LSTM, hard sample mining

1 Introduction

Sentiment analysis, a subfield of Natural Language Processing (NLP), focuses on extracting subjective information from text data [1, 2]. This task has gained increasing importance with the exponential growth of user-generated content on social media platforms [2]. Accurately determining public sentiment from this data is crucial for businesses, policymakers, and researchers to understand public opinion, track emerging trends, and make informed decisions [2]. Twitter offers a rich real-time data source reflecting public sentiment on various topics, making it a valuable resource for sentiment analysis studies [1, 3].

In recent years, Bidirectional Long Short-Term Memory (BiLSTM) networks have been widely used in NLP tasks due to their ability to capture long-term dependencies in both forward and backward directions. Graves et al. [4] demonstrated the effectiveness of BiLSTMs in sequence labelling tasks [4, 5]. Similarly, Bidirectional Gated Recurrent Unit (BiGRU) networks, proposed by Cho et al. [6], have shown comparable performance

to BiLSTMs with a simpler structure, making them efficient for various sequence modelling tasks. BiLSTM and BiGRU have been effectively employed in sentiment analysis, machine translation, and speech recognition tasks, proving their robustness in handling sequential data [6].

The advent of transformer-based models has marked a significant milestone in the field of Natural Language Processing. Bidirectional Encoder Representations from Transformers (BERT), introduced by Devlin et al. [7], uses a transformer-based architecture to pre-train on a large corpus and fine-tune specific tasks [7]. BERT's bidirectional approach allows it to understand the context from both directions, significantly improving the performance of NLP tasks [8]. Building on the success of BERT, researchers have continued to refine and improve its architecture. The robustly Optimized BERT Pre-training Approach (RoBERTa) was proposed by Liu et al. [9] and builds on BERT by optimizing the pre-training process [9]. It uses more data and computation, removes the next sentence prediction objective, and dynamically changes the masking pattern applied to the training data, resulting in substantial performance improvements over BERT [9–11]. DistilBERT, developed by Sanh et al. (2019), is a distilled version of BERT that retains 97% of BERT's language understanding while being 60% faster and 40% smaller, making it suitable for scenarios requiring reduced computational resources while maintaining high-performance [12].

This paper presents a novel Dual Joint Classifier (DJC) model, which integrates advanced techniques to enhance sentiment analysis performance and robustness, particularly in scenarios with imbalanced datasets and limited resources [13, 14]. The proposed DJC model leverages the combined strengths of RoBERTa and BERT architectures, coupled with BiLSTM, dropout layers and BiGRU layers to capture complex sequential dependencies and improve the model's understanding of context in both forward and backward directions [4, 5]. Dropout layers, introduced by Srivastava et al. (2014) [15], are incorporated to prevent overfitting by randomly deactivating a fraction of neurons during training [15]. Additionally, joint training is employed to integrate these multiple architectures listed above, ensuring that the model benefits from the strengths of each while mitigating their weaknesses. Meanwhile, the proposed model (DJC) also combines techniques such as Focal Loss and Hard Sample Mining to address class imbalance and enhance model learning. First, Focal Loss, introduced by Lin et al. [16], addresses the issue of class imbalance by focusing on hard-to-classify samples, reducing the impact of easy-to-classify examples [16]. In an imbalanced dataset, classes with many samples can dominate the loss function, causing the model to be biased towards these classes and perform poorly on underrepresented classes [16]. Focal Loss addresses this issue by reducing the loss contribution from well-classified examples and putting more focus on hard, misclassified examples, thus improving the model's performance on minority classes [16, 17]. This adjustment improves the model's performance on minority classes by dynamically scaling the loss associated with each sample, making it particularly effective for handling imbalanced datasets [16, 18]. Hard Sample Mining [19], proposed in the context of computer vision by Shrivastava et al. (2016), further augments this by identifying and re-training misclassified samples from previous epochs [20]. This ensures that the model learns from its mistakes and improves its accuracy on challenging data points [19]. By focusing on the hard examples that the model previously struggled with, this technique helps refine the model's understanding and handling of complex cases, thus enhancing overall robustness and accuracy [19, 21, 22].

To evaluate the effectiveness of the proposed DJC model, two widely used public datasets from Kaggle are selected and extensive comparisons are conducted with a baseline BERT-based model, recent state-of-the-art models, and traditional machine learning models. The effectiveness of the proposed DJC model is demonstrated through extensive experiments on diverse benchmark datasets, highlighting its robustness and accuracy in sentiment analysis tasks.

The contributions of this paper can be summarized as follows:

1. The Proposed model employs joint training of RoBERTa and BERT architectures.
2. The model architecture includes additional layers such as BiGRU, BiLSTM, and Dropout layers, which enhance the model's ability to capture sequential dependencies and prevent overfitting.
3. The interaction between Focal Loss and Hard Sample Mining (HSM) techniques enhances the model's ability to handle imbalanced datasets by focusing on hard-to-classify samples.

The rest of this paper is organized as follows. Section 2 reviews related works in sentiment analysis. Section 3 details the DJC model's design and methodology. Section 4 evaluates the DJC model's performance against benchmarks. Section 5 concludes with findings and future work.

2 Related works

Wang et al. [23] introduced a regional CNN-LSTM model for dimensional sentiment analysis, targeting the prediction of valence-arousal (VA) ratings in texts. Their model divides input text into regions, using individual sentences as regions, and applies a regional CNN to extract local affective features, followed by an LSTM to integrate these features sequentially for VA prediction. The model was tested on the Stanford Sentiment Treebank (SST) and Chinese Valence-Arousal Texts (CVAT) datasets, achieving improvements over lexicon-based, regression-based, and conventional NN-based methods, with the best results showing RMSE values of 1.341 for valence and 0.874 for arousal on the CVAT dataset [23]. Joulin et al. [24] introduced FastText, a model designed for efficient text classification, including sentiment analysis. FastText utilizes a bag-of-words approach combined with word vectorization, enabling rapid processing of large-scale datasets. While it offers computational efficiency. The model demonstrated competitive accuracy on benchmark datasets such as Amazon [25] and Yelp reviews [26] but may struggle with the nuanced sentiment found in social media data. Similarly, Singh et al. [27] explored the application of various machine-learning techniques for sentiment analysis to predict outbreaks and epidemics using health-related tweets. They analyzed nearly one and a half million tweets to track illness over time and measure behaviour risk factors, symptoms of diseases, and medication usage. The study employed supervised classification techniques such as Support Vector Machine (SVM), Naïve Bayes, Random Forest, and Decision Tree models.

Their system demonstrated a high correlation with CDC data, achieving 85% accuracy in detecting influenza-related messages [27]. Extending the scope of sentiment analysis to multiple languages, Can et al. [28] developed a multilingual sentiment analysis framework using RNNs, tested on the SemEval-2016 Challenge Task 5 dataset. Their framework achieved accuracies of 84.21% for Spanish, 74.36% for Turkish, 81.77% for Dutch, and 85.61% for Russian, demonstrating the versatility and effectiveness of their RNN-based approach in handling sentiment analysis across different languages.

Loureiro et al. [29] introduced the TimeLMs model, a transformer-based language model pretrained on historical Twitter data to capture temporal changes in language. The model achieved strong performance across various TweetEval tasks, particularly excelling in emotion recognition and offensive language detection, demonstrating its adaptability to evolving social media language. Further advancing the field, Xiang et al. [30] proposed an affection-driven neural network model for sentiment analysis by integrating affective knowledge from the Affect Control Theory (ACT). This approach incorporates an affective lexicon with Evaluation, Potency, and Activity (EPA) values into Long Short-Term Memory (LSTM) models to enhance sentiment classification. The study demonstrated that incorporating these EPA values as numerical influence weights significantly improved the performance of conventional LSTM models, achieving 1.0%–1.5% higher improvements across three large benchmark datasets (Twitter, airline customer reviews, and IMDB movie reviews) [30]. In another recent study, Talaat [31] explored the performance of eight hybrid models combining DistilBERT and RoBERTa with BiGRU and BiLSTM layers for sentiment analysis across three datasets (Airlines, CrowdFlower, and Apple). The study utilized models such as DistilBERT-GLG, RoBERTa-3G, and RoBERTa-LGL, evaluating their accuracy with and without emojis. The results indicated that hybridizing BiGRU and BiLSTM layers improved model performance. Notably, DistilBERT-GLG achieved the highest accuracy for the Airlines dataset (83.74% with emojis, 83.47% without emojis) and RoBERTa-3G showed the best performance for the emoji case (86%) [31].

3 Proposal method

3.1 Data preprocessing

The data preprocessing step involves cleaning and normalizing the text data to prepare it for model training. This includes replacing null values with empty strings, normalizing Unicode characters, and removing hashtags, mentions, URLs, digits, emojis, and non-alphanumeric characters. Additionally, leading and trailing whitespace is stripped from the text.

The sentiment analysis models were tested using two primary datasets from Kaggle: The Apple Twitter Sentiment Dataset [32] and the Twitter US Airline Sentiment Dataset [33]. The Apple dataset includes tweets about Apple products and services, categorised into positive, negative, and neutral sentiments, offering insights into public perception over time. The US Airline dataset involves tweets about major US airlines, similarly categorised, detailing reasons for negative sentiments such as delayed flights or poor service. The TweetEval Sentiment [34] subset was also selected from the broader, multilingual TweetEval dataset to provide additional evaluation across diverse, balanced sentiment categories. Since the pre-trained models in this research were initially trained on TweetEval, using this subset allows for further validation aligned with the training data.

3.2 Dataset overview and splitting

All datasets were simplified to contain only two columns: “text” and “sentiment.” The “sentiment” column was standardised with labels 0 for negative, 1 for neutral and 2 for positive sentiments. Detailed information about the specific tweets in these datasets is shown in Table 1. The Stratified K-Fold cross-validation method is applied to split the dataset. Initially, the dataset was divided into 90% for training and 10% for testing. Then, the training data is applied to 10-fold Stratified K-Fold cross-validation. This means that in each fold, 90% of the training data was used for actual training and 10% for validation. This method ensures that each fold maintains the same class proportions as the entire dataset.

However, the distribution of the three sentiments is not balanced, as shown in Table 1. This imbalance can result in biased model training and reduced performance in accurately predicting minority classes. Strategies to mitigate this issue will be discussed in the following sections.

3.3 Combined model training approach

The sentiment analysis model utilized in this study is based on a combined model training approach that incorporates two primary models available on Hugging Face, referred to as **Roberta** and **BERTweet**.

1. **“CardiffNLP/twitter-roberta-base-sentiment-latest”** is a RoBERTa-base model trained on tweets from January 2018 to December 2021 and fine-tuned for sentiment analysis using the TweetEval benchmark. It classifies English tweets into three sentiments: 0 for Negative, 1 for Neutral, and 2 for Positive.
2. **“FiniteAutomata/bertweet-base-sentiment-analysis”** is another model that leverages the BERTweet architecture and is pre-trained on a large dataset of English tweets for sentiment classification. It also categorizes tweets into three sentiments: 0 for Negative, 1 for Neutral, and 2 for Positive.

3.4 Proposed models- dual joint classifier architecture

Two distinct joint components were developed and combined into a single final model (Dual Joint classifier), as illustrated in Figure 1. The details of these components and layers followed by the two models mentioned in Section 3.3 are provided in Table 2 below.

3.4.1 Joint component 1 architecture

The first joint model component integrates two parallel branches shown in Figure 1: one branch uses RoBERTa, while the

TABLE 1 Twitter sentiment datasets overview.

Datasets	Positive tweets	Neutral tweets	Negative tweets	Total
Apple Twitter Sentiment [32]	686	801	143	1,630
Twitter US Airline Sentiment [33]	2,363	3,099	9,178	14,640
TweetEval Sentiment [35]	10,895	28,958	20,046	59,899
Total	13,944	32,858	29,367	74,702

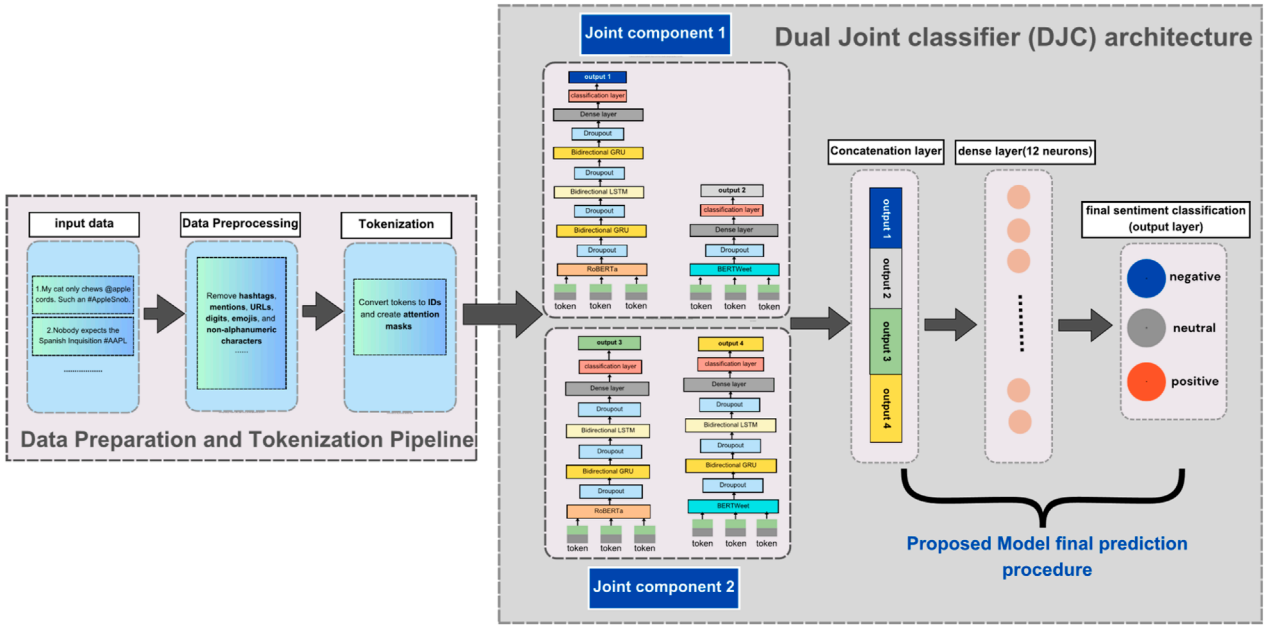

FIGURE 1
The architecture of the proposed model integrating two joint components.

TABLE 2 Overview of joint component architectures and model abbreviations.

Component	Model abbreviation	Layer architecture
Joint component 1	RoBERTa-GLG	Dropout → BiGRU → Dropout → BiLSTM → Dropout → BiGRU → Dropout → Classification Layer
	BERTweet	Dropout → Dense → Classification Layer
Joint component 2	RoBERTa-GL	Dropout → BiGRU → Dropout → BiLSTM → Dropout → Classification Layer
	BERTweet-GL	Dropout → BiGRU → Dropout → BiLSTM → Dropout → Classification Layer

other branch uses BERTweet. The detailed layer configurations and parameter settings for the two base models, RoBERTa and BERTweet, are compiled in Tables 3, 4.

3.4.2 Joint component 2 architecture

Joint component 2, depicted in Figure 1, similarly integrates two parallel branches with RoBERTa on the left and BERTweet on the right. This structure mirrors Joint Component 1, ensuring consistent architecture across components. Detailed layer configurations and parameter settings for RoBERTa and BERTweet are compiled in Tables 5, 6.

The two joint components, joint component 1 (RoBERTa-GLG with BERTweet) and joint component 2 (RoBERTa-GL with BERTweet-GL), are combined into a single final model by

TABLE 3 Joint component 1—detailed architecture of RoBERTa branch.

Layer name	Input dimension	Hidden units	Output dimension
RoBERTa Model	—	—	768
Dropout	768	—	768
Bi-GRU Layer 1	768	256	512 (bidirectional)
Dropout	—	—	512
Bi-LSTM Layer	512	256	512
Dropout	—	—	512
Bi-GRU Layer 2	512	256	512
Dropout	—	—	512
Classifier	512	—	3

TABLE 4 Joint component 1—detailed architecture of BERTweet branch.

Layer name	Input dimension	Hidden units	Output dimension
BERTweetModel	—	—	768
Dropout	—	—	768
Classifier	768	—	3

TABLE 5 Joint component 2—detailed architecture of RoBERTa branch.

Layer name	Input dimension	Hidden units	Output dimension
RoBERTa Model	—	—	768
Dropout	768	—	768
Bi-GRU Layer	768	256	512 (bidirectional)
Bi-LSTM Layer	512	256	512
Dropout	512	—	512
Classifier	512	—	3

TABLE 6 Joint component 2—detailed architecture of BERTweet branch.

Layer name	Input dimension	Hidden units	Output dimension
BERTweetModel	—	—	768
Dropout	768	—	768
Bi-GRU Layer	768	256	512 (bidirectional)
Bi-LSTM Layer	512	256	512
Dropout	512	—	512
Classifier	512	—	3

TABLE 7 Hyperparameters used in model training.

Hyperparameter	Value
Optimizer	AdamW
Learning Rate	1.5e-5
Weight Decay	0.01
Epsilon	1e-8
Number of Epochs	10
Batch Size	16
Dropout Rate	0.3
Loss Function	Focal Loss
Alpha (Focal Loss)	1
Gamma (Focal Loss)	3

first obtaining the outputs (logits) from each component. These logits are concatenated to form a combined representation as represented in Figure 1, which is then passed through a meta-classifier layer. This layer reduces the concatenated logits into a final set of logits corresponding to the sentiment classes. To convert these final logits into predictions, the **SoftMax** function is applied, which transforms the logits into a probability distribution over the classes. The class with the highest probability is selected as the predicted class, indicating the model's final sentiment classification.

3.5 Model training parameters setup

The following sections describe the hyperparameters used for model training and the method for dataset splitting to ensure balanced and robust training. The models are initialized with the following predefined parameters, which are compiled in Table 7:

The hyperparameters listed above were determined through systematic tuning. The learning rate was optimized via grid search within 1e-5~5e-5 for stable convergence, while batch sizes of 16 and 32 were tested to balance memory use and efficiency. AdamW was selected for its adaptive learning rate and weight decay to reduce overfitting. Each setting was refined based on validation outcomes for optimal performance. The selection of Focal Loss parameters, Alpha and Gamma, will be detailed in Section 4.1.1.

Focal Loss (Equation 1) was chosen because the samples were imbalanced, which is mentioned in Section 3.2, making it effective for focusing on harder-to-classify samples. The formula for Focal Loss is given by:

$$FL(p_t) = -a_t(1-p_t)^{\gamma}\log(p_t) \tag{1}$$

Where p_t is the predicted probability for the true class, a_t is a weighting factor for the class t, and γ is the focusing parameter that reduces the relative loss for well-classified examples ($p_t > 0.5$) and puts more focus on hard, misclassified examples.

3.6 Hard sample mining technique

In the training procedure, a technique called hard sample mining is employed to enhance the model's training efficacy. The dataset is first divided into training and testing sets, as mentioned in Section 3.2, to ensure that no data from the test set leaks into the training process. During training, special attention is given to samples on which the model previously performed poorly, known as "hard samples." Specifically, hard samples refer to those instances from the training set that were incorrectly predicted in the previous epoch.

3.6.1 Interaction of focal loss and hard sample mining

The core of the Hard Sample Mining technique is to focus additional attention on samples where the model previously performed poorly during training, by re-training on these samples to improve the model's performance. On the other hand, Focal Loss already adjusts the weight of these hard-to-classify samples at the loss function level, giving them greater significance in the loss calculation. When combined with Hard Sample Mining, these samples identified as difficult by Focal Loss are further emphasized in subsequent training iterations, enhancing the model's ability to learn from these challenging cases.

The Hard Sample Mining process involves identifying and focusing on "hard samples"—those incorrectly classified by the model—across multiple training iterations. Initially, these hard samples are identified based on their classification errors and are given additional attention in subsequent training iterations. As training progresses, the model adjusts to better handle these challenging samples, resulting in improvement. Over time, a greater portion of these hard samples are correctly classified, signifying enhanced model performance. Throughout this process, the model improves progressively as it re-evaluates and correctly classifies these hard samples, moving them from a misclassified status to a correctly classified one. Each training iteration brings the model closer to robustness by reducing misclassifications, particularly for samples that are difficult to categorize. This iterative process ultimately contributes to a more accurate model that can effectively address classification difficulties in the data.

3.7 Comparative evaluation of BERT-Based models

To underscore the efficacy of the Proposed Model (DJC), several widely recognized BERT-based models were employed for comparison. These models, include RoBERTa, BERT, and DistilBERT variants used in the sentiment analysis tasks. The DistilBERT model used in the evaluation is "DistilBERT-Base-Uncased-Emotion" from Hugging Face. The dataset split ratios and the hyperparameters for these models, such as loss function and training setup, were kept consistent with those used for the Proposed Model (DJC). This ensures a fair comparison and accurately reflects the Proposed Model's effectiveness.

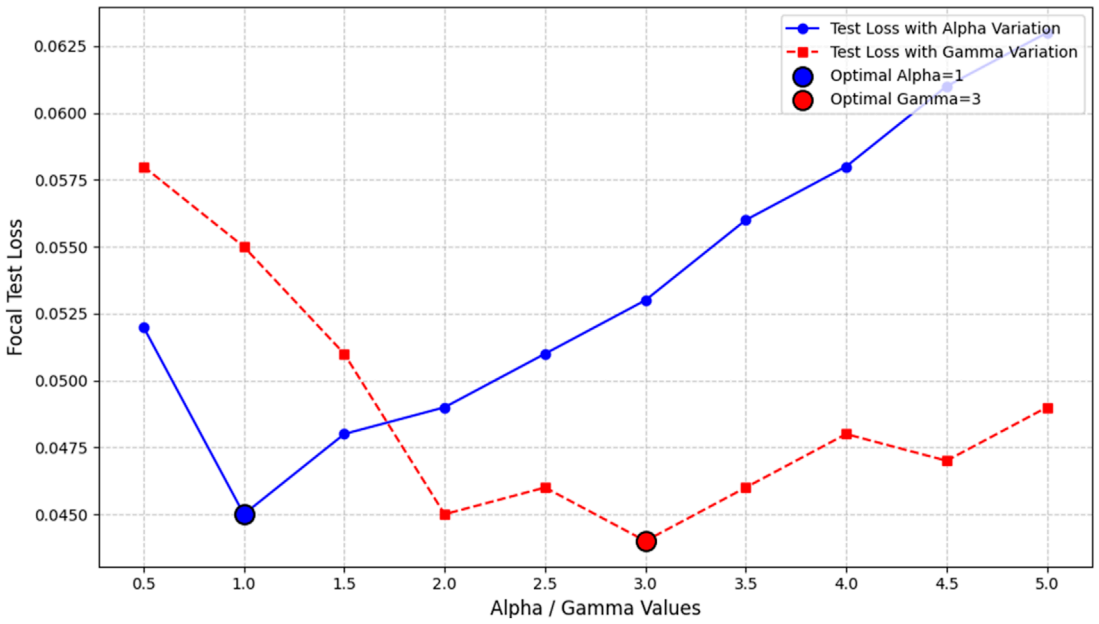

FIGURE 2
Impact of Alpha and Gamma tuning on focal loss performance during model evaluation.

4 Experimental analysis and results

All experiments were conducted on a cloud server with the following configuration: Ubuntu 20.04.2 LTS operating system, NVIDIA Tesla V100 GPU, 64GB RAM, Python version 3.8.5, and PyTorch 1.7.1 as the deep learning framework. Parameters like batch size, learning rate, and optimizer settings were kept consistent to ensure fair comparisons of training time across models.

4.1 Results and charts

4.1.1 Optimal tuning of focal loss hyperparameters: Alpha (α) and Gamma (γ)

Alpha (α) balances sample distribution across classes, assigning a higher weight to minority classes to improve the model's ability to learn from these samples. **Gamma (γ)** adjusts the model's focus on hard-to-classify samples; higher Gamma values place more emphasis on challenging samples, enhancing robustness to noisy or extreme cases. Through grid and progressive tuning, Alpha and Gamma were adjusted over a practical range. The lowest test loss was achieved at **Alpha = 1** and **Gamma = 3**, as highlighted in Figure 2. This optimal combination aligns well with the key properties of focal loss by down-weighting well-classified samples and emphasizing more challenging cases, which counters class imbalance effectively. The test loss value was achieved on the Twitter US Airline Sentiment dataset.

4.1.2 Impact of joint model and BiLSTM-BiGRU on performance

The proposed DJC architecture's innovation in combining the outputs of these two joint components into a single final model has led to significant improvements in performance metrics, as shown in Table 8; Figure 3. The concatenated logits from both components provide a richer representation of the data, which is then processed by a meta-classifier layer to generate the final sentiment predictions. Specifically, on the Twitter US Airline Sentiment Dataset, DJC achieved the highest accuracy (87.22%), precision (84.67%), recall (87.74%), and F1 score (86.03%). On the Apple Twitter Sentiment Dataset, DJC also demonstrated superior performance with the highest accuracy (93.87%), precision (93.10%), recall (87.84%), and F1 score (90.08%). The interaction between the BiGRU and BiLSTM layers in both components ensures that the model effectively captures both short-term and long-term dependencies, leading to superior performance across various datasets. However, there is a slight decline in recall on the Apple Twitter Sentiment Dataset compared to DistilBERT. This could be due to DJC's tendency to be more precise at the expense of missing some relevant instances, indicating a trade-off between precision and recall. In addition to the above datasets, the larger TweetEval Sentiment dataset was included to further validate the robustness of the DJC model across diverse and larger datasets. DJC achieved an accuracy of 81.67%, precision of 70.38%, recall of 77.25%, and F1 score of 78.50% on TweetEval Sentiment, surpassing other models tested on the same dataset as shown in Figure 3.

However, due to the larger data volume, training times increased significantly. Specifically, DJC's training time was 7.50 h as shown in Table 9, compared to 6.66 h for RoBERTa and 5.50 h for DistilBERT, representing a 12.6% and 36.4% increase, respectively, in training duration. These increases reflect the computational demands of DJC's dual architecture but are offset by its superior accuracy and robustness, as evidenced by its performance across all datasets. n addition to the original comparisons, RoBERTa-GL and RoBERTa-GLG models, which incorporate LSTM and GRU

TABLE 8 Performance metrics on twitter US airline sentiment dataset.

Datasets	Models	Accuracy	Precision	Recall	F1 score
Airline Twitter	Distillbert	81.63%	76.48%	86.93%	78.88%
	Bertweet	84.63%	80.80%	79.23%	79.37%
	RoBERTa	85.86%	81.56%	82.17%	81.80%
	DJC (proposed)	87.22%	84.67%	87.74%	86.03%
Apple Twitter	Distillbert	90.18%	93.05%	83.31%	86.83%
	Bertweet	91.41%	93.86%	86.11%	89.15%
	RoBERTa	90.80%	88.24%	83.73%	85.62%
	DJC (proposed)	93.87%	93.10%	87.84%	90.08%
TweetEval Sentiment	Distillery	63.50%	64.64%	63.50%	63.52%
	Bertweet	73.00%	70.39%	72.23%	70.63%
	RoBERTa	75.50%	73.65%	70.54%	71.51%
	RoBERTA-GL	78.00%	77.88%	74.51%	75.88%
	RoBERTa-GLG	79.27%	75.63%	74.60%	74.84%
	DJC (proposed)	81.67%	70.38%	77.25%	78.50%

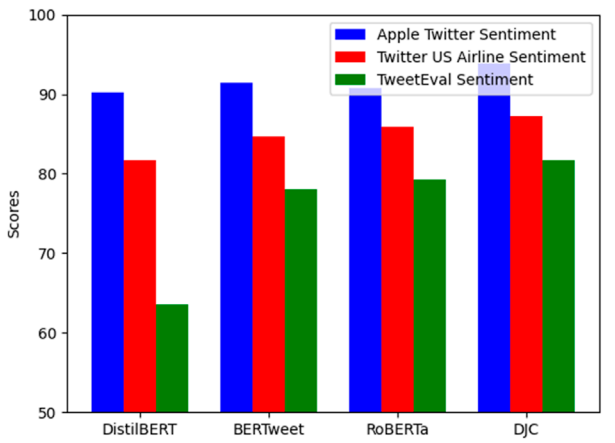

FIGURE 3
Performance comparison of BERT-Based models on different datasets.

layers, demonstrated training times of 7.30 h on the larger TweetEval Sentiment dataset. This reflects the additional computational load introduced by sequential dependency layers. Training times can also vary based on hardware and implementation. Modern GPUs, like the NVIDIA Tesla V100 or A100, have high peak FLOPS but rarely achieve full utilization due to memory bandwidth limitations, data transfer times, and non-optimal batch processing. Studies indicate that utilization rates are often as low as 30%–40% in real-world settings for large models, which impacts efficiency despite powerful hardware [36]. As such, the slightly extended training time of DJC is a calculated investment, reflecting the model's complexity and its ability to deliver superior performance across diverse and challenging datasets.

4.1.3 Performance analysis of proposed model with hard sample mining

Table 10 llustrate the performance of the proposed model (DJC) on different datasets with and without Hard Sample Mining (HSM). The performance metrics indicate that incorporating Hard Sample Mining (HSM) improves the model's performance across both datasets. Specifically, the Apple Twitter Sentiment dataset shows a notable increase in Recall (from 84.63% to 87.84%) and F1 Score (from 88.18% to 90.08%), despite a slight decrease in Precision. The overall accuracy also improves from 92.02% to 93.87%. Similarly, in the Twitter US Airline Sentiment dataset, the Recall and F1 Score significantly improve with HSM, suggesting that the model becomes better at identifying true positives and maintaining a balance between precision and recall. The Accuracy also increases from 85.86% to 87.22%. For the TweetEval Sentiment dataset, the inclusion of HSM results in an increase in Recall (from 77.25% to 86.93%) and F1 Score (from 78.50% to 78.88%), though the overall impact on Accuracy is minimal, showing a slight increase from 81.63% to 81.67%. This indicates that while HSM enhances the model's performance in identifying relevant instances across all datasets, the effect is most pronounced in datasets with a higher imbalance, as seen in the Twitter US Airline Sentiment dataset.

The datasets exhibit significant imbalances, particularly in the Twitter US Airline Sentiment dataset, where negative

TABLE 9 Training time comparison (hours) across datasets and models.

Model	Apple twitter	Airline twitter	TweetEval sentiment
DistillBERT	0.20	2.00	5.50
BERTweet	0.25	2.50	6.66
RoBERTa	0.35	2.83	7.08
RoBERTA-GL	0.35	2.90	7.30
RoBERTa-GLG	0.35	3.00	7.30
DJC (Proposed)	0.35	3.15	7.50

TABLE 10 Performance comparison with and without hard sample mining (HSM) on 3 datasets.

Datasets	Accuracy	Precision	Recall	F1 score	Test loss
Airline Twitter	87.22%	84.67%	87.74%	86.03%	0.0053
Airline (No HSM)	85.86%	82.06%	81.45%	81.51%	0.0046
Apple Twitter	93.87%	93.10%	87.84%	90.08%	0.0063
Apple (No HSM)	92.02%	94.38%	84.63%	88.18%	0.0035
TweetEval Sentiment	81.67%	76.48%	86.93%	78.88%	0.0075
TweetEval (No HSM)	81.63%	70.38%	77.25%	78.50%	0.0064

TABLE 11 Performance metrics by category with and without HSM for imbalanced datasets.

Datasets	Category	Accuracy	Precision	Recall	F1 score
Apple Twitter	Negative	92.64%	55.04%	87.84%	67.67%
	Neutral	90.51%	92.48%	87.84%	67.67%
	Positive	90.88%	90.24%	87.84%	89.03%
Apple (No HSM)	Negative	93.52%	59.15%	84.63%	69.63%
	Neutral	89.58%	93.56%	84.63%	88.87%
	Positive	90.27%	91.62%	84.63%	87.98%
Airline Twitter	Negative	86.59%	90.58%	87.74%	89.13%
	Neutral	85.32%	60.58%	87.74%	71.67%
	Positive	85.17%	52.42%	87.74%	65.63%
Airline (No HSM)	Negative	81.68%	88.41%	81.45%	84.78%
	Neutral	81.93%	54.93%	81.45%	65.61%
	Positive	81.96%	46.63%	81.45%	59.31%

tweets dominate. The Apple Twitter Sentiment dataset has 1,630 total tweets, with 686 positive, 801 neutral, and 143 negative tweets. The Twitter US Airline Sentiment dataset contains 14,640 total tweets, with 2,363 positive, 3,099 neutral, and 9,178 negative tweets. Table 11 reflects these imbalances and their impact on performance metrics. For the Apple Twitter dataset, HSM enhances the performance metrics, with noticeable improvements in recall and F1 score for all sentiment categories, particularly negative sentiment. In the Twitter US Airline dataset, where positive and neutral tweets are significantly fewer, HSM substantially boosts

TABLE 12 Comparative evaluation of tweet sentiment analysis models.

Dataset	Used model	Accuracy
Twitter US Airlines	RNN/LSTM (ULMFiT) [28]	77.8%
	LSTM, CNN [23]	79.64%
	MultinomialNB [27]	±80%
	BiLSTM(P) [30]	82%
	RoBERTa-GLG [31]	85.93%
	DJC (proposed)	87.22%
Apple Twitter	SVM [31]	84.05%
	DistillBERT-GLG [31]	88.04%
	RoBERTa-GLG [31]	90.18%
	RoBERTa-LGL [31]	90.49%
	RoBERTa-3G [31]	91.72%
	DJC (proposed)	93.87%
TweetEval Sentiment	LSTM [34]	58.37%
	BLSTM [34]	58.34%
	SVM [34]	62.91%
	FastText [34]	62.98%
	RoBERTa-Twitter [34]	69.14%
	TimeLMs-19 [29]	73.20%
	TimeLMs-21 [29]	73.70%
	DJC (proposed)	81.67%

these categories' precision, recall, and F1 scores, indicating better handling of imbalanced data. In contrast, the impact of HSM is less pronounced in the **TweetEval Sentiment dataset** due to its relatively balanced distribution, and thus, it is excluded from this table for clearer comparative insights.

4.1.4 Comparison with other models

To benchmark the performance of the proposed DJC model, it was compared against several existing models in the field of sentiment analysis represented in Table 12, the bold models are the proposed models in this research. The superior performance of the proposed DJC model can be attributed to its innovative architecture that effectively combines the strengths of multiple approaches. Specifically, the DJC model integrates two joint components, RoBERTa and BERTweet, using BiGRU and BiLSTM layers. This integration allows the model to capture both short-term and long-term dependencies in the data, leading to more accurate sentiment predictions.

Previous models, such as those utilizing LSTM, CNN, or MultinomialNB, focus on either sequential dependencies or simple probabilistic methods, which may not fully capture the complex patterns in sentiment data. For instance, the RNN/LSTM (ULMFiT) and LSTM-CNN models leverage sequential dependencies but may lack the depth provided by combining multiple advanced architectures. The BiLSTM(P) model and RoBERTa-GLG improve performance by integrating bidirectional layers and transformer-based embeddings, yet they still fall short of the comprehensive approach used in the DJC model. The TweetEval Sentiment dataset highlights the limitations of various traditional and simpler models in capturing sentiment complexity. LSTM and BLSTM, known for handling sequential data, struggle with long-range dependencies and may overlook sentiment nuances, especially in larger datasets. Their performance tends to fall short in identifying indirect expressions common in tweets. SVM, effective for smaller datasets, suffers in scalability with larger data like TweetEval Sentiment due to high

computational costs, which reduces its ability to manage complex sentiment cues and nonlinear relationships essential for accurate sentiment analysis.

FastText, although efficient, relies on word vectorisation and cannot capture word order or syntactic subtleties. This limitation hampers its ability to detect sentiment in contexts where order and subtleties shape the emotional tone, such as sarcasm or indirect sentiment expression. RoBERTa-Twitter, fine-tuned on Twitter data, shows an advantage over traditional models with its transformer-based approach, achieving better accuracy; however, it lacks additional handling of nuanced and ambiguous sentiment due to its single-model design. The TimeLMs-19 and TimeLMs-21 models are based on a RoBERTa architecture, pretrained on a large corpus of tweets from 2019 to 2021, respectively. This temporal pretraining allows them to specialize in capturing evolving language trends specific to certain periods, such as emerging slang, hashtags, and abbreviations on Twitter. Structurally, both TimeLMs versions rely on the standard RoBERTa architecture without modifications for deeper sequence modelling or class imbalance handling, which means they perform well on temporally contextualized language but may lack the versatility for broader sentiment analysis tasks. The DJC model's use of Hard Sample Mining (HSM) and Focal Loss further enhances its ability to handle imbalanced data, a common challenge in sentiment analysis. By focusing on hard-to-classify samples and adjusting the loss function to prioritize these cases, the DJC model achieves higher recall and F1 scores, as well as overall accuracy.

5 Conclusion

The proposed Dual Joint Classifier (DJC) model demonstrates significant advancements in sentiment analysis, attributed to its innovative architecture and the incorporation of Hard Sample Mining (HSM) and Focal Loss techniques. By combining RoBERTa and BERT models with BiGRU and BiLSTM layers, the DJC model effectively captures both short-term and long-term dependencies in data, leading to superior performance metrics. Specifically, the DJC model achieved the highest accuracy (87.22%), precision (84.67%), recall (87.74%), and F1 score (86.03%) on the Twitter US Airline Sentiment dataset. On the Apple Twitter Sentiment dataset, it also demonstrated superior performance with the highest accuracy (93.87%), precision (93.10%), recall (87.84%), and F1 score (90.08%). The inclusion of the larger TweetEval Sentiment dataset further validates the DJC model's robustness across diverse datasets, highlighting its superior accuracy and F1 score over conventional models like LSTM, BLSTM, and SVM, which often struggle with complex social media sentiment. Despite a 12.6% and 36.4% increase in training time compared to RoBERTa and DistilBERT, DJC's enhanced performance justifies the computational cost, particularly in handling imbalanced data effectively. The use of HSM and Focal Loss addresses class imbalance issues, enhancing the model's ability to accurately identify and predict sentiments. This comprehensive approach results in outperforming existing models and setting a new benchmark in sentiment analysis. Future work will explore the application of the DJC model to other domains and languages, further enhancing its robustness and versatility. To further improve DJC's efficiency, future implementations could incorporate selective weight updates and optimized parallelism in distributed training, reducing training time and computational resource demands.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: Apple Twitter Sentiment Dataset: https://www.kaggle.com/seriousran/appletwittersentimenttexts Twitter US Airline Sentiment Dataset: https://www.kaggle.com/crowdflower/twitter-airline-sentiment.

Ethics statement

Ethical approval was not required for the study involving human data in accordance with the local legislation and institutional requirements. Written informed consent was not required, for either participation in the study or for the publication of potentially/indirectly identifying information, in accordance with the local legislation and institutional requirements. The social media data was accessed and analyzed in accordance with the platform's terms of use and all relevant institutional/national regulations.

Author contributions

LH: Writing–original draft, Writing–review and editing, Conceptualization, Data curation, Formal Analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

1. Wankhade M, Rao ACS, Kulkarni C. A survey on sentiment analysis methods, applications, and challenges. *Artif Intelligence Rev*(2022) 55(7):5731–80. doi:10.1007/s10462-022-10144-1

2. Yue L, Chen W, Li X, Zuo W, Yin M. A survey of sentiment analysis in social media. *Knowledge Inf Syst*(2019) 60:617–63. doi:10.1007/s10115-018-1236-4

3. Medhat W, Hassan A, Korashy H. Sentiment analysis algorithms and applications: a survey. *Ain Shams Eng J*(2014) 5(4):1093–113. doi:10.1016/j.asej.2014.04.011

4. Graves A, Fernández S, Schmidhuber J. Bidirectional LSTM networks for improved phoneme classification and recognition. In: *International conference on artificial neural networks*. Warsaw, Poland: Springer(2005).

5. Zhou P, Shi W, Tian J, Qi Z, Li B, Hao H, et al. Attention-based bidirectional long short-term memory networks for relation classification. In: *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: short papers)*(2016).

6. Cho K, Van Merriënboer B, Bahdanau D, Bengio Y. On the properties of neural machine translation: encoder-decoder approaches(2014) arXiv preprint arXiv:14091259.

7. Devlin J, Chang M-W, Lee K, Toutanova K. Bert: pre-training of deep bidirectional transformers for language understanding(2018) arXiv preprint arXiv:181004805.

8. Koroteev MV. BERT: a review of applications in natural language processing and understanding(2021). arXiv preprint arXiv:210311943.

9. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. Roberta: a robustly optimized bert pretraining approach(2019) arXiv preprint arXiv:190711692.

10. Delobelle P, Winters T, Berendt B. Robbert: a Dutch roberta-based language model(2020) arXiv preprint arXiv:200106286.

11. Warstadt A, Zhang Y, Li H-S, Liu H, Bowman SR. Learning which features matter: RoBERTa acquires a preference for linguistic generalizations (eventually)(2020) p. 217–35. arXiv preprint arXiv:201005358. doi:10.18653/v1/2020.emnlp-main.16

12. Sanh V, Debut L, Chaumond J, Wolf T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter(2019) arXiv preprint arXiv:191001108.

13. Chen W, Du J, Zhang Z, Zhuang F, He Z. A hierarchical interactive network for joint span-based aspect-sentiment analysis(2022) arXiv preprint arXiv:220811283.

14. Li Y, Yang Z, Yin C, Pan X, Cui L, Huang Q A joint model for aspect-category sentiment analysis with shared sentiment prediction layer. In: Chinese Computational Linguistics: 19th China National Conference, CCL 2020, Hainan, China, October 30–November 1, 2020, Proceedings 19. Springer(2020).

15. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. *J machine Learn Res*(2014) 15(1):1929–58. doi:10.5555/2627435.2670313

16. Lin T-Y, Goyal P, Girshick R, He K, Dollár P. In: *Focal loss for dense object detection. Proceedings of the IEEE international conference on computer vision*(2017).

17. Shuang K, Lyu Z, Loo J, Zhang W. Scale-balanced loss for object detection. *Pattern Recognition*(2021) 117:107997. doi:10.1016/j.patcog.2021.107997

18. Mukhoti J, Kulharia V, Sanyal A, Golodetz S, Torr P, Dokania P. Calibrating deep neural networks using focal loss. *Adv Neural Inf Process Syst*(2020) 33:15288–99.

19. Sheng H, Zheng Y, Ke W, Yu D, Cheng X, Lyu W, et al. Mining hard samples globally and efficiently for person reidentification. *IEEE Internet Things J*(2020) 7(10):9611–22. doi:10.1109/jiot.2020.2980549

20. Shrivastava A, Gupta A, Girshick R. Training region-based object detectors with online hard example mining. In: Proceedings of the IEEE conference on computer vision and pattern recognition(2016).

21. Baker W, Colditz JB, Dobbs PD, Mai H, Visweswaran S, Zhan J, et al. Classification of twitter vaping discourse using BERTweet: comparative deep learning study. *JMIR Med Inform*(2022) 10(7):e33678. doi:10.2196/33678

22. Chen K, Chen Y, Han C, Sang N, Gao C. Hard sample mining makes person re-identification more efficient and accurate. *Neurocomputing*(2020) 382:259–67. doi:10.1016/j.neucom.2019.11.094

23. Wang J, Yu L-C, Lai KR, Zhang X. Dimensional sentiment analysis using a regional CNN-LSTM model. In: *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: short papers)*(2016).

24. Joulin A, Grave E, Bojanowski P, Mikolov T. Bag of tricks for efficient text classification(2016) arXiv preprint arXiv:160701759.

25. Hou Y, Li J, He Z, Yan A, Chen X, McAuley J. Bridging language and items for retrieval and recommendation(2024) arXiv preprint arXiv:240303952.

26. Rendle S, Krichene W, Zhang L, Anderson J. Neural collaborative filtering vs. matrix factorization revisited. In: Proceedings of the 14th ACM Conference on Recommender Systems(2020) p. 240–8. doi:10.1145/3383313.3412488

27. Singh R, Singh R, Bhatia A. Sentiment analysis using Machine Learning technique to predict outbreaks and epidemics. *Int J Adv Sci Res*(2018) 3(2):19–24.

28. Can EF, Ezen-Can A, Can F. Multilingual sentiment analysis: an RNN-based framework for limited data(2018) arXiv preprint arXiv:180604511.

29. Loureiro D, Barbieri F, Neves L, Anke LE, Camacho-Collados J. TimeLMs: diachronic language models from twitter. arXiv preprint arXiv:220203829. 2022.

30. Xiang R, Long Y, Wan M, Gu J, Lu Q, Huang C-R. Affection driven neural networks for sentiment analysis. In: Proceedings of the Twelfth Language Resources and Evaluation Conference(2020).

31. Talaat AS. Sentiment analysis classification system using hybrid BERT models. *J Big Data*(2023) 10(1):110. doi:10.1186/s40537-023-00781-w

32. Seriousran. Apple twitter sentiment texts(2019). Available from: https://www.kaggle.com/seriousran/appletwittersentimenttexts Accessed August 15, 2024.

33. CrowdFlower. Twitter US airline sentiment(2015). Available from: https://www.kaggle.com/crowdflower/twitter-airline-sentiment Accessed August 15, 2024.

34. Barbieri F, Camacho-Collados J, Neves L, Espinosa-Anke L. Tweeteval: unified benchmark and comparative evaluation for tweet classification(2020) arXiv preprint arXiv:201012421.

35. Qiu L, Zhao Y, Shi W, Liang Y, Shi F, Yuan T, et al. Structured attention for unsupervised dialogue structure induction(2020) p. 1889–99. arXiv preprint arXiv:200908552. doi:10.18653/v1/2020.emnlp-main.148

36. Mittal S, Vaishay S. A survey of techniques for optimizing deep learning on GPUs. *J Syst Architecture*(2019) 99:101635. doi:10.1016/j.sysarc.2019.101635