e-JURNAL PENYELIDIKAN DAN INOVASI
ISLAMIC INNOVATION CENTER FOR TECHNOLOGY, UIS
https://ejpi.uis.edu.my/index.php/ejpi
e-ISSN 2289-7909

VOL. 13 NO. 2 (AUGUST, 2026) Special Edition AIDI
Page 174-187
DOI: https://doi.org/10.53840/e-jpi.v13i2.422

Received: 3 August 2026 Accepted: 18 August 2026 Published: 31 August 2026

Comparative Analysis of Machine Learning and Deep Learning Approaches for Sentiment Analysis across Heterogeneous Datasets under Class Imbalance Conditions

Analisis Perbandingan Pendekatan Pembelajaran Mesin dan Pembelajaran Mendalam untuk Analisis Sentimen Merentas Set Data Heterogen dalam Keadaan Ketidakseimbangan Kelas

Shakirah Mohd Sofi[1], Ali Selamat[2], Zatul Alwani Shaffiei[3]

[1] Department of Computing, Faculty of Creative Multimedia & Computing, Selangor Islamic University (UIS), 43000 Bandar Seri Putra, Kajang, Selangor, Malaysia
[1, 2, 3] Malaysia-Japan International Institute of Technology (MJIIT), Universiti Teknologi Malaysia Kuala Lumpur, Jalan Sultan Yahya Petra, Kuala Lumpur 54100, Malaysia
[2] Faculty of Computing, Faculty of Engineering, & Media and Games Center of Excellence (MagicX), Universiti Teknologi Malaysia, Skudai 81310, Johor Bahru, Malaysia
Email: [1]syakirah@uis.edu.my, [2]aselamat@utm.my, [3]zatulalwani.kl@utm.my

ABSTRACT

Sentiment analysis across heterogeneous datasets remains a significant challenge because datasets from different domains exhibit substantial variations in vocabulary, writing style, text length, and class distribution, making it difficult to develop models that perform consistently across diverse datasets. To address this issue, this paper presents a comparative analysis of machine learning and deep learning approaches for sentiment analysis across heterogeneous datasets on three different datasets COVIDSENTI, Yelp reviews and IMDb movie reviews. These datasets were intentionally selected to reflect different text-length categories, namely short, medium, and long texts. Such variation allows a more comprehensive evaluation of model performance under different linguistic characteristics and textual complexities across domains. Three class imbalance handling strategies were studied with no balancing, ROS, and ROS combined with class weighting. The models were evaluated including LR, SVM, NB, CNN, BiLSTM and BERT. The results show that ROS greatly enhanced the F1-score for imbalanced datasets and COVIDSENTI achieved the greatest improvement +8.83% for LR. The results indicated that BERT outperformed machine learning and deep learning baselines with an F1-score of 95.32%, 73.79%, and 75.76% on COVIDSENTI, Yelp, and IMDb accordingly. An extra benefit of class weighting was shown in deep learning models on COVIDSENTI but not for machine learning models as an improvement after balancing with ROS. Future work will construct a hybrid BERT+BiLSTM+Attention framework for improving generalization across heterogeneous datasets.

Keywords: Sentiment analysis, heterogeneous, machine learning, deep learning, BERT, class imbalance, Random OverSampling, CNN, BiLSTM

INTRODUCTION

Sentiment analysis, also known as opinion mining, is one of the more well-established applications of natural language processing (NLP). The process comprises the classification of text's affective polarity a positive, negative, or neutral tone. Research in this field has been significantly expedited by the rapid expansion of electronic commerce, social media, and online consumer review platforms, which generate large volumes of text that are rich with personal perspectives and opinions. As a result, the ability to automatically analyze and interpret sentiment across diverse sources and domains has become increasingly critical for both academic research and real-world applications.

One of the more persistent problems in sentiment analysis is that models trained on a specific domain frequently fail to generalize effectively when applied to a different domain. This phenomenon, widely referred to as domain shift (Du et al., 2020; Kumar et al., 2025) is particularly disruptive when source and target domains differ significantly. This performance degradation is particularly pronounced when the source and target domains exhibit substantial differences in vocabulary, writing style, and text length, as is the case between health-related social media posts and feature-length movie reviews. This challenge is further complicated by the diversity of text length across domains, as short texts, such as tweets, require fundamentally different feature representations than long review texts (Tsirmpas et al., 2024). Current approaches often fail to generalize across heterogeneous domains with varying linguistic features, although significant efforts are being made to address domain shift through transfer learning and domain adaptation techniques (Alahmadi et al., 2025; Mohd Sofi et al., 2026).

Class imbalance remains a common issue in practical sentiment analysis applications. Many datasets rarely exhibit balanced class distributions, and neutral sentiments are often underrepresented relative to positive and negative classes. This imbalance issue becomes more challenging because irregular distribution can lead models to focus on the majority classes, resulting in poorer prediction performance for minority classes (Permataning Tyas et al., 2023; Talukder et al., 2025).

Although class imbalance has been studied from the perspective of data level or algorithm level, most assessments have been performed on a single dataset or a particular type of model. There is still limited empirical evidence regarding how different imbalance handling strategies behave consistently across heterogeneous datasets with varying text lengths and across different model families, including machine learning, deep learning, and transformer-based approaches. Therefore, this study systematically investigates the impact of three imbalance handling strategies under multiple sentiment analysis scenarios.

In preliminary studies on sentiment analysis, most researchers have employed traditional machine learning (ML) approaches, such as Logistic Regression (LR), Support Vector Machines (SVM) and Naive Bayes (NB), with Term Frequency-Inverse Document Frequency (TF-IDF) features (Bordoloi & Biswas, 2023; Mao et al., 2024). Performance would be negatively impacted when used for multiple domains, although effective in only one domain. Furthermore, deep learning (DL) models such as Convolutional Neural Networks (CNN) and Bidirectional Long Short-Term Memory (BiLSTM) were introduced to capture deeper semantic features. CNN is good for local features and BiLSTM captures bidirectional sequence dependencies (Abdullah & Ahmet, 2023; Younesi et al., 2024). However, these models still rely on static word embeddings which are limited for different domains. Recently, transformer-based models such as Bidirectional Encoder Representations from Transformers (BERT) have shown superior performance in a variety of NLP tasks including sentiment analysis (He, 2023; Kumar et al., 2025). BERT uses a richer contextual representation than previous approaches. However, most previous studies have only tested in a single domain. Systematic comparisons of ML, DL and BERT across heterogeneous domains with varying text length are still understudied (Alahmadi et al., 2025; Tsirmpas et al., 2024).

This paper addresses this challenge through a systematic comparative study of ML and DL approaches for sentiment analysis across domains. The main contributions of this work are:

i. A comprehensive evaluation of the performance of 6 models (LR, SVM, NB, CNN, BiLSTM, BERT) across 3 heterogeneous datasets COVIDSENTI, Yelp Reviews and IMDb representing short, medium and long text lengths.
ii. Providing empirical evidence on the effectiveness of three class imbalance handling strategies across heterogeneous datasets using machine learning, deep learning, and transformer-based models.
iii. A comprehensive empirical comparison and guidance on selecting appropriate strategies for comparative sentiment analysis across multiple domains.

The remainder of this paper is arranged as follows: The paper is organized as follows. Section 2 reviews the relevant work. Section 3 explains the approach. Section 4 analyzes the results. Section 5 closes the study and presents future work.

RELATED WORK

Sentiment Analysis Using Machine Learning

Despite the development of DL algorithms, traditional ML classifiers such as LR and SVM are still performing well with TF-IDF feature extraction (Mao et al., 2024; Mohd Sofi & Selamat, 2023). Due to effectiveness and simplicity, these approaches are commonly used as baseline models to prove benchmarks against which the performance of more sophisticated models can be evaluated (Bordoloi & Biswas, 2023). NB is computationally efficient, but its performance is often limited when dealing with complex linguistic patterns and imbalanced class distributions (Agrawal et al., 2025). Traditional ML methods have achieved satisfactory results in single-domain scenarios by relying on TD-IDF representations. However, there are limitations in their ability to recognize contextual semantic relationships (Siino et al., 2024). Therefore, the use of DL approaches has been recommended (Mao et al., 2024). Despite their competitive performance in balanced datasets, traditional ML models remain highly dependent on handcrafted features such as TF-IDF, limiting their ability to capture contextual semantics across heterogeneous datasets.

Deep Learning for Sentiment Analysis

DL approaches have shown superiority over traditional approaches for sentiment analysis because they can learn appropriate features from text input automatically without requiring any human feature engineering (Abdullah & Ahmet, 2023). CNN are efficient in learning local n-gram features using convolutional filters that are useful in detecting short-term emotion patterns in text (Younesi et al., 2024). The BiLSTM is an outstanding choice for longer texts, as it captures sequence dependencies in both forward and backward directions. This is especially useful for longer writings such as product reviews and movie reviews (Ramaziyah & Setiawan, 2024). It has been shown that the integration of CNN and BiLSTM is more successful because CNN identifies local characteristics while BiLSTM captures global context (He, 2023; Yang et al., 2025). However, DL models still rely on static word embeddings that cannot capture contextual meaning (Tsirmpas et al., 2024). Although CNN and BiLSTM improve semantic representation, their reliance on static embeddings restricts their adaptability to domain-specific vocabulary and varying text lengths.

BERT and Transformer-Based Approaches

BERT introduced by Devlin et al. (2019), represents a breakthrough for NLP, providing deep contextual representations by pre-training on a large corpus (Devlin et al., 2019; He, 2023). Unlike static word embeddings, CNN and BERT generate dynamic contextual representations that depend on the context of the phrase, which is beneficial for domain-diverse sentiment classification. Fine-tuned BERT has the best performance on state-of-the-art results on sentiment analysis benchmarks (Kumar et al., 2025). Hybrid designs that combine BERT with CNN and BiLSTM are more successful in exploiting the benefits of context and simultaneous sequence modeling (Wijayashantha et al., 2026). But thorough

comparisons of ML, DL and BERT in varied domains with different text length and class imbalance are still understudied, which is the aim of our work. Existing studies primarily report performance improvements within individual datasets rather than providing comprehensive comparisons across heterogeneous datasets under different imbalance conditions.

Sentiment Analysis across Multiple Domains

Most previous studies focus on transfer learning and domain adaptation. However, fewer studies provide systematic comparisons of different sentiment analysis models across heterogeneous datasets using consistent experimental settings. This is especially challenging because the vocabulary, writing style and text length differ across domains, leading to considerable degradation in model performance across domains (Du et al., 2020; Kumar et al., 2025). To reduce the domain gap, domain adaptation methods such as transfer learning and adversarial training have been examined (Zou & Wang, 2025). However, most research has only considered binary classification in comparable domains and there are still no systematic comparisons across diverse domains with text of varied lengths (Alahmadi et al., 2025; Wijayashantha et al., 2026).

Class Imbalance Handling

Sentiment analysis has a frequent problem of class imbalance, where the neutral classes are often significantly fewer than the positive and negative classes (Talukder et al., 2025). This is addressed by Random OverSampling (ROS), which reproduces the minority class samples in the training set. ROS greatly enhanced BERT-LSTM performance on unbalanced COVID-19 data (Talukder et al., 2025). Algorithm-level techniques, such as class weighting, modify the loss function to prevent misclassification of the minority classes (Ganganwar & Rajalakshmi, 2024). Hybrid approaches can offer complementary advantages, notably for the neural network-based models (Fujiwara, 2024; Permataning Tyas et al., 2023).

METHODOLOGY

Research Framework Overview

The proposed research framework is shown in Figure 1. The study is composed of seven parts, for example: dataset preparation, text preprocessing, feature representation, addressing class imbalance, model construction, performance assessment, and comparison analysis. The effectiveness of ML, DL and transformer-based techniques is evaluated under distinct class imbalance scenarios using three datasets from diverse fields. Standard classification metrics are used to compare the experimental findings and to choose the best effective model across heterogeneous datasets.

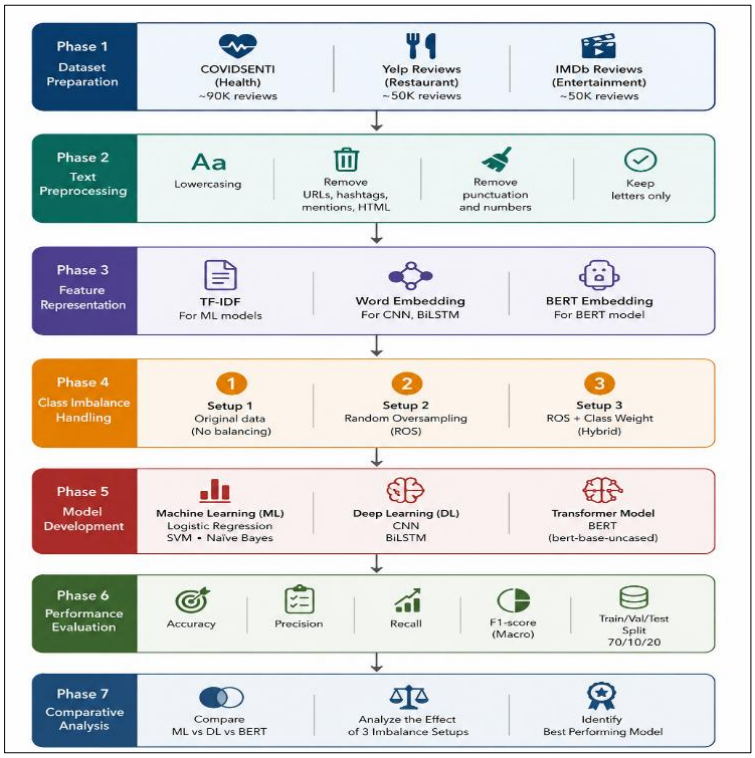

Figure 1: Research Methodology Framework

Dataset Preparation

Three publicly available datasets were selected to represent distinct domains and text length categories, enabling a comprehensive evaluation of sentiment classification performance across heterogeneous datasets. Table 1 presents the dataset characteristics.

Table 1: Dataset Characteristics

Dataset	Domain	Source	Size	Avg. Length	Sentiment Labels
COVIDSENTI	Health	(Naseem et al., 2021) – GitHub	~90,000 tweets	~20 words	Positive, Neutral, Negative
Yelp Review Full	Restaurant	(Zhang & LeCun, 2015) – Hugging Face	650,000 train + 50,000 tests	~80 words	Convert from rating 1-5★
IMDb Movie Reviews	Entertainment	(Maas et al., 2011) – Hugging Face	50,000 reviews	~250 words	VADER-based three-class labels

COVIDSENTI: The dataset consists of about 90K tweets that were gathered in February and March 2020. The tweets were carefully classified as favorable, neutral or negative. The dataset is severely biased with neutral tweets accounting for 74.9% of the total number of occurrences. The data were collected from the authors' GitHub repository.

Yelp Review Full: The dataset was downloaded from the Hugging Face repository. It comprises 650k training examples and 50k testing examples of company evaluations from the Yelp Dataset Challenge 2015. The original five-star ratings were converted into sentiment labels for this study.

IMDb Movie Reviews: The dataset from the Hugging Face repository, which contains 50,000 movie reviews. The dataset does not contain three-class sentiment labels, hence positive, neutral and negative labels were produced utilizing the VADER sentiment analyzer with a sentiment threshold of ±0.3.

Each dataset was trained and evaluated independently using its own stratified train/validation/test split. Therefore, this study performs comparative sentiment classification across heterogeneous datasets rather than source-to-target cross-domain transfer learning.

Text Preprocessing

Text preparation was performed to improve data quality and to maintain uniformity between the three datasets. In the preprocessing step, all text was converted to lowercase, the URLs, punctuation marks, special characters and unnecessary white space were removed. Where applicable, non-textual features such as emoticons and user remarks were also stripped. These techniques are applied to minimize the noise and normalize the textual input before feature extraction and model training.

Feature Extraction

The learning method adopted in this work led to the usage of several feature representation strategies. The textual data was converted into numerical vectors using the TF-IDF technique for the ML models (LR, SVM, and NB). TF-IDF applies weights to the words depending on their relevance in the document and in the corpus.

For DL models (CNN and BiLSTM), we used pre-trained word embeddings to represent words in terms of dense vector representations that capture semantic links between phrases. The transformer-based model employed contextual embeddings derived from BERT to maintain the contextual information and the meanings of words in various textual settings (He, 2023).

Class Imbalance Handling

Three experimental setups were designed based on Table 2, to evaluate the effectiveness of class imbalance handling strategies on model performance across heterogeneous datasets:

i. Setup 1 (Original dataset): Baseline performance was measured without applying any balancing techniques on the original dataset.
ii. Setup 2 (Random Oversampling - ROS): Random OverSampling was employed to duplicate existing instances to increase the number of instances in the minority class until a more balanced class distribution is reached.
iii. Setup 3 (ROS + Class Weight): This was a combination of the Random Oversampling and class weighting. To further reduce bias in favor of majority classes, the design assigned additional weights to minority classes during model training.

In the context of multi-domain sentiment analysis, the experimental settings allowed us to investigate in detail the effectiveness of the imbalance management strategies for machine learning and deep learning models.

Table 2: Class Imbalance Handling Setup Strategies

Setup	Strategy	Description	Reference
Setup 1	No Balancing	Baseline – original distribution	-
Setup 2	ROS	Random OverSampling on minority classes	(Talukder et al., 2025)
Setup 3	ROS + Class Weight	ROS + weighted loss function	(Ganganwar & Rajalakshmi, 2024)

Model Architecture

This study employed three categories of models: conventional ML, DL and transformer-based, with a total of six models, as shown in Table 3. For conventional ML baselines, we use L2 regularization penalty for LR, LinearSVC for SVM, and Multinomial classification using bigram (ngram=1-2, max=10K) TF-IDF features for NB (Mohd Sofi & Selamat, 2023). As for DL baselines, CNN adopts 3 parallel Conv1D filters of 128 filters with kernel size [3,4,5], followed by a global max pooling and dropout 0.3 (Younesi et al., 2024). The BiLSTM consists of 2 bidirectional LSTM layers with a hidden size of 128 and dropout of 0.3 to capture the bidirectional sequence dependencies(Ramaziyah & Setiawan, 2024). BERT (bert-base-uncased) is fine-tuned using a learning rate of 2e-5, AdamW optimizer, linear warmup scheduler and batch size of 16-32 for 3 epochs, as illustrated in Figure 2 (Devlin et al., 2019). The training method of all models is shown in Algorithm 1 and Algorithm 2.

Table 3: Model Architecture and Configuration

Model	Category	Feature Extraction	Description
Logistic Regression (LR)	ML Baseline	TF-IDF	Linear classifier with L2 regularization
SVM (LinearSVC)	ML Baseline	TF-IDF	Linear support vector classification
Naive Bayes (NB)	ML Baseline	TF-IDF	Multinomial Naive Bayes classifier
CNN	DL Baseline	Word Embedding	Conv1D with kernel sizes [3,4,5], 128 filters
BiLSTM	DL Baseline	Word Embedding	2-layer bidirectional LSTM, hidden_dim=128
BERT	Transformer	Contextual embeddings (fine-tuned)	Pre-trained transformer; fine-tuned with lr=2e-5, max_len=128, 3 epochs

Algorithm 1 Pseudocode: Sentiment Classification Training (ML & DL)

```
Require: Training set X_train, labels y_train, Setup ∈ {1,2,3}
Ensure:  Trained model M, Evaluation metrics

if Setup == 2 then
    X_train, y_train ← ROS(X_train, y_train)                    ▷ Oversample minority
else if Setup == 3 then
    X_train, y_train ← ROS(X_train, y_train)                    ▷ Oversample minority
    W ← compute_class_weight(y_train)                           ▷ Compute class weights
end if
f ← extract_features(X_train)                                   ▷ TF-IDF or Word Embedding
M ← train(f, y_train, loss=CrossEntropyLoss(W))                 ▷ Train model
y_pred ← M.predict(X_test)                                      ▷ Generate predictions
return evaluate(y_test, y_pred)                                 ▷ Accuracy, Precision, Recall, F1
```

Algorithm 2 Pseudocode: BERT Fine-tuning

```
Require: X_train, y_train, X_test, Setup ∈ {1,2,3} lr=2e-5, epochs=3, max_len=128/256
Ensure: Fine-tuned BERT, Evaluation metrics

if Setup == 2 or Setup == 3 then
      X_train, y_train ← ROS(X_train, y_train)                  ▷ Oversample minority
end if
W ← compute_class_weight(y_train) if Setup==3                   ▷ Compute class weights
tokens ← BertTokenizer(X_train, max_len)                        ▷ Tokenize input
for epoch = 1 to 3 do
      logits ← BERT(tokens) ▷ Forward pass
      loss ← CrossEntropyLoss(logits, y, W)                     ▷ Compute loss
      loss.backward() → optimizer.step()                        ▷ Backpropagation
      save_best_model(F1_macro)                                 ▷ Save if improved
end for
M ← load_best_model()
return evaluate(M, X_test, y_test)                              ▷ Accuracy, Precision, Recall, F1
```

Figure 2: BERT Fine-tuning Architecture for Three-Class Sentiment Classification

Evaluation Metrics

The model performance is measured by four metrics: Accuracy (1), Precision (2), Recall (3) and F1-score (4). The macro-averaged F1-score was chosen as the major assessment measure because it gives equal weight to all classes regardless of their frequencies, which is especially useful for unbalanced multi-class classification problems. To ensure uniform class distribution throughout all splits, the dataset was divided using a stratified train/validation/test split of 70/10/20.

i. $$Accuracy = \frac{TP+TN}{TP + FP + FN + TN} \quad (1)$$

ii. $$Precision = \frac{TP}{TP+FP} \quad (2)$$

iii. $$Recall = \frac{TP}{TP+FN} \quad (3)$$

iv. $$F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision+Recall} \quad (4)$$

given by,

TP: True Positives, TN: True Negatives, FP: False Positives, FN: False Negatives.

RESULTS AND DISCUSSIONs

Machine Learning Baseline Results

Table 4 presents the F1-score (macro) results of ML baseline models on three datasets and three experimental setups. Figure 3 also shows a visualization of the effect of class imbalance handling approaches across domains, presenting the comparative performance of ML models.

Table 4: ML Baseline F1-Score (Macro) Results

Dataset	Model	Setup 1 (No Balance)	Setup 2 (ROS)	Setup 3 (ROS+CW)	Best F1
COVIDSENTI	LR	74.65%	83.48%	83.48%	**83.48% (S2)**
COVIDSENTI	SVM	81.93%	82.11%	82.11%	82.11% (S2)
COVIDSENTI	NB	51.71%	62.23%	62.23%	62.23% (S2)
Yelp	LR	71.44%	72.45%	72.45%	72.45% (S2)
Yelp	SVM	70.34%	70.16%	70.16%	70.34% (S1)
Yelp	NB	64.46%	68.53%	68.53%	68.53% (S2)
IMDb	LR	49.73%	52.68%	52.68%	52.68% (S2)
IMDb	SVM	50.52%	50.91%	50.91%	50.91% (S2)
IMDb	NB	44.97%	51.85%	51.85%	51.85% (S2)

Figure 3 and Table 4 indicate that ROS (Setup 2) increased F1-scores for unbalanced datasets with COVIDSENTI showing the most improvement with +8.83% in LR (74.65% → 83.48%) which is similar with the results of Talukder et al. (2025). Interestingly, Setup 2 and Setup 3 yielded the same results for

all ML models, since class weighting becomes uniform after ROS balances the training data, suggesting that ROS alone is adequate for ML-based sentiment classification. The lowest F1-scores (44.97% - 52.68%) were attained by IMDb because of VADER-generated noise in the neutral label and the limits of TF-IDF in modeling complex semantic linkages in lengthy texts (~250 words) which supports the use of contextual embedding techniques.

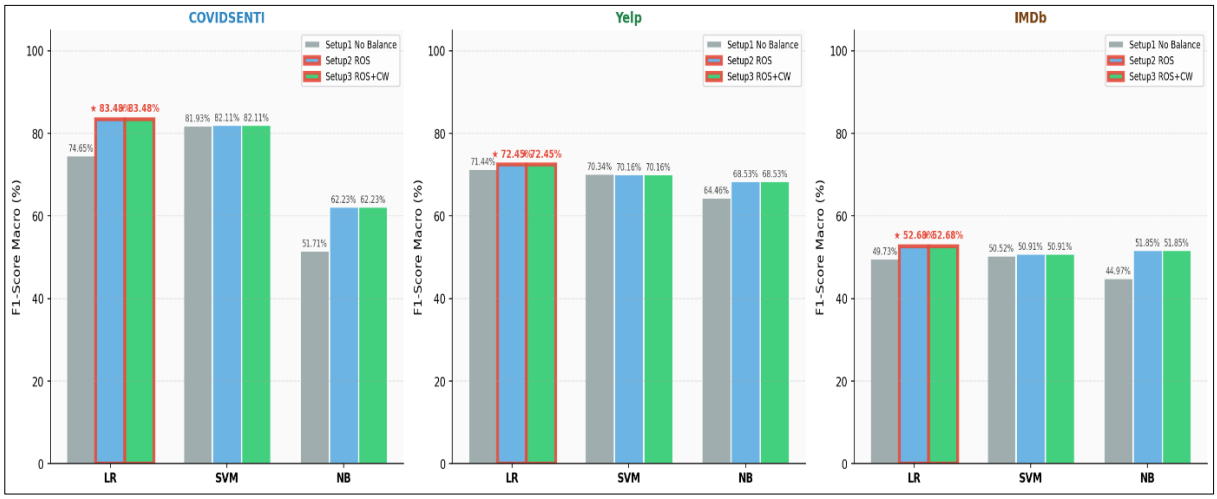

Figure 3: ML Baseline F1-Score (Macro) Comparison across Three Datasets and Three Experimental Setups

Deep Learning Results

Table 5 shows that DL models consistently outperform ML baselines in all three datasets, where BiLSTM performs best with an overall performance of 89.17% F1-score on COVIDSENTI (Setup 3), 62.79% on Yelp (Setup 2), and 57.40% on IMDb (Setup 2). CNN, as predicted, performed well but was always lagging behind BiLSTM as sequential modeling is more appropriate for sentiment-bearing text than local n-gram feature extraction. This is due to noise in the neutral labels that were obtained from VADER, and the inability of static word embeddings to capture long-range semantic dependencies in longer texts. This also serves as empirical motivation for the following BERT experiments.

Table 5: DL Baseline (CNN + BiLSTM) F1-Score (Macro) Results

Dataset	Model	Setup 1 (No Balance)	Setup 2 (ROS)	Setup 3 (ROS+CW)	Best F1
COVIDSENTI	CNN	85.37%	86.54%	86.31%	86.54% (S2)
COVIDSENTI	BiLSTM	88.59%	88.76%	89.17%	**89.17% (S3)**
Yelp	CNN	61.10%	60.88%	61.62%	61.62% (S2)
Yelp	BiLSTM	61.88%	62.79%	61.49%	62.79% (S2)
IMDb	CNN	49.16%	52.25%	51.97%	52.25% (S2)
IMDb	BiLSTM	55.40%	57.40%	55.90%	57.40% (S2)

BERT Fine-tuning Results

The BERT fine-tuning findings on three datasets and three experimental settings are shown in Table 6. The F1-score comparison of DL and BERT models is shown in Figure 4. BERT showed significant improvement over the ML and DL baselines in all areas. The largest improvement was recorded on

IMDb where BERT showed a +18.36% increase over BiLSTM (57.40% → 75.76%), demonstrating the important benefit of contextual embeddings for long-form content categorization. As can be seen from Figure 4, there were different optimal imbalance management options for the different datasets. Setup 2 performed best on IMDb (75.76%) whereas Setup 3 performed best on COVIDSENTI (95.32%) and Yelp Reviews (73.79%) in terms of F1-score. Fluctuations indicate the degree of imbalance and that the text length characteristics of each domain influence the effectiveness of class weighting and this effectiveness varies between datasets.

Table 6: BERT Fine-tuning F1-Score (Macro) and Accuracy Results

Dataset	Setup 1 F1 / Acc	Setup 2 F1 / Acc	Setup 3 F1 / Acc	Best F1	Best Acc
COVIDSENTI	94.85% / 96.87%	95.13% / 96.96%	95.32% / 97.10%	95.32% (S3)	97.10% (S3)
Yelp	73.66% / 78.16%	73.63% / 77.84%	73.79% / 77.83%	73.79% (S3)	78.16% (S1)
IMDb	74.71% / 84.77%	75.76% / 84.15%	75.45% / 83.92%	75.76% (S2)	84.77% (S1)

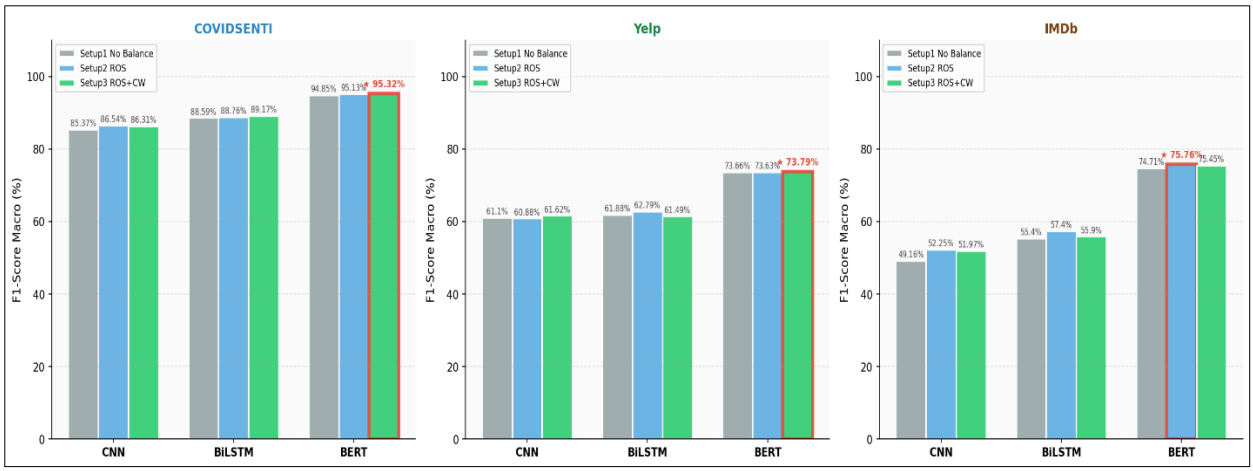

Figure 4: DL and BERT F1-Score Comparison across Three Datasets and Three Setups

Comparative Analysis – All Models

Table 7: Best F1-Score Summary — All Models across 3 Datasets

Model	Category	COVIDSENTI	Yelp	IMDb	Overall Avg.	Rank
LR	ML	83.48%	72.45%	52.68%	69.54%	3rd
SVM	ML	82.11%	70.34%	50.91%	67.79%	4th
NB	ML	62.23%	68.53%	51.85%	60.87%	6th
CNN	DL	86.54%	61.62%	52.25%	66.80%	5th
BiLSTM	DL	89.17%	62.79%	57.40%	69.79%	2nd
BERT	Transformer	95.32%	73.79%	75.76%	**81.62%**	1st ★

Table 7 summarizes the best F1-score of each model on all three datasets while Figure 5 shows the performance progression from ML to DL to BERT. BERT earned the overall highest average of 81.62% F1-score, significantly beating DL (BiLSTM with 69.79%) and ML (LR with 69.54%) techniques. The

total performance ranking is BERT > BiLSTM ≈ LR > SVM > CNN > NB. As shown in Figure 5, longer texts such as IMDb increase the performance difference significantly, supporting the effectiveness of contextual embeddings for sentiment classification across heterogeneous datasets with different text lengths.

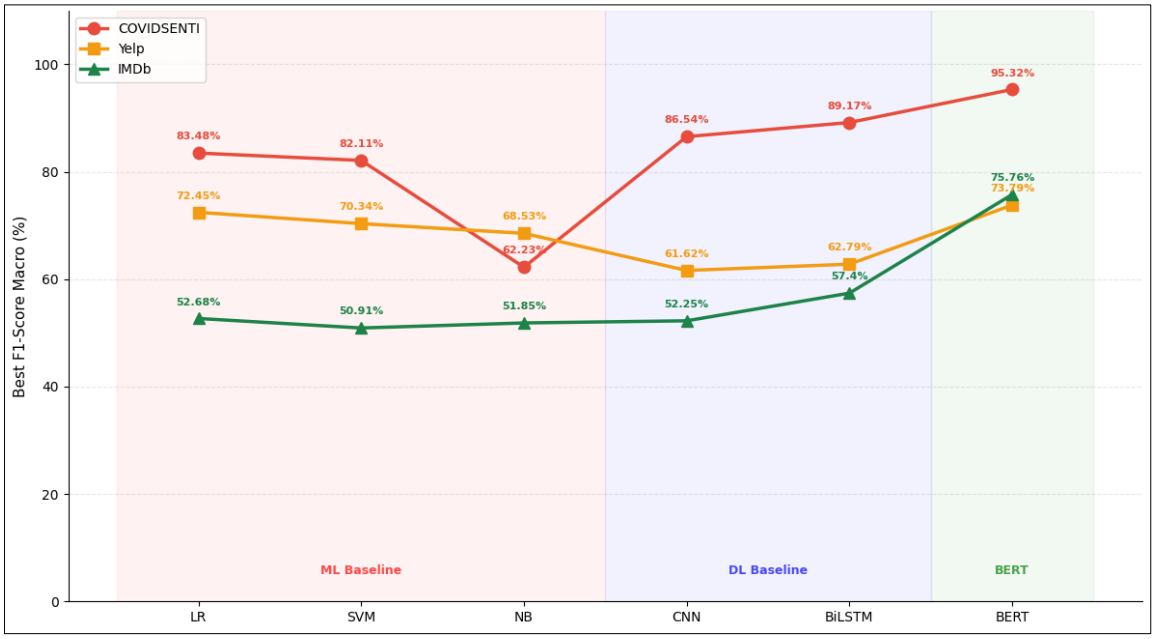

Figure 5: Model Performance Progression: ML → DL → BERT

Effect of Class Imbalance Handling

Figure 5 and Tables 5, 6 and 7 provide three important observations from the comparison of class imbalance strategies. First, for extremely unbalanced COVIDSENTI, ROS significantly enhanced the performance, with LR gaining +8.83%. Second, ML models produced the same results for Setup 2 and Setup 3 which indicates that ROS alone is enough for ML-based sentiment classification. Third, for DL models, Setup 3 exhibited a slight additional benefit on COVIDSENTI, with BiLSTM achieving +0.41% over Setup 2, indicating that algorithm-level class weighting supports data-level balancing specifically for gradient-based neural networks, a key novel finding of this study.

CONCLUSION

This paper presented a comprehensive comparative study of ML, DL, and transformer-based approaches for multiple-domain sentiment analysis on three heterogeneous datasets with different text lengths. The main results are:

i. ROS achieved large performance improvements for severely unbalanced datasets COVIDSENTI increase 8.83% for LR, but minimal gains for moderately imbalanced datasets Yelp increase 1.01%.
ii. For the standard ML models, data balancing using ROS is enough for class imbalance handling, as the class weighting does not bring any further improvement after data balancing. However, DL models benefit little from the combined technique, with BiLSTM scoring +0.41% on COVIDSENTI in Setup 3.
iii. BERT showed a consistent improvement over ML and DL baselines across all domains with F1-scores of 95.32%, 73.79% and 75.76% on COVIDSENTI, Yelp and IMDb respectively, with most improvements on lengthier texts, an increase +18.36% over BiLSTM on IMDb. The length of text is a very important factor in the hardness of sentiment classification across

heterogeneous datasets. For long texts, contextual embeddings have much larger benefits over TF-IDF and static word embedding methods.

The results show that pre-trained transformer models consistently outperform ML and DL baselines across heterogeneous sentiment datasets with different text-length characteristics. Future work will extend this study to true cross-domain sentiment classification by training models on one source domain and evaluating them on unseen target domains. Comparisons of binary vs. three-class classification across domains will also be made to quantify the influence of the presence of the neutral class on the performance across domains.

ACKNOWLEDGEMENT

This publication forms part of the ongoing doctoral research conducted at the Malaysia-Japan International Institute of Technology (MJIIT), Universiti Teknologi Malaysia (UTM). The authors would like to express their sincere appreciation to Universiti Islam Selangor (UIS) for providing financial support for the publication of this manuscript.

REFERENCES

Abdullah, T., & Ahmet, A. (2023). Deep Learning in Sentiment Analysis: Recent Architectures. *ACM Computing Surveys*, *55*(8). https://doi.org/10.1145/3548772

Agrawal, R., Majumder, M., Yadav, I., Taneja, N., Hamdare, S., & Hemnani, P. (2025). Evaluating sentiment analysis models: A comparative analysis of vaccination tweets during the COVID-19 phase leveraging DistilBERT for enhanced insights. *MethodsX*, *14*. https://doi.org/10.1016/j.mex.2025.103407

Alahmadi, K., Alharbi, S., Chen, J., & Wang, X. (2025). Generalizing sentiment analysis: a review of progress, challenges, and emerging directions. In *Social Network Analysis and Mining* (Vol. 15, Number 1). Springer. https://doi.org/10.1007/s13278-025-01461-8

Bordoloi, M., & Biswas, S. K. (2023). Sentiment analysis: A survey on design framework, applications and future scopes. *Artificial Intelligence Review*, *56*(11), 12505–12560. https://doi.org/10.1007/s10462-023-10442-2

Devlin, J., Chang, M.-W., Lee, K., & Kristina Toutanova. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. https://doi.org/10.18653/v1/N19-1423

Du, Y., He, M., Wang, L., & Zhang, H. (2020). Wasserstein based transfer network for cross-domain sentiment classification. *Knowledge-Based Systems*, *204*. https://doi.org/10.1016/j.knosys.2020.106162

Fujiwara, K. (2024). Knowledge distillation with resampling for imbalanced data classification: Enhancing predictive performance and explainability stability. *Results in Engineering*, *24*. https://doi.org/10.1016/j.rineng.2024.103406

Ganganwar, V., & Rajalakshmi, R. (2024). Employing synthetic data for addressing the class imbalance in aspect-based sentiment classification. *Journal of Information and Telecommunication*, *8*(2), 167–188. https://doi.org/10.1080/24751839.2023.2270824

He, Y. (2023). BERT-CNN-BiLSTM: A Hybrid Deep Learning Model for Accurate Sentiment Analysis. *2023 IEEE 5th International Conference on Power, Intelligent Computing and Systems, ICPICS 2023*, 921–926. https://doi.org/10.1109/ICPICS58376.2023.10235335

Kumar, A., Abhishek, K., & Shafeeq B M, A. (2025). SentXFormer: a transformer-enhanced hybrid deep learning framework for cross-domain sentiment analysis of customer reviews. *Scientific Reports*. https://doi.org/10.1038/s41598-025-33526-1

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011). Learning Word Vectors for Sentiment Analysis. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, 142–150.

Mao, Y., Liu, Q., & Zhang, Y. (2024). Sentiment analysis methods, applications, and challenges: A systematic literature review. In *Journal of King Saud University - Computer and Information Sciences* (Vol. 36, Number 4). King Saud bin Abdulaziz University. https://doi.org/10.1016/j.jksuci.2024.102048

Mohd Sofi, S., & Selamat, A. (2023). Aspect Based Sentiment Analysis: Feature Extraction using Latent Dirichlet Allocation (LDA) and Term Frequency - Inverse Document Frequency (TF-IDF) in Machine Learning (ML). *Malaysian Journal of Information and Communication Technology (MyJICT)*, 169–179. https://doi.org/10.53840/myjict8-2-102

Mohd Sofi, S., Selamat, A., Alwani Shaffiei, Z., Kuala Lumpur, M., Sultan Yahya Petra, J., & Lumpur, K. (2026). An Adaptive Ensemble Machine Learning Classifier for Sentiment Analysis on Twitter. *Journal of Advanced Research Design Journal Homepage: Journal of Advanced Research Design*, *136*, 340–357. https://doi.org/10.37934/ard.136.1.340357

Naseem, U., Razzak, I., Khushi, M., Eklund, P. W., & Kim, J. (2021). COVIDSenti: A Large-Scale Benchmark Twitter Data Set for COVID-19 Sentiment Analysis. *IEEE Transactions on Computational Social Systems*, *8*(4), 976–988. https://doi.org/10.1109/TCSS.2021.3051189

Permataning Tyas, S. M., Sarno, R., Haryono, A. T., & Rossa Sungkono, K. (2023). A Robustly Optimized BERT using Random Oversampling for Analyzing Imbalanced Stock News Sentiment Data. *ICCoSITE 2023 - International Conference on Computer Science, Information Technology and Engineering: Digital Transformation Strategy in Facing the VUCA and TUNA Era*, 897–902. https://doi.org/10.1109/ICCoSITE57641.2023.10127725

Ramaziyah, Y. A., & Setiawan, E. B. (2024). Hybrid Deep Learning CNN and BiLSTM with FastText as Feature Expansion for Sentiment Analysis in President Election 2024. *COMNETSAT 2024 - IEEE International Conference on Communication, Networks and Satellite*, 176–183. https://doi.org/10.1109/COMNETSAT63286.2024.10862946

Siino, M., Tinnirello, I., & La Cascia, M. (2024). Is text preprocessing still worth the time? A comparative survey on the influence of popular preprocessing methods on Transformers and traditional classifiers. *Information Systems*, *121*. https://doi.org/10.1016/j.is.2023.102342

Talukder, M. A., Uddin, M. A., Roy, S., Ghose, P., Sarker, S., Khraisat, A., Kazi, M., Rahman, M. M., & Hakimi, M. (2025). A hybrid deep learning model for sentiment analysis of COVID-19 tweets with class balancing. *Scientific Reports*, *15*(1). https://doi.org/10.1038/s41598-025-97778-7

Tsirmpas, D., Gkionis, I., Papadopoulos, G. T., & Mademlis, I. (2024). Neural natural language processing for long texts: A survey on classification and summarization. In *Engineering Applications of Artificial Intelligence* (Vol. 133). Elsevier Ltd. https://doi.org/10.1016/j.engappai.2024.108231

Wijayashantha, A. K. S., Samarasinghe, U. S., Perera, H. A. D. U., & Subhashini, L. D. C. S. (2026). An Attention Enhanced CNN BERT Model with TF IDF Feature Fusion for Cross Domain Sentiment Analysis. *2026 6th International Conference on Advanced Research in Computing: Responsible AGI: Balancing Intelligence, Responsibility and Sustainability, ICARC 2026 - Conference Proceedings*. https://doi.org/10.1109/ICARC68737.2026.11453951

Yang, S., Xing, J., Dong, Z., & Liu, Z. (2025). Heterogeneous Ensemble Sentiment Classification Model Integrating Multi-View Features and Dynamic Weighting. *Electronics (Switzerland)*, *14*(21). https://doi.org/10.3390/electronics14214189

Younesi, R. T., Tanha, J., Namvar, S., & Mostafaei, S. H. (2024). A CNN-BiLSTM based deep learning model to sentiment analysis. *2024 20th CSI International Symposium on Artificial Intelligence and Signal Processing, AISP 2024*. https://doi.org/10.1109/AISP61396.2024.10475311

Zhang, X., & LeCun, Y. (2015). Text Understanding from Scratch. *Advances in Neural Information Processing Systems 28 (NIPS 2015)*. http://arxiv.org/abs/1502.01710

Zou, H., & Wang, Y. (2025). Large language model augmented syntax-aware domain adaptation method for aspect-based sentiment analysis. *Neurocomputing*, *625*. https://doi.org/10.1016/j.neucom.2025.129472