

Class-Aware Semantic Hybrid Data Augmentation for Imbalanced Sentiment Classification Using Multiple Transformer Models

Prashant Upadhyaya

Chandigarh University

G.L. Saini

`gaindi.saini@jaipur.manipal.edu`

Manipal University Jaipur

Atul Makrariya

Bundelkhand University

Research Article

Keywords: CSHDA, Class Imbalance, Sentiment Classification, Transformer-based models

Posted Date: August 27th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10437610/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Class-Aware Semantic Hybrid Data Augmentation for Imbalanced Sentiment Classification Using Multiple Transformer Models

Prashant Upadhyaya^{1*}, G.L. Saini^{2*} and Atul Makrariya³

^{1*}Department of ECE, University Institute of Engineering Chandigarh
University, Gharuan, Mohali, 140413, Punjab, India.

²Department of IoT and Intelligent Systems, Manipal University,
Jaipur, 303007, Rajasthan, India.

³Department of ECE, Institute of Engineering & Technology,
Bundelkhand University, Jhansi, 284003, Uttar Pradesh, India.

*Corresponding author(s). E-mail(s): upadhyaya.prashant@gmail.com;
gaindi.saini@jaipur.manipal.edu;
Contributing authors: makrariya@rediffmail.com;

Abstract

Class imbalance remains one of the major challenges in sentiment classification because transformer-based models are often biased toward majority classes, leading to poor recognition of minority sentiments. Although existing data augmentation techniques increase the number of training samples, they generally apply identical transformations across all classes without considering class-specific characteristics, which may introduce semantic drift, redundant samples, and limited improvements in minority-class learning. To address these limitations, this paper proposes a Class-Aware Semantic Hybrid Data Augmentation (CSHDA) framework for imbalanced sentiment classification. The proposed framework combines class-aware augmentation probabilities with four complementary augmentation techniques, namely BERT-based paraphrasing, synonym replacement, random insertion, and random deletion. To ensure the quality of generated samples, semantic consistency is validated using Sentence-BERT cosine similarity, while duplicate-aware filtering removes redundant augmented instances. The framework augments only minority classes, thereby preserving the original class distribution characteristics while reducing learning bias. The proposed framework was evaluated using BERT, RoBERTa, DistilBERT, and ALBERT

transformer-based architectures on two benchmark datasets with different characteristics: the CUFC corpus and the Twitter Airline dataset. Experimental results demonstrate consistent improvements across both datasets. On the CUFC dataset, Macro F1-score improved from 60.80% to 81.10% for BERT and from 55.43% to 80.01% for DistilBERT. On the Twitter dataset, BERT improved from 60.65% to 69.82%, while DistilBERT increased from 42.98% to 71.21%. The proposed framework also substantially improved minority-class recall, particularly for the neutral sentiment class, while paired statistical analysis confirmed that the observed improvements were statistically significant. These results demonstrate that the proposed CSHDA framework effectively mitigates class imbalance by generating semantically consistent and diverse minority-class samples, leading to improved robustness and cross-dataset generalization of transformer-based sentiment classification models.

Keywords: CSHDA, Class Imbalance, Sentiment Classification, Transformer-based models

1 Introduction

The concept of Sentiment Analysis (SA) has a long history within the framework of Natural Language Processing (NLP) [1, 2]. It is a technique which offers a way to extract and understand the opinions expressed in text [3, 4] by classifying text inputs in terms of sentiment categories, which are positive, negative or neutral [5, 6]. Therefore, being an efficient tool to derive meaningful insights from data, enabling usage of it for E-commerce domain, like Amazon and Flipkart [7–9], food delivery [10, 11], identification of learning difficulties in industry, and enhanced student engagement [12, 13]. Also, it consists of review of the product websites [2, 14], educational discussion sites [15, 16], social networking sites [1, 5] and the biomedical domain [17]. However, with the increase in the amount of data and the developments of new technologies [5, 6], their expansions have seen an increase for the analysis of huge volumes of texts. Therefore, requirement of the data driven decisions has to be implemented which is very important both for research and industry sectors to understand and interpreted the user generated content. Moreover, trend monitoring of sentiment analysis can help the organization to forecast the demands of the user, guide their policy, and make plans accordingly [2, 7, 14].

Sentiment analysis mainly uses four main approaches [18, 19], namely, lexicon-based, machine learning (ML), deep learning (DL) based, and their hybrid approach. Each of these approaches brings its own set of advantages in terms of different kinds of datasets and applications [8, 9]. One of these approaches is the lexicon-based approach which uses sentiment lexicons that contain lists of words associated with predefined polarities. In sentiment analysis, the process takes place by determining the sentiment in the texts through summing up the scores of words used in the texts. Since no labeling process is needed for the method, the approach becomes very useful in cases

of unavailability of labeled data or when such labeling is costly. However, the lexicon-based method faces problems when it comes to dealing with context such as sarcasm, idiomatic language, technical terms, and emotional expressions [19, 20].

Classical machine learning techniques like Naive-Bayes (NB) classifiers i.e., to be adopted because of its simple structure and efficiency to deal with high dimensional textual information [21, 22]. The major strength of NB is the probabilistic approach, making it a good fit for dealing with the small size of training sets. On the other hand, Support Vector Machines (SVM) [22], provide the optimal way to utilize margins between different sentiment categories due to their capability of constructing discriminating hyperplanes. Decision Trees (DT) [4, 23] give rule-based and hierarchical structure that explains the effect of specific words or phrases on the final sentiment classification. Despite possible over-fitting in case of using high dimensional feature spaces, they were quite useful in revealing patterns within data and the impact of certain linguistic features. Based on that, the idea of ensemble learning with techniques like Random Forests (RF) and Gradient Boosting (GB) [24] arose, improving robustness by combining several simpler classifiers. However, despite establishing some base in the development of sentiment analysis, those methods had difficulties understanding the context and required much feature engineering. This motivated the transition to ML, DL and Transformer Models (TM) [2, 7, 25–27].

The latest technology that makes use of Bidirectional Encoder Representations from Transformers (BERT) has helped in enhancing this improvement through allowing the model to pay attention to the contextual relationships [28]. In this case, the architectural analyses not only include local but also the whole text sequences. Moreover, hybrid sentiment analysis is one of the other approaches that aims at taking advantage of lexicon based approaches along with machine learning and deep learning approaches in order to enhance sentiment analysis through hybrid approach or architecture.

Though, remarkable improvement has been reported in the field of sentiment analysis, however, the biggest problem lies in dealing with short corpora. In this the main problems are related to the specific vocabulary and the language ambiguity. Moreover, imbalance in the dataset is another crucial problem, due to which the accuracy varies depending on the imbalance of sentiments. Data augmentation becomes a practical method to increase the size of the training datasets and improve model generalization [29]. Hence, it is the proper method to work with real world tasks or small datasets. Data augmentation has contributed significantly to NLP systems since it has improved the reliability of the system in terms of increasing data variability and richness even with the small amount of labeled data [30, 31].

New data samples can be generated by making slight changes to existing samples in the dataset using basic operations like synonym substitution, word insertion, deletion, or paraphrasing. This will help improve the robustness and generalization of the model. It is great because it will not only increase diversity among the labeled data samples but also avoid overfitting. Thus, this research paper highlights the findings of the experiments conducted by analyzing the problem of data imbalance and evaluating the efficiency of using state-of-the-art augmentation methods to improve model performance. While various augmentation strategies including EDA, back-translation, and

contextual paraphrasing were proven to be effective in NLP tasks, current techniques lack three main drawbacks when used for handling imbalanced sentiment datasets:

1. Inadequate preservation of semantic consistency,
2. Generation of redundant or noisy samples, and
3. Lack of class-aware augmentation control for minority sentiment categories,

Since, most existing augmentation frameworks apply transformations uniformly across classes without considering sentiment sensitivity, which may distort contextual meaning and reduce minority-class reliability. Therefore, there remains a need for a semantically controlled and class-aware augmentation framework capable of improving minority-class learning while preserving contextual fidelity. The major contributions of this work are summarized as follows:

1. A class-aware semantic hybrid augmentation framework is proposed for imbalanced sentiment classification, where augmentation probabilities are dynamically assigned according to minority sentiment characteristics.
2. Unlike existing EDA-based methods, the proposed CSHDA simultaneously
 - adapts augmentation probabilities according to minority class characteristics,
 - validates semantic consistency using SBERT,
 - removes duplicate generated samples,
 - performs class-aware augmentation only for minority classes.
3. A semantic consistency mechanism using SBERT similarity thresholds is introduced to preserve contextual meaning while improving sentence diversity and preventing semantic drift.

Further, for evaluating the performance, we have selected the two corpus i.e., the CUFC corpus, [23] and Twitted dataset [32] for cross data generalization using four transformer architectures under multiple random seeds. The rest of this paper is structured as follows. Section II provides a survey of the relevant literature on sentiment classification and data augmentation approaches. Section III introduces the proposed CSHDA approach. Section IV describes the experiments conducted, and reports on the results of the experiments. Section V analyzes the statistical significance and computational efficiency of the experiments conducted. Section VI discusses the results of the experiments. Finally, Section VII concludes the paper.

2 Literature Survey

In the past few years, there has been a tremendous advancement in the field of sentiment analysis. The major reason behind such advancements is not only the increased amount of user-generated content but also because of advancements in natural language processing. In past research studies, the methodology used was mainly based on the lexicon-based approach with traditional ML algorithms. Thus, the deep learning as well as transformer model approaches were adopted by the researcher due to their improved ability to deal with complex textual structures.

The work proposed by Pang and Lee [4] provides a overview of approaches for identifying and analyzing information in textual data while clearly distinguishing opinion-oriented information retrieval from conventional fact-based retrieval. The survey discusses key tasks, including subjectivity classification and sentiment polarity detection, and highlights several open research challenges. It includes the lack of quality labeled data sets and the challenges associated with cross-domain adaptation of SA models. Another point made by the authors is the intricacy of the connection between the topic content and sentiment expressions, as well as ethical and economic aspects of using opinion-aware technologies.

The [33] author suggested performing the sentiment analysis of the Twitter datasets through distant supervision approach. Different ML models like NB, ME, and SVM were used and they obtained accuracy over 80% for classifying positive and negative sentiments. The author further compared the use of unigrams and bigrams with other features usually used in the traditional review datasets and highlighted difficulties in handling the microblogs, which include shortness of message, informality and change in topics at a fast pace. The author further demonstrated that some simple pre-processing steps, including the normalization of usernames, URLs, and repetition of characters, along with unigrams and bigrams helped improve classification.

The document modelling for classification using IMDB and Yelp Dataset using sentiments neural network model has been discussed in [34]. In this approach classification was done using the Bi GatedNN Avg. The suggested model uses CNN/LSTM approach for capturing the relations and sequences of sentences. The suggested approach was able to achieve the accuracy of 67.6% and 45.3% on Yelp 2015 and IMDB datasets respectively. Another study that discusses the classification of sentiment of a balanced data set of movies review with the use of NB and SVM classifier has been published in [22]. Textual data has been preprocessed and vectorized with the use of CountVectorizer and TF-IDF approaches before training. The performance of both models was tested with the 10-fold CV (cross validation) technique and achieve accuracy of 94% while NB got 89.5%.

In the year between 2018-2020 was the Transformer Era in which transformer model architecture brought a breakthrough in the field of NLP. BERT model and its variations like RoBERTa, XLNet, and ALBERT have shown better performance than CNN and RNN models for sentiment classification [35–38]. Contextual embedding was achieved by means of self-attention mechanism in these models. Hence, there is no need for sequence dependency/recurrence. In the experiment using BERT, new benchmarks were set by showing improvements in 11 tasks that include a 7.7-point increase in GLUE benchmark, improvements in SQuAD benchmark ($F1 = 93.2$ on SQuAD1.1 Test), and significant improvements in sentence-level and token-level benchmark tasks.

On the other hand, for the small corpus size the Easy Data Augmentation (EDA) method is most widely used on the original dataset for the generation of secondary forms of a sentence while preserving the label of the primary sentence [39]. The most popular EDA methods are the Synonym Replacement (SR), Random Insertion (RI), Random Swapping (RS) and Random Deletion (RD) [39]. As stated in [40], it may become difficult to predict the sentence correctly as positive or neutral due to class imbalance. Thus, they have used data augmentation and showed the improvement

of 81% in F1 score on the binary classification task using the RoBERTa model on SemEval-2022 Task 4. Hence, it has been experimentally shown that the data augmentation helps to significantly improve the transformers. A further research paper built on EDA framework indicates the usefulness of data augmentation technique for handling limited data scenarios [41]. The data augmentation techniques used by the author are of two types, that is, token level and sentence level. These include substitution, random insertion, paraphrasing, and round-trip translation. From the experimental findings, it was found that the former was most useful in a supervised setting while the latter was useful in a semi-supervised setting.

3 Methodology

This section presents the proposed Class-Aware Semantic Hybrid Data Augmentation (CSHDA) framework for addressing class imbalance in sentiment classification. The overall methodology consists of datasets used for this research, problem formulation, dataset preprocessing and minority-class augmentation techniques.

3.1 Dataset Description

For this research study, two imbalanced datasets are selected i.e., benchmark Twitter Airline Sentiment dataset [32] and CUFC (Chandigarh University food courts) corpus [23] consists of user feedback about food items. The objective is to evaluate the robustness, cross-domain generalization, and adaptability of the proposed augmentation framework across different imbalanced sentiment datasets. “Table. 1” shows the datasets statistics of the CUFC corpus and Twitter datasets. From the given dataset it can be seen that for CUFC corpus the positive sample are the majority whereas for the twitter the negative sample were majority. From the given imbalance ratio for both datasets, it can be seen that the dataset is highly imbalance. Therefore data augmentation is required to handle this class imbalance problem. Hence, the proposed framework is designed to generate semantically consistent augmented samples exclusively for minority sentiment classes from the original training datasets while preserving the integrity of the validation and testing datasets.

Table 1: Statistical dataset analysis of CUFC and Twitter corpora.

Statistic	CUFC Corpus	Twitter Dataset
Total reviews/tweets	1,174	14206
Positive samples	769 (65.50%) (Majority)	2192 (15.43%)
Neutral samples	222 (18.91%)	2951 (20.77%)
Negative samples	183 (15.59%)	9063 (63.80%) (Majority)
Imbalance Ratio (Minority/Majority)	0.238 (Severe)	0.242 (Severe)

3.2 Dataset Preprocessing Pipeline

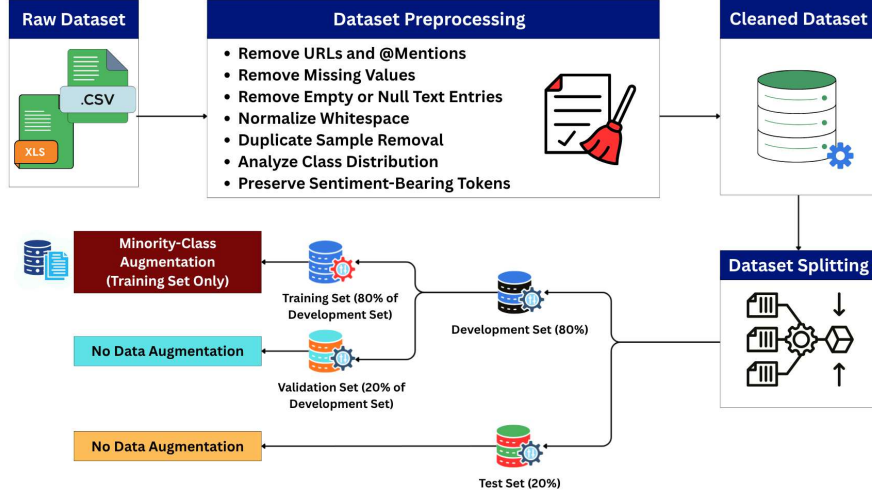


Fig. 1: Overview of the data preprocessing, sentiment-preserving cleaning, and dataset splitting pipeline used in the proposed framework.

The completed data pre-processing and data splitting pipeline is shown in “Fig. 1”. In this case, the initial steps of data cleansing include removal of any special characters and white space except for punctuation relevant to sentiment, such as exclamation marks, question marks, and full stops. This is followed by the normalization of text to lower case as well as removal of URLs and user mentions from the data.

Further, it also detect the duplicate sentences to produce a high-quality cleaned datasets. Finally, the splitting of the data is performed on cleaned datasets using stratified sampling to preserve class distribution. Initially, the data was split into development set and test set using the 80:20 ratio. Further, the development set was split into training set (80% of development set) and validation set (20% of development set) as shown in “Fig. 1”. The data augmentation is applied on the training subsets (on minority class only to prevent information leakage), while the validation and test subsets remain unaltered to ensure unbiased evaluation.

3.3 Proposed Class-Aware Semantic Hybrid Data Augmentation Framework

As shown in Table 1, both the CUFC corpus and the Twitter Airline dataset exhibit severe class imbalance, where one sentiment class significantly outnumbers the remaining classes. Such imbalance often biases transformer-based classifiers toward the majority class, resulting in reduced minority-class recall and macro-level performance. To address this limitation, a Class-Aware Semantic Hybrid Data Augmentation

(CSHDA) framework is proposed. The objective of the proposed framework is to generate diverse yet semantically consistent training samples exclusively for the minority classes while preserving the original sentiment polarity. Unlike conventional Easy Data Augmentation (EDA) techniques, the proposed framework combines rule-based perturbation methods with transformer-based paraphrasing, semantic similarity validation, and duplicate-aware filtering within a unified augmentation pipeline.

3.3.1 Problem Formulation

Sentiment classification aims to learn a mapping function that assigns a sentiment label to an input text sample. Let the original training dataset be represented as

$$D = \{(s_i, y_i)\}_{i=1}^N, \quad (1)$$

where s_i denotes the i^{th} preprocessed text sample and $y_i \in \mathcal{C}$ is its corresponding sentiment label, with $\mathcal{C} = \{\text{positive}, \text{neutral}, \text{negative}\}$. The dataset is assumed to be class-imbalanced, where the number of samples belonging to the majority class is significantly higher than those of one or more minority classes.

For each class $c \in \mathcal{C}$, class-specific subset is obtained using

$$D_c = \{s_i \mid y_i = c\}, \quad n_c = |D_c| \quad (2)$$

To obtain a balanced dataset, augmentation is performed only on the minority classes to achieve a balanced class distribution.

$$T = \max_{c \in \mathcal{C}} n_c. \quad (3)$$

Accordingly, augmented samples required for each minority class is computed as

$$k_c = T - n_c. \quad (4)$$

Augmentation is applied only to classes where $k_c > 0$.

Most existing data augmentation techniques apply identical transformations across all sentiment classes without considering class imbalance or sentiment-specific characteristics. Such approaches may introduce redundant samples, semantic drift, or class-specific noise, thereby limiting their effectiveness for minority-class learning.

To address these limitations, this work proposes a Class-Aware Semantic Hybrid Data Augmentation (CSHDA) framework. Given a minority-class sample s_i , the proposed augmentation operator $\mathcal{A}(\cdot)$ generates a candidate sample according to the class-aware probability distribution P_c ,

$$s'_i = \mathcal{A}(s_i, P_c), \quad (5)$$

where P_c determines the probability of selecting a particular augmentation technique for sentiment class c .

For each minority class, the probability vector is given as

$$P_c = \{p_{BP}, p_{SR}, p_{RI}, p_{RD}\}, \quad (6)$$

where p_{BP} , p_{SR} , p_{RI} , and p_{RD} denote the probabilities of selecting BERT-based paraphrasing, synonym replacement, random insertion, and random deletion, respectively. The values of P_c are determined empirically according to the characteristics of the dataset and the minority sentiment class.

To ensure semantic consistency, each generated sample is validated using Sentence-BERT (SBERT) cosine similarity. Let $\phi(\cdot)$ denote the SBERT embedding function that maps an input sentence into a dense semantic vector. Furthermore, let s denote a sample selected from the minority-class subset D_c . The semantic similarity between s and s' is computed as

$$\text{sim}(s, s') = \frac{\phi(s) \cdot \phi(s')}{\|\phi(s)\| \|\phi(s')\|}. \quad (7)$$

where $\text{sim}(\cdot)$ denotes the cosine similarity computed using SBERT embeddings. The generated sentence is accepted as a valid augmented sample only if

$$\tau_{\min} \leq \text{sim}(s, s') \leq \tau_{\max}. \quad (8)$$

The similarity thresholds $\tau_{\min} = 0.65$ and $\tau_{\max} = 0.97$ were empirically determined through pilot experiments to achieve an appropriate balance between semantic preservation and sentence diversity. This constraint filters out augmentations that either deviate significantly from the original semantic context or are nearly identical to the source sentence. Candidate sentences that do not satisfy the semantic similarity criterion or are identified as duplicates are discarded. To further improve the quality and diversity of the augmented dataset, duplicate-aware validation is performed by maintaining a repository containing all previously accepted sentences.

$$S = \{s_i \mid (s_i, y_i) \in D\} \quad (9)$$

A generated sentence s' is accepted only if

$$s' \notin S. \quad (10)$$

Let A_c denote the set of valid augmented samples generated for class c . The final augmented set is then defined as

$$A = \bigcup_{c \in \mathcal{C}} A_c. \quad (11)$$

Finally, the augmented dataset used for model training is constructed as

$$D^* = D \cup A. \quad (12)$$

By construction,

$$|D_c^*| \approx T, \quad \forall c \in \mathcal{C}, \quad (13)$$

which produces a nearly balanced dataset while preserving semantic fidelity and textual diversity.

Consequently, the proposed framework constructs a nearly balanced training dataset while preserving semantic fidelity and sentence diversity. By combining

rule-based perturbation techniques, transformer-based paraphrasing, semantic similarity validation, and duplicate-aware filtering, the CSHDA framework generates high-quality minority-class samples that improve the robustness and generalization capability of downstream transformer-based sentiment classifiers [39, 42]. Algorithm 1 summarizes the complete Class-Aware Semantic Hybrid Data Augmentation (CSHDA) procedure adopted in this study.

Algorithm 1 Proposed Class-Aware Semantic Hybrid Data Augmentation (CSHDA)

```

1: Input: Training dataset  $D = \{(s_i, y_i)\}_{i=1}^N$ 
2: Output: Balanced augmented training dataset  $D^*$ 
3: Compute class frequencies for all sentiment classes in  $D$ 
4: Determine target class size  $T = \max_c |D_c|$ 
5: Initialize augmented dataset  $A \leftarrow \emptyset$ 
6: Initialize sentence repository  $S$  with all original sentences
7: for each minority class  $c$  do
8:   Initialize  $A_c \leftarrow \emptyset$ 
9:    $k_c \leftarrow T - |D_c|$ 
10:  while  $|A_c| < k_c$  do
11:    Randomly select a sample  $s$  from  $D_c$ 
12:    Select augmentation technique  $t$  according to class-aware probability
    distribution  $P_c$ 
13:    Generate augmented candidate sentence  $s'$  using  $t$ 
14:    if  $s' \neq s$  then
15:      Compute SBERT similarity  $\delta = \text{sim}(s, s')$ 
16:      if  $\tau_{\min} \leq \delta \leq \tau_{\max}$  then
17:        if  $s' \notin S$  then
18:          Add  $s'$  to  $A_c$ 
19:          Add  $s'$  to  $S$ 
20:        end if
21:      end if
22:    end if
23:  end while
24:   $A \leftarrow A \cup A_c$ 
25: end for
26:  $D^* \leftarrow D \cup A$ 
27: return Balanced augmented dataset  $D^*$ 

```

The proposed CSHDA framework integrates four complementary augmentation techniques, namely SR, RD, RI, and BERT-based Paraphrasing (BP), as illustrated in Fig. 2. The framework employs a class-aware sampling strategy to select the most appropriate augmentation technique for each minority sentiment class, thereby preserving sentiment polarity while increasing textual diversity. The augmentation

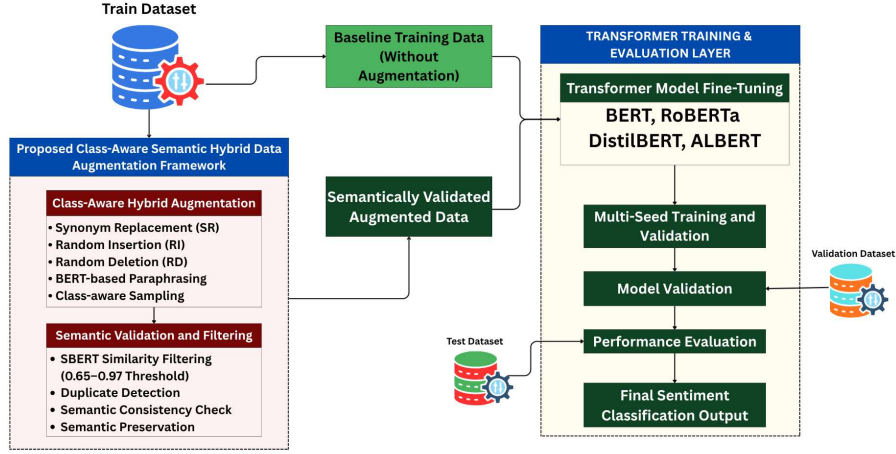


Fig. 2: Overview of the proposed Class-Aware Semantic Hybrid Data Augmentation (CSHDA) framework illustrating minority-class augmentation, semantic consistency validation, duplicate-aware filtering, and transformer-based sentiment classification.

parameters were configured as follows: one synonym replacement ($n = 1$), one random insertion ($n = 1$), a deletion probability of $p = 0.05$ for random deletion, and a transformer-based paraphrasing model for BP.

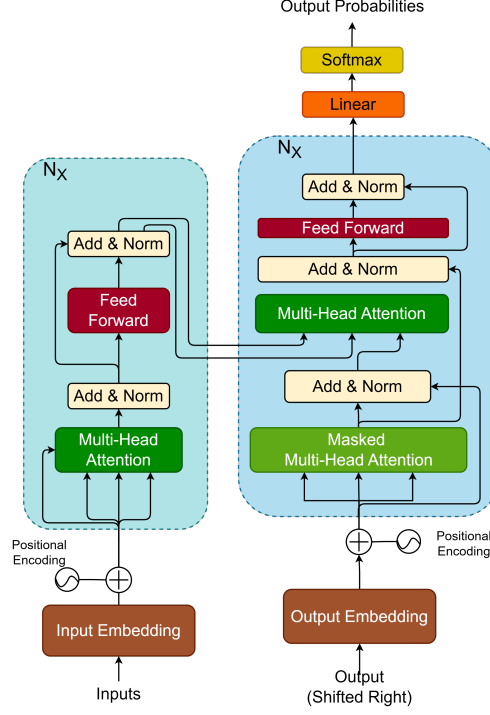


Fig. 3: Overview of the transformer encoder architecture based on the self-attention mechanism proposed by Vaswani et al. [28].

The balanced training dataset generated by the proposed CSHDA framework is subsequently used for fine-tuning transformer-based sentiment classification models. To evaluate the effectiveness and generalizability of the proposed augmentation strategy, four representative transformer architectures, namely BERT, RoBERTa, DistilBERT, and ALBERT, are employed. The corresponding transformer-based classification framework is described in the following subsection.

3.3.2 Transformer-Based Sentiment Classification

For the effectiveness of the proposed Class-Aware Semantic Hybrid Data Augmentation (CSHDA) framework evaluation, four pre-trained transformer-models BERT, RoBERTa, DistilBERT, and ALBERT, were selected. These models represent different trade-offs between contextual representation capability, computational complexity, and parameter efficiency while sharing the common transformer encoder architecture based on multi-head self-attention [28]. The balanced training dataset generated by the proposed CSHDA framework was independently used to fine-tune each transformer model under identical experimental settings, thereby enabling a fair comparison of classification performance. The transformer encoder architecture proposed by Vaswani et al. [28] is illustrated in Fig. 3.

Let consider a input sentence s that represented as a tokens sequence

$$s = \{s_1, s_2, \dots, s_n\}, \quad (14)$$

where each token is representation of continuous vector. The embedding matrix $E \in \mathbb{R}^{n \times d}$. The data is then sent through a series of stacked transformer encoder layers, where each encoder layer has a multi-headed self-attention (MHA) block along with a feed-forward network (FFN) block. The self-attention model is represented as

$$\text{Attention}(Q, K, V) = \text{softmax} \left(\frac{QK^\top}{\sqrt{d_k}} \right) V, \quad (15)$$

with Q , K , and V being the query, key, and value matrices respectively obtained from the embedding matrix E , and d_k representing the dimensions of the key vectors. The multi-head attention mechanism builds upon this idea by performing attention computations on multiple parallel heads.

$$\text{MHSA}(E) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W^O. \quad (16)$$

3.4 Experimental Setup

The experiments were performed using Google Colab with an NVIDIA Tesla T4 GPU that had 16 GB of GDDR6 memory. As the primary programming language Python (version 3.9.10) was used whereas as deep learning framework PyTorch (version 1.12.1) was adopted. For implementation of the transformer models Hugging Face Transformers (version 4.24.0) was used. Text processing, evaluation, data manipulation and numerical computations were performed using NLTK (version 3.7), Scikit-learn (version 1.1.3), Pandas (version 1.5.1) and NumPy (version 1.23.4), correspondingly. In order to make fair comparison between transformer models, all of them were fine-tuned using the same hyperparameters as shown in Table 2. The choice of pre-trained models included BERT-base-uncased, RoBERTa-base, DistilBERT-base-uncased, and ALBERT-base-v2. The characteristics of the models' architecture are presented in Table 3. To ensure the reproducibility of the experiments each model was trained and evaluated using five different random seeds $s \in \{7, 42, 99, 123, 2024\}$, and the performance is reported in term of mean and standard deviation (SD) across these runs.

3.5 Evaluation Metrics

Because the datasets considered in this study are class-imbalanced, Macro F1-score is regarded as the primary evaluation metric since it assigns equal importance to each sentiment class irrespective of class frequency. To provide a comprehensive performance assessment, Accuracy, Precision, Recall, Weighted F1-score, Macro F1-score, class-wise Recall, Confusion Matrix, and Inference Time were also evaluated. The evaluation metrics are defined as follows.

Accuracy

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (17)$$

Table 2: Training configuration used for fine-tuning transformer-based architectures.

Parameter	Value	Rationale
Learning Rate	2×10^{-5}	For fine-tuning
Batch Size	16	Balance between speed/memory
Epochs	4	Sufficient for convergence
Model Selection	Best validation checkpoint	Prevent overfitting
Warmup Steps	10% of total training steps	Stabilizes early training
Weight Decay	0.01	L2 regularization
Max Sequence Length	128	Covers majority of review lengths
Optimizer	AdamW	Decoupled weight decay
Loss Function	Cross-Entropy	Standard for classification
Dropout	0.1	Prevent overfitting
Gradient Clipping	1.0	Stabilizes training and prevents exploding gradients

Table 3: Architectural characteristics of the pre-trained transformer models used in this study [43].

Model	Layers	Hidden Size	Attention Heads	Parameters (Millions)
BERT-base-uncased	12	768	12	110M
RoBERTa-base	12	768	12	125M
DistilBERT-base-uncased	6	768	12	66M
ALBERT-base-v2	12	768	12	12M

Precision

$$\text{Precision} = \frac{TP}{TP + FP} \quad (18)$$

Recall

$$\text{Recall} = \frac{TP}{TP + FN} \quad (19)$$

F1-Score

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (20)$$

$$F1_{macro} = \frac{1}{C} \sum_{i=1}^C F1_i \quad (21)$$

$$F1_{weighted} = \sum_{i=1}^C w_i F1_i \quad (22)$$

where C denotes the number of sentiment classes, $F1_i$ is the F1-score corresponding to class i . and w_i represents the proportion of samples belonging to class i .

Since the primary objective of this work is to improve minority-class classification, particular emphasis is placed on Macro F1-score and class-wise Recall, as these metrics provide a more reliable assessment of classifier performance on imbalanced datasets than overall Accuracy. In addition, the Confusion Matrix is analyzed to investigate class-specific prediction errors, while Inference Time (ms/sample) is reported to assess computational efficiency.

4 Results and Analysis

For experimental analysis, two experimental settings were considered. First, the transformer models were trained on the original imbalanced datasets without augmentation. Second, the same models were trained using the datasets generated by the proposed Class-Aware Semantic Hybrid Data Augmentation (CSHDA) framework. Model performance was evaluated using overall metrics (Accuracy, weighted Precision, weighted Recall, weighted F1-score, and Macro F1-score) together with class-wise Recall and F1-score to provide a comprehensive assessment.

4.1 Performance on the CUFC Dataset

The performance analysis of the CUFC dataset is discussed in this part. At first, the performance analysis of the four transformer-based models is performed with the original imbalanced training set to find out the baseline performance. Later, the CSHDA model is used for generating a balanced training set by performing augmentation on the minority classes only.

4.1.1 Results on CUFC Corpus (without augmentation)

Table 4 shows the performance evaluation of the baseline transformer models using three classes: Negative, Neutral, and Positive. All four models, as shown in Table 4, achieve overall accuracy between 0.7728 and 0.8298. This means that roughly 78 to 83 out of every 100 predictions are correct. The highest performance is obtained with RoBERTa, which achieves the highest accuracy (0.8298 ± 0.0225), weighted F1-score (0.8111 ± 0.0394) and macro F1-score (0.7079 ± 0.0631), while the lowest performance is observed with DistilBERT, with an accuracy of (0.7728 ± 0.0302), weighted F1-score (0.7141 ± 0.0541) and macro F1-score (0.5543 ± 0.0948). These results indicate that RoBERTa provides stronger contextual representations, leading to higher overall classification performance on the CUFC dataset. The BERT model shows accuracy of (0.7855 ± 0.0196), weighted F1-score (0.7417 ± 0.0374) and macro F1-score (0.6080 ± 0.0664), which are slightly lower when compared with the RoBERTa. ALBERT shows comparable performance with an accuracy of (0.7804 ± 0.0177), weighted F1-score (0.7379 ± 0.0347) and macro F1-score (0.5916 ± 0.0583). The comparatively lower performance of DistilBERT can be attributed to its compressed architecture obtained through knowledge distillation, which reduces model capacity while improving computational efficiency. DistilBERT contains fewer encoder layers and substantially fewer parameters than RoBERTa, resulting in a trade-off between computational efficiency and classification performance.

Table 4: Baseline performance of transformer models on the CUFC dataset without data augmentation (Mean $\pm$ Std across five seeds).

Model	Accuracy	Precision (weighted)	F1 (weighted)	F1 (macro)	Recall (macro)
BERT	0.7855 $\pm$ 0.0196	0.8125 $\pm$ 0.0078	0.7417 $\pm$ 0.0374	0.6080 $\pm$ 0.0664	0.5799 $\pm$ 0.0490
ALBERT	0.7804 $\pm$ 0.0177	0.7757 $\pm$ 0.0218	0.7379 $\pm$ 0.0347	0.5916 $\pm$ 0.0583	0.5812 $\pm$ 0.0370
DistilBERT	0.7728 $\pm$ 0.0302	0.7428 $\pm$ 0.0885	0.7141 $\pm$ 0.0541	0.5543 $\pm$ 0.0948	0.5517 $\pm$ 0.0692
RoBERTa	0.8298 $\pm$ 0.0225	0.8374 $\pm$ 0.0159	0.8111 $\pm$ 0.0394	0.7079 $\pm$ 0.0631	0.6907 $\pm$ 0.0678

To further investigate model behaviour on imbalanced data, the per-class performance reported in Table 5 was analyzed. From the results reported in Table 5, RoBERTa achieves the highest class-wise recall for positive, neutral and negative which was computed as (0.4324 $\pm$ 0.2514, 0.6682 $\pm$ 0.0443, and 0.9714 $\pm$ 0.0187).

Table 5: Per-class performance comparison of transformer models on the CUFC dataset without data augmentation (Mean $\pm$ Std across five seeds).

Model	Recall _{neg}	Recall _{neu}	Recall _{pos}	F1 _{neg}	F1 _{neu}	F1 _{pos}
BERT	0.2162 $\pm$ 0.1911	0.5273 $\pm$ 0.0466	0.9961 $\pm$ 0.0058	0.3049 $\pm$ 0.2068	0.6448 $\pm$ 0.0352	0.8744 $\pm$ 0.0158
ALBERT	0.1676 $\pm$ 0.1169	0.5955 $\pm$ 0.0102	0.9805 $\pm$ 0.0092	0.2545 $\pm$ 0.1642	0.6376 $\pm$ 0.0106	0.8827 $\pm$ 0.0160
DistilBERT	0.1135 $\pm$ 0.2097	0.5455 $\pm$ 0.0278	0.9961 $\pm$ 0.0058	0.1510 $\pm$ 0.2546	0.6416 $\pm$ 0.0244	0.8701 $\pm$ 0.0218
RoBERTa	0.4324 $\pm$ 0.2514	0.6682 $\pm$ 0.0443	0.9714 $\pm$ 0.0187	0.5022 $\pm$ 0.1798	0.7064 $\pm$ 0.0322	0.9152 $\pm$ 0.0238

The per-class results further indicate that all four transformer models accurately classify the majority positive class but exhibit lower recall for the minority negative and neutral classes. In addition, as the negative recall score (the minority class) is extremely low, the four models are unable to detect negative classes. Also, sometime it may also fail to identify the neutral class, thereby resulting in lower macro F1 score. The confusion matrices of all four models are shown in Fig. 4, which provide further insights into misclassification patterns.

Across all baseline models, a clear bias toward the majority positive class is observed. While positive reviews are classified with high accuracy, many negative and neutral instances are misclassified as positive, indicating that class imbalance adversely affects the models’ ability to learn minority-class representations. Among the evaluated architectures, RoBERTa provides relatively better discrimination of minority sentiment classes by correctly identifying more negative and neutral samples than BERT, DistilBERT, and ALBERT. However, noticeable confusion among the minority classes remains across all baseline models, highlighting the need for class imbalance mitigation in transformer-based sentiment classification.

Since the negative class represents a minority class, the relatively low recall observed for several models indicates that they struggle to identify negative sentiment instances. This limitation can reduce their reliability in real-world applications where minority class detection is important. The low scores recorded in light or parameter efficient models highlight the challenge of attaining high precision and recall when augmentation is not utilized. This demonstrates that more sophisticated methods

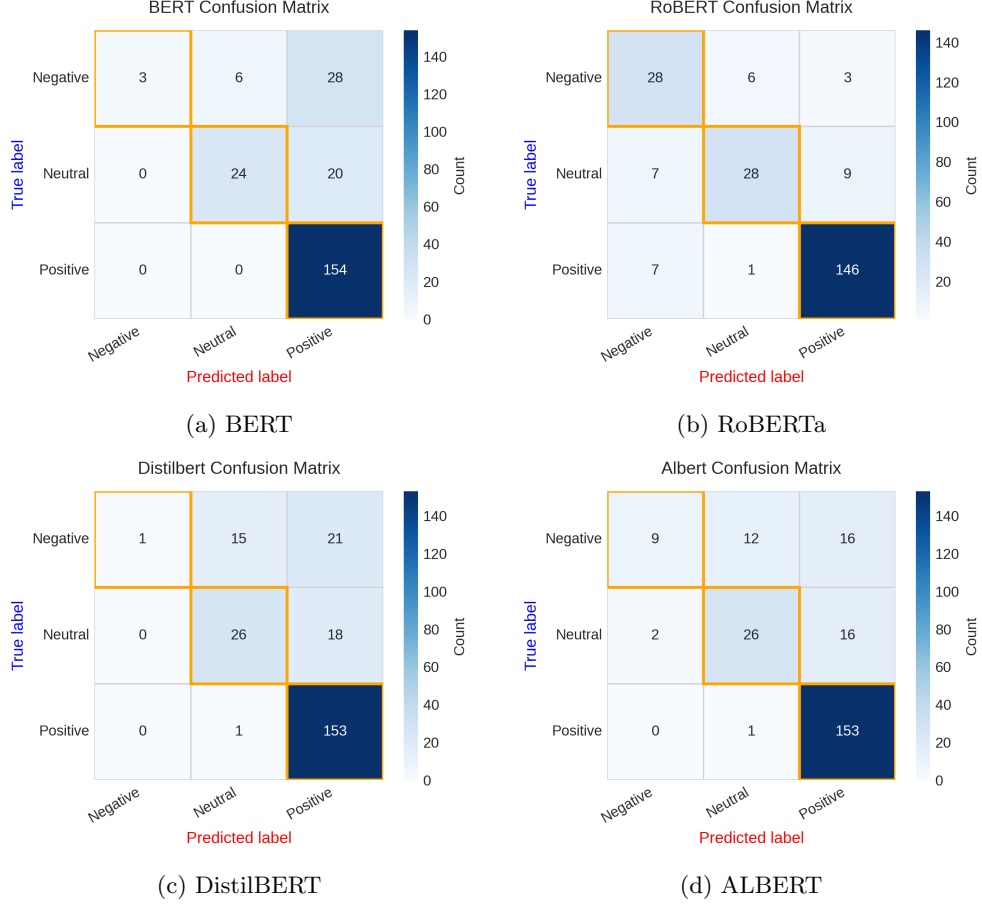


Fig. 4: Confusion matrices of transformer-based models trained without augmentation on the CUFC dataset.

of training, including augmentation, will be necessary to increase the detection of minority classes. Therefore, in next section we presented the results on the proposed Class-Aware Semantic Hybrid Augmentation CUFC Corpus.

4.1.2 Performance of the Proposed Class-Aware Semantic Hybrid Augmentation CUFC Corpus

For the CUFC corpus, augmentation is applied only to the minority sentiment classes (negative and neutral), while the majority positive class remains unchanged. Therefore, the class-wise augmentation probabilities are used for each minority class as shown in Table 6.

Table 6: Proposed class-aware augmentation probabilities used for minority classes for CUFC datasets.

Technique	Negative Class	Neutral Class
Random Deletion (RD)	0.50	0.20
BERT Paraphrasing (BP)	0.30	0.40
Random Insertion (RI)	0.15	0.30
Synonym Replacement (SR)	0.05	0.10

Higher probability is assigned to random deletion for the negative class because small lexical removals help maintain the original sentiment polarity while introducing minor structural variations. A relative high probability is assigned to BERT-based paraphrasing, to generate semantically equivalent sentences without altering the semantic meaning. Similarly, for the neutral class, BERT paraphrasing and random insertion are given higher probabilities to increase contextual diversity, since neutral sentences generally contain less explicit sentiment indicators. On the other hand, synonym replacement is less common in both classes to prevent excessive lexical changes that could distort the original meaning or sentiment. This class-specific weighting approach enables controlled augmentation, maintaining semantic accuracy while diversifying the dataset. The augmented training dataset generated by the proposed

Table 7: Performance comparison of transformer models on the CUFC dataset using the proposed Class-Aware Semantic Hybrid Data Augmentation (CSHDA) framework (Mean $\pm$ Std across five seeds).

Model	Accuracy	Precision (weighted)	F1 (weighted)	F1 (macro)	Recall (macro)
BERT	0.8715 $\pm$ 0.0251	0.8731 $\pm$ 0.0274	0.8718 $\pm$ 0.0262	0.8110 $\pm$ 0.0411	0.8180 $\pm$ 0.0468
ALBERT	0.8366 $\pm$ 0.0077	0.8418 $\pm$ 0.0064	0.8378 $\pm$ 0.0054	0.7660 $\pm$ 0.0078	0.7730 $\pm$ 0.0146
DistilBERT	0.8689 $\pm$ 0.0139	0.8711 $\pm$ 0.0166	0.8685 $\pm$ 0.0151	0.8001 $\pm$ 0.0253	0.8034 $\pm$ 0.0267
RoBERTa	0.8723 $\pm$ 0.0120	0.8772 $\pm$ 0.0111	0.8733 $\pm$ 0.0103	0.8122 $\pm$ 0.0159	0.8234 $\pm$ 0.0162

class-aware semantic hybrid augmentation framework was used to train and evaluate the transformer models in order to examine the impact of the augmentation strategy on sentiment classification performance. Table 7 shows the performance evaluation of all four models using the proposed CSHDA framework while Table 8 reports the corresponding per-class performance.

Starting with BERT (with proposed CSHDA framework), Table 7 shows that applying the proposed augmentation strategy improves the performance compared with the baseline model. The model achieves an accuracy of 0.8715 ± 0.0251 and a macro F1-score of 0.8110 ± 0.0411 . On the other hand, as shown in Table 8 a notable improvement is observed in minority class detection (both negative and neutral), where the negative recall increases to 0.8216 ± 0.0622 . The per-class performance is as follows: Positive: (Recall 0.9325, F1 0.9392), Neutral: (Recall 0.7000, F1 0.7083), and Negative: (Recall 0.8216, F1 0.7853). For the BERT model improvement of 8.60%

Table 8: Per-class performance comparison of transformer models on the CUFC dataset using the proposed Class-Aware Semantic Hybrid Data Augmentation (CSHDA) framework (Mean $\pm$ Std across five seeds).

Model	Recall _{neg}	Recall _{neu}	Recall _{pos}	F1 _{neg}	F1 _{neu}	F1 _{pos}
BERT	0.8216 $\pm$ 0.0622	0.7000 $\pm$ 0.0826	0.9325 $\pm$ 0.0098	0.7853 $\pm$ 0.0538	0.7083 $\pm$ 0.0626	0.9392 $\pm$ 0.0113
ALBERT	0.7568 $\pm$ 0.0541	0.6545 $\pm$ 0.0102	0.9078 $\pm$ 0.0231	0.7145 $\pm$ 0.0052	0.6674 $\pm$ 0.0230	0.9161 $\pm$ 0.0044
DistilBERT	0.7946 $\pm$ 0.0491	0.6727 $\pm$ 0.0998	0.9429 $\pm$ 0.0107	0.7683 $\pm$ 0.0180	0.6880 $\pm$ 0.0599	0.9441 $\pm$ 0.0068
RoBERTa	0.8541 $\pm$ 0.0363	0.6864 $\pm$ 0.0493	0.9299 $\pm$ 0.0192	0.7943 $\pm$ 0.0324	0.7008 $\pm$ 0.0295	0.9415 $\pm$ 0.0056

and 20.30% is obtained in accuracy and macro F1 score when compared with its corresponding baseline model. DistilBERT also benefits from the proposed augmentation strategy. The model achieves an accuracy of 0.8689 ± 0.0139 . The negative recall improves to 0.7946 ± 0.0491 , indicating better detection of minority sentiment samples. The weighted precision and macro recall reach 0.8711 ± 0.0166 and 0.8034 ± 0.0267 , respectively. The class-wise performance is: Positive: (Recall 0.9429, F1 0.9441), Neutral: (Recall 0.6727, F1 0.6880), and Negative: (Recall 0.7946, F1 0.7683). The results obtained for the DistilBERT model indicate the best overall performance (in terms of percentage) with improvement of 9.61% and 24.58% for respectively.

On the other hand, the improvement in RoBERTa model showed the consistency performance for the baseline model. Here also the RoBERTa (with proposed CSHDA framework) performs the best among all four models. RoBERTa achieves an accuracy of 0.8723 ± 0.0120 and a weighted F1-score of 0.8733 ± 0.0103 . In addition, the negative recall reaches 0.8541 ± 0.0363 , demonstrating strong capability in detecting minority class instances. The weighted precision and macro recall are 0.8772 ± 0.0111 and 0.8234 ± 0.0162 , respectively. The class-wise results are: Positive: (Recall 0.9299, F1 0.9415), Neutral: (Recall 0.6864, F1 0.7008), and Negative: (Recall 0.8541, F1 0.7943). Similarly, ALBERT also shows noticeable improvements when trained using the proposed augmentation framework. The model achieves an accuracy of 0.8366 ± 0.0077 and a weighted F1-score of 0.8378 ± 0.0054 . The negative recall increases to 0.7568 ± 0.0541 , while the weighted precision and macro recall reach 0.8418 ± 0.0064 and 0.7730 ± 0.0146 , respectively. The class-wise performance is: Positive: (Recall 0.9078, F1 0.9161), Neutral: (Recall 0.6545, F1 0.6674), and Negative: (Recall 0.7568, F1 0.7145).

Thus, the proposed class-aware semantic hybrid augmentation shows the consistent improvement over all transformer models. The proposed augmentation technique results demonstrate that the proposed augmentation strategy helps the models learn more balanced sentiment representations and reduces cross-class confusion and improving the over all macro F1 score. Fig. 5 presents the confusion matrices obtained after applying the proposed CSHDA framework. Compared with the baseline models, the diagonal elements become substantially stronger across all three sentiment classes, indicating an increase in correctly classified samples. In particular, the number of negative samples incorrectly predicted as positive is considerably reduced, while the recognition of neutral samples is noticeably improved. These improvements indicate that the proposed augmentation strategy effectively enhances minority-class representation without degrading the classification performance of the majority class.

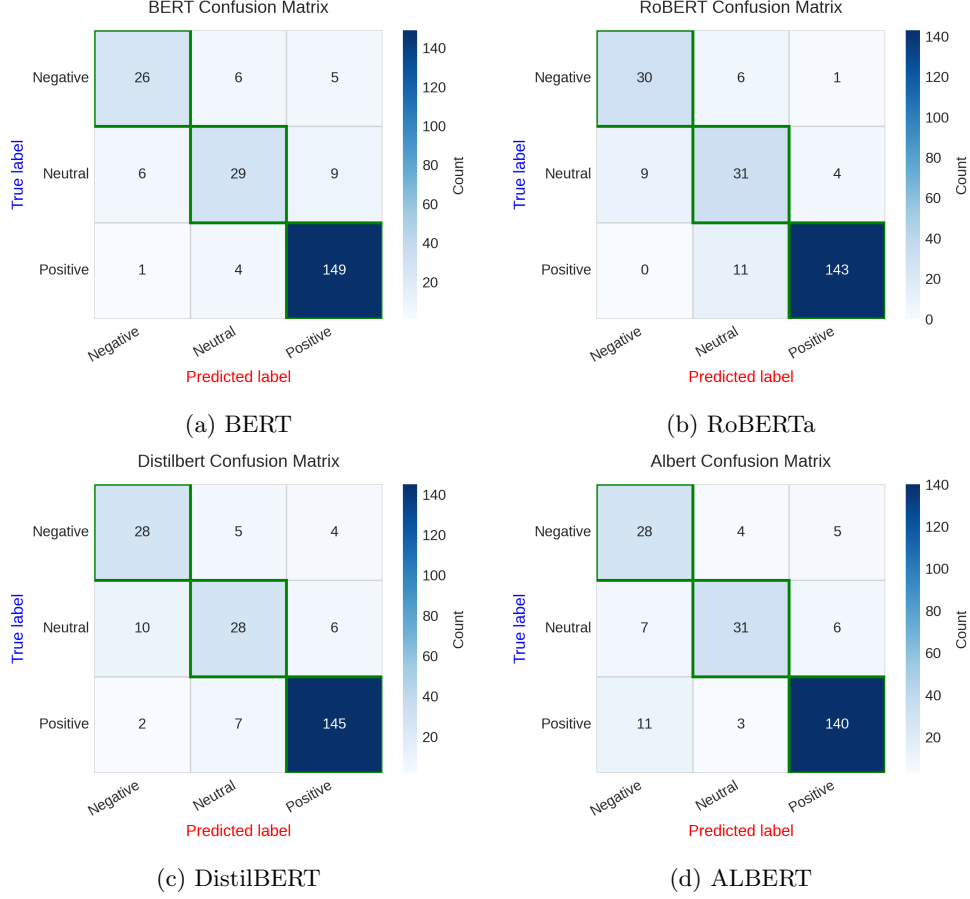
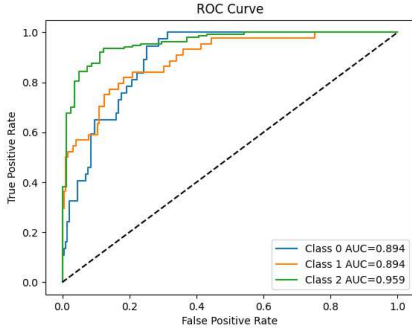


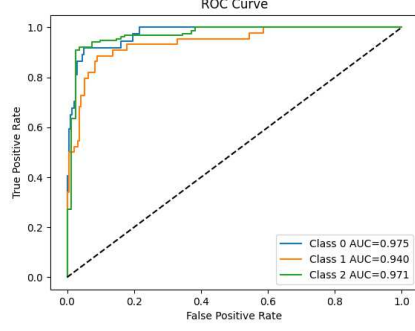
Fig. 5: Confusion matrices of transformer models trained with the proposed CSHDA framework on the CUFC dataset.

Consequently, the confusion matrices confirm that CSHDA produces a more balanced decision boundary and alleviates the prediction bias caused by the original class imbalance.

The ROC curves of the baseline BERT model and the proposed CSHDA-enhanced BERT model on the CUFC dataset is reported in Fig. 6. The proposed framework consistently shifts the ROC curves toward the upper-left corner, indicating a higher true positive rate (TPR) for the same false positive rate (FPR). This improvement demonstrates that the augmented training data enables the classifier to achieve better discrimination among sentiment classes. Furthermore, the Area Under the Curve (AUC) increases across all classes after applying the proposed CSHDA framework, confirming its effectiveness in learning more discriminative decision boundaries. For the negative sentiment class (Class 0), the AUC increases from 0.894 in the baseline



(a) ROC curve for baseline BERT without augmentation



(b) ROC curve for BERT with Class-Aware Semantic Hybrid Data Augmentation (CSHDA) framework

Fig. 6: ROC curve comparison of the baseline BERT model and the proposed Class-Aware Semantic Hybrid Data Augmentation (CSHDA) framework on the CUFC dataset. The augmented model demonstrates improved discriminative capability with higher AUC values across sentiment classes.

model to 0.975 after applying CSHDA, indicating a noticeable improvement in distinguishing minority negative samples from the remaining classes. Similarly, the neutral class (Class 1) improves from 0.894 to 0.940, demonstrating enhanced recognition of the most ambiguous sentiment category. The positive class (Class 2) also exhibits a performance gain, with the AUC increasing from 0.959 to 0.971. These consistent improvements across all sentiment classes confirm that the proposed augmentation strategy enhances class separability while preserving the strong performance of the majority class.

4.1.3 Comparative Ablation Analysis of Individual Augmentation Techniques For CUFC Corpus

In the first part of this study, we determined the performance of four transformer architectures based on the original datasets, which are unbalanced. It was found that the RoBERTa model had the best performance compared to other baseline models. However, it was reported from the result that all baseline models struggled significantly with the Negative and Neutral class. This was due to imbalanced data, and during training, the model was biased toward the majority class. Therefore, we proposed a class aware smart hybrid approach to mitigate this bias. The proposed framework shows the competitive results in terms of accuracy and macro F1 score. Also, the proposed scheme improve the overall performance of the minority class. However, we should also compare the proposed frame work with the existing EDA techniques. Using the same training configuration and to have a fair comparison, we test the four EDA augmentation scheme (SR,RI,RD,RS) and compare them with the proposed

Table 9: Ablation study comparing individual augmentation techniques and the proposed Class-Aware Semantic Hybrid Data Augmentation (CSHDA) across transformer models.

Technique	BERT	ALBERT	DistilBERT	RoBERTa
Accuracy				
SR	0.8681 ± 0.0090	0.8051 ± 0.0358	0.8604 ± 0.0126	0.8664 ± 0.0130
RI	0.8587 ± 0.0133	0.8323 ± 0.0253	0.8502 ± 0.0133	0.8757 ± 0.0092
RD	0.8706 ± 0.0065	0.8221 ± 0.0174	0.8698 ± 0.0140	0.8672 ± 0.0174
RS	0.8553 ± 0.0043	0.8400 ± 0.0174	0.8621 ± 0.0149	0.8638 ± 0.0141
Proposed	0.8715 ± 0.0251	0.8366 ± 0.0077	0.8689 ± 0.0139	0.8723 ± 0.0120
F1-Macro				
SR	0.8019 ± 0.0224	0.7282 ± 0.0515	0.7849 ± 0.0192	0.8030 ± 0.0180
RI	0.7930 ± 0.0189	0.7513 ± 0.0360	0.7744 ± 0.0197	0.8154 ± 0.0071
RD	0.8067 ± 0.0143	0.7507 ± 0.0154	0.8054 ± 0.0205	0.8088 ± 0.0233
RS	0.7938 ± 0.0033	0.7500 ± 0.0516	0.7909 ± 0.0237	0.7980 ± 0.0119
Proposed	0.8110 ± 0.0411	0.7660 ± 0.0078	0.8001 ± 0.0253	0.8122 ± 0.0159
Recall-Macro				
SR	0.7937 ± 0.0305	0.7278 ± 0.0521	0.7751 ± 0.0231	0.8069 ± 0.0093
RI	0.7834 ± 0.0214	0.7415 ± 0.0518	0.7573 ± 0.0283	0.8136 ± 0.0097
RD	0.7972 ± 0.0163	0.7675 ± 0.0152	0.8196 ± 0.0148	0.8058 ± 0.0129
RS	0.7840 ± 0.0141	0.7259 ± 0.0677	0.7821 ± 0.0334	0.7999 ± 0.0135
Proposed	0.8180 ± 0.0468	0.7730 ± 0.0146	0.8034 ± 0.0267	0.8234 ± 0.0162

Table 10: Per-class F1-score comparison of augmentation techniques across transformer models (Mean $\pm$ Std).

Augmentation	Class	BERT	ALBERT	DistilBERT	RoBERTa
SR	Negative	0.7720 ± 0.0382	0.6151 ± 0.1158	0.7408 ± 0.0307	0.7717 ± 0.0378
	Neutral	0.6980 ± 0.0389	0.6807 ± 0.0322	0.6770 ± 0.0356	0.6988 ± 0.0186
	Positive	0.9355 ± 0.0032	0.8887 ± 0.0206	0.9370 ± 0.0128	0.9385 ± 0.0066
RI	Negative	0.7693 ± 0.0243	0.6603 ± 0.0589	0.7453 ± 0.0200	0.7909 ± 0.0142
	Neutral	0.6819 ± 0.0358	0.6825 ± 0.0420	0.6529 ± 0.0357	0.7129 ± 0.0106
	Positive	0.9277 ± 0.0078	0.9112 ± 0.0255	0.9251 ± 0.0122	0.9423 ± 0.0070
RD	Negative	0.7818 ± 0.0080	0.6673 ± 0.0169	0.7447 ± 0.0326	0.7862 ± 0.0516
	Neutral	0.7034 ± 0.0382	0.6779 ± 0.0295	0.7270 ± 0.0267	0.7087 ± 0.0250
	Positive	0.9350 ± 0.0031	0.9070 ± 0.0155	0.9444 ± 0.0046	0.9315 ± 0.0075
RS	Negative	0.7548 ± 0.0303	0.5931 ± 0.1384	0.7552 ± 0.0209	0.7547 ± 0.0198
	Neutral	0.7061 ± 0.0307	0.7512 ± 0.0256	0.6808 ± 0.0460	0.7020 ± 0.0261
	Positive	0.9205 ± 0.0085	0.9057 ± 0.0127	0.9367 ± 0.0120	0.9373 ± 0.0123
Proposed	Negative	0.7853 ± 0.0538	0.7145 ± 0.0052	0.7683 ± 0.0180	0.7943 ± 0.0324
	Neutral	0.7083 ± 0.0626	0.6674 ± 0.0230	0.6880 ± 0.0599	0.7008 ± 0.0295
	Positive	0.9392 ± 0.0113	0.9161 ± 0.0044	0.9441 ± 0.0068	0.9415 ± 0.0056

framework. Table 9 shows the results obtained using the EDA techniques (SR, RI, RS and RD).

From Table 9, proposed method provided improvement in accuracy for BERT, whereas for ALBERT model RS provide the best accuracy, RD for DistilBERT and RI for RoBERTa. However, comparing the results based on macro F1 score the proposed augmentation shows highest F1 score for the BERT and ALBERT model whereas for DistilBERT and RoBERTa is obtained for RD and RI EDA techniques. Similarly, the recall score for the proposed framework shows improvement for the BERT, ALBERT and RoBERTa model, RD for DistilBERT model. This shows the consistency of proposed CSHDA framework techniques. However, RD sometimes improve the performance of model but excessive deletion may change or lose the semantic information. The results in Table 10 demonstrate that proposed CSHDA framework provides the improvement across all models for the minority classes (Negative). This again shows the robustness of the proposed framework.

4.2 Cross-Domain Validation Using the Twitter Sentiment Dataset

To further evaluate the generalization capability of the proposed Class-Aware Semantic Hybrid Data Augmentation (CSHDA) framework, experiments were conducted on the Twitter Airline Sentiment dataset [32]. Unlike the CUFC dataset, where the positive class is dominant, the Twitter dataset is dominated by negative sentiment, thereby providing a complementary benchmark for evaluating the robustness of the proposed framework under a different class imbalance distribution.

In Twitter dataset the percentage of negative, neutral and positive are 62.70%, 21.17% and 16.14% respectively. Further, to simulate a low-resource scenario we restrict the dataset to 1500 samples and evaluate performance across multiple transformer-based models, including BERT, ALBERT, DistilBERT, and RoBERTa, using five different random seeds using the same experimental settings used for the CUFC dataset. However, in CUFC dataset we have Positive reviews (65.50%), Neutral reviews (18.91%) and Negative reviews (15.59%) i.e., the positive class was the majority class. On the other hand for Twitter Airline Sentiment the negative are the majority. Therefore, the class-aware semantic hybrid augmentation for the Twitter dataset is given in From Table 11.

Table 11: Proposed class-aware augmentation probabilities used for minority classes for Twitter datasets.

Technique	Positive Class	Neutral Class
BERT Paraphrasing (BP)	0.45	0.60
Synonym Replacement (SR)	0.25	0.25
Random Deletion (RD)	0.05	0.05
Random Insertion (RI)	0.25	0.10

The experimental results on the Twitter dataset using the baseline model and the proposed augmentation is presented in Table 12. For BERT, the application of proposed CSHDA framework leads to a consistent improvement in F1-macro (60.65% to

Table 12: Performance comparison of baseline and proposed CSHDA framework on the Twitter dataset. Results are reported as mean $\pm$ standard deviation over 5 seeds.

	Model	Accuracy	Precision (weighted)	F1 (weighted)	F1 (macro)	Recall (macro)
Baseline	BERT	0.7517 $\pm$ 0.0092	0.7308 $\pm$ 0.0206	0.7056 $\pm$ 0.0192	0.6065 $\pm$ 0.0311	0.6114 $\pm$ 0.0271
	ALBERT	0.7302 $\pm$ 0.0439	0.7472 $\pm$ 0.0311	0.6902 $\pm$ 0.0430	0.5833 $\pm$ 0.0684	0.5732 $\pm$ 0.0706
	DistilBERT	0.6879 $\pm$ 0.0575	0.5776 $\pm$ 0.1640	0.5952 $\pm$ 0.1057	0.4298 $\pm$ 0.1671	0.4481 $\pm$ 0.1150
	RoBERTa	0.8188 $\pm$ 0.0156	0.8230 $\pm$ 0.0130	0.8166 $\pm$ 0.0121	0.7628 $\pm$ 0.0143	0.7592 $\pm$ 0.0314
Proposed	BERT	0.7732 $\pm$ 0.0202	0.7775 $\pm$ 0.0126	0.7704 $\pm$ 0.0171	0.6982 $\pm$ 0.0221	0.7083 $\pm$ 0.0207
	ALBERT	0.7664 $\pm$ 0.0397	0.7792 $\pm$ 0.0210	0.7669 $\pm$ 0.0322	0.6965 $\pm$ 0.0320	0.7068 $\pm$ 0.0310
	DistilBERT	0.7812 $\pm$ 0.0065	0.7814 $\pm$ 0.0128	0.7783 $\pm$ 0.0094	0.7121 $\pm$ 0.0167	0.7196 $\pm$ 0.0275
	RoBERTa	0.7980 $\pm$ 0.0139	0.7957 $\pm$ 0.0113	0.7955 $\pm$ 0.0123	0.7316 $\pm$ 0.0204	0.7345 $\pm$ 0.0187

69.82%) and recall-macro (61.14% to 70.83%), indicating enhanced balanced learning across sentiment classes. This improvement is primarily driven by a substantial increase in neutral class recall, which rises from 0.1746 to 0.4984 as shown in Table 13. A similar trend is observed for ALBERT, where F1-macro improves by over 11%, alongside significant gains in both neutral and positive class recall. These results suggest that the proposed augmentation effectively mitigates class imbalance and enriches semantic diversity in the training data.

Table 13: Per-class performance comparison of transformer models on the Twitter dataset using the proposed Class-Aware Semantic Hybrid Data Augmentation (CSHDA) framework (Mean $\pm$ Std across five seeds).

	Model	Recall _{neg}	Recall _{neu}	Recall _{pos}	F1 _{neg}	F1 _{neu}	F1 _{pos}
Baseline	BERT	0.9574 $\pm$ 0.0150	0.1746 $\pm$ 0.0877	0.7021 $\pm$ 0.1329	0.8555 $\pm$ 0.0038	0.2572 $\pm$ 0.0974	0.7068 $\pm$ 0.0533
	ALBERT	0.9372 $\pm$ 0.0662	0.2889 $\pm$ 0.1852	0.4936 $\pm$ 0.2764	0.8358 $\pm$ 0.0253	0.3488 $\pm$ 0.0966	0.5652 $\pm$ 0.2164
	DistilBERT	0.9947 $\pm$ 0.0092	0.0984 $\pm$ 0.1395	0.2511 $\pm$ 0.2352	0.8122 $\pm$ 0.0375	0.1401 $\pm$ 0.1872	0.3371 $\pm$ 0.3111
	RoBERTa	0.9011 $\pm$ 0.0613	0.6190 $\pm$ 0.1023	0.7574 $\pm$ 0.0923	0.8919 $\pm$ 0.0163	0.6260 $\pm$ 0.0350	0.7705 $\pm$ 0.0257
Proposed	BERT	0.8691 $\pm$ 0.0526	0.4984 $\pm$ 0.1020	0.7574 $\pm$ 0.0466	0.8702 $\pm$ 0.0133	0.5275 $\pm$ 0.0511	0.6967 $\pm$ 0.0272
	ALBERT	0.8511 $\pm$ 0.0852	0.5460 $\pm$ 0.1109	0.7234 $\pm$ 0.0796	0.8621 $\pm$ 0.0339	0.5478 $\pm$ 0.0492	0.6796 $\pm$ 0.0293
	DistilBERT	0.8723 $\pm$ 0.0296	0.5206 $\pm$ 0.0879	0.7660 $\pm$ 0.0752	0.8704 $\pm$ 0.0074	0.5494 $\pm$ 0.0448	0.7166 $\pm$ 0.0272
	RoBERTa	0.8904 $\pm$ 0.0234	0.5429 $\pm$ 0.0284	0.7702 $\pm$ 0.0571	0.8842 $\pm$ 0.0056	0.5758 $\pm$ 0.0185	0.7347 $\pm$ 0.0466

The largest relative improvement is observed in DistilBERT, which exhibits a performance boost across all evaluation metrics. Specifically, F1-macro increases from 42.98% to 71.21%, while recall-macro improves from 44.81% to 71.96%. Notably, neutral class recall rises from a severely limited 0.0984 to 0.5206, and positive recall from 0.2511 to 0.7660. This indicates that, in the baseline setting, DistilBERT suffers from a strong bias toward the majority class, producing predictions that are heavily biased toward the majority class. The proposed augmentation successfully alleviates this issue, enabling the model to learn a more balanced decision boundary. These findings highlight the effectiveness of the approach for low-capacity models operating under limited data conditions.

RoBERTa already demonstrates strong baseline performance on the Twitter dataset i.e., (with an average accuracy of 81.88%) and having a low standard deviation of $\pm$ 0.0156 across five random seeds. From the baseline model, it can be seen from the results that recall performance of the negative, neutral and positive are 90.11%, 61.90% and 75.74% respectively i.e., RoBERTa already achieves a balanced representation across sentiment classes. This suggests that RoBERTa has already learned sufficiently

discriminative representations from the original training data. Thus, suggesting that the model may already captured sufficient linguistic variability from training data. Consequently, the proposed augmentation results in a slight reduction in the overall performance metrics. Therefore, using the augmentation the performance of the proposed augmentation, shows a slight decrease in overall metrics (F1-weighted drops from 81.66% to 79.55%). The observed reduction in Macro F1-score (approximately 3.1%) is relatively small, suggesting that the proposed augmentation remains robust even when applied to a model that already exhibits strong baseline performance.

The ablation study comparing individual augmentation techniques with the proposed Class-Aware Semantic Hybrid Data Augmentation (CSHDA) framework on the Twitter dataset is presented in Table 14. For the BERT model, Random Deletion (RD) achieves the highest performance with an accuracy of 78.26% and an F1-Macro score of 70.10%. The proposed CSHDA approach shows a small decrease of nearly 1.15% in accuracy compared with RD, while the difference in Recall-Macro remains limited to only 0.78%. Similarly, the F1-Macro score obtained using the hybrid framework (69.82%) remains very close to RD, with only a marginal reduction of 0.39%. These observations suggest that the hybrid strategy is able to preserve balanced class-wise performance while maintaining semantic consistency in the generated samples.

Table 14: Ablation study comparing individual augmentation techniques and the proposed Class-Aware Semantic Hybrid Data Augmentation (CSHDA) across transformer models for Twitter Datasets.

Technique	BERT	ALBERT	DistilBERT	RoBERTa
Accuracy				
SR	0.7705 $\pm$ 0.0188	0.7805 $\pm$ 0.0177	0.7671 $\pm$ 0.0135	0.7953 $\pm$ 0.0142
RI	0.7765 $\pm$ 0.0155	0.7000 $\pm$ 0.1937	0.7805 $\pm$ 0.0144	0.7805 $\pm$ 0.0255
RD	0.7826 $\pm$ 0.0060	0.7443 $\pm$ 0.0577	0.7738 $\pm$ 0.0218	0.7886 $\pm$ 0.0293
RS	0.7537 $\pm$ 0.0140	0.7765 $\pm$ 0.0216	0.7027 $\pm$ 0.0357	0.8007 $\pm$ 0.0321
Proposed	0.7732 $\pm$ 0.0202	0.7664 $\pm$ 0.0397	0.7812 $\pm$ 0.0065	0.7980 $\pm$ 0.0139
F1-Macro				
SR	0.6969 $\pm$ 0.0217	0.7065 $\pm$ 0.0171	0.6925 $\pm$ 0.0246	0.7294 $\pm$ 0.0218
RI	0.6874 $\pm$ 0.0168	0.6606 $\pm$ 0.1345	0.6973 $\pm$ 0.0251	0.7123 $\pm$ 0.0198
RD	0.7010 $\pm$ 0.0250	0.6840 $\pm$ 0.0323	0.6995 $\pm$ 0.0264	0.7195 $\pm$ 0.0250
RS	0.6064 $\pm$ 0.0223	0.6832 $\pm$ 0.0394	0.4779 $\pm$ 0.0989	0.7348 $\pm$ 0.0226
Proposed	0.6982 $\pm$ 0.0221	0.6965 $\pm$ 0.0320	0.7121 $\pm$ 0.0167	0.7316 $\pm$ 0.0204
Recall-Macro				
SR	0.7026 $\pm$ 0.0286	0.7079 $\pm$ 0.0241	0.6976 $\pm$ 0.0365	0.7412 $\pm$ 0.0229
RI	0.6815 $\pm$ 0.0230	0.6911 $\pm$ 0.0840	0.6906 $\pm$ 0.0318	0.7281 $\pm$ 0.0322
RD	0.7038 $\pm$ 0.0307	0.6933 $\pm$ 0.0203	0.6934 $\pm$ 0.0312	0.7345 $\pm$ 0.0156
RS	0.6149 $\pm$ 0.0386	0.6663 $\pm$ 0.0473	0.4838 $\pm$ 0.0795	0.7297 $\pm$ 0.0230
Proposed	0.7083 $\pm$ 0.0207	0.7068 $\pm$ 0.0310	0.7196 $\pm$ 0.0275	0.7345 $\pm$ 0.0187

For the ALBERT model, the proposed Class-Aware Semantic Hybrid Data Augmentation (CSHDA) framework achieves performance comparable to the best

individual augmentation technique. The obtained F1-Macro score of 69.65% is approximately 1.41% lower than the best-performing Synonym Replacement (SR) approach, which achieves 70.65%. Despite the slight reduction, the hybrid framework provides more stable performance across all evaluation metrics without introducing excessive variations. In the case of DistilBERT, the proposed CSHDA framework demonstrates the best overall improvement among all augmentation methods. The hybrid framework improves accuracy by approximately 0.93% compared with the best individual augmentation method, Random Insertion (RI). Similarly, the F1-Macro score improves by nearly 2.40% compared with RD and by more than 48% compared with RS. In addition, the Recall-Macro score increases by approximately 3.25% over RD. These results indicate that the hybrid augmentation strategy significantly enhances minority-class learning and contextual generalization, particularly for lightweight transformer architectures such as DistilBERT.

For the RoBERTa model, Random Swap (RS) achieves the highest accuracy (80.07%) and F1-Macro score (73.48%). However, the proposed CSHDA framework achieves closely comparable performance, with only a 0.34% decrease in accuracy and a 0.44% reduction in F1-Macro. Furthermore, the Recall-Macro score obtained using the hybrid framework differs by only 0.65% from the best-performing RD method. Overall, the ablation results demonstrate that the proposed Class-Aware Semantic Hybrid Data Augmentation (CSHDA) framework provides consistent and generalized performance across different transformer architectures. By combining multiple augmentation strategies, the framework effectively balances diversity generation and semantic preservation, leading to improved robustness in transformer-based sentiment classification on imbalanced Twitter datasets.

5 Statistical Significance and Computational Efficiency Analysis

To assess whether the improvements achieved by the proposed Class-Aware Semantic Hybrid Data Augmentation (CSHDA) framework are statistically reliable, a statistical significance analysis was performed using the Macro F1-score obtained on the CUFC dataset. Since both the baseline and proposed CSHDA augmented models were evaluated using identical experimental settings and random seeds, a paired statistical analysis was conducted to compare their performance. For this, the null hypothesis (H_0) states that there is no statistically significant difference between the performance of the baseline models and the proposed Class-Aware Semantic Hybrid Data Augmentation (CSHDA) framework models, whereas hypothesis (H_1) assumes that a statistically significant difference exists between the two approaches. The hypotheses are formulated as:

$$\begin{aligned} H_0 : \quad & \mu_{\text{baseline}} = \mu_{\text{CSHDA}} \\ H_1 : \quad & \mu_{\text{baseline}} \neq \mu_{\text{CSHDA}} \end{aligned} \tag{23}$$

where μ_{baseline} and μ_{CSHDA} represent the mean performance of the baseline and augmented models, respectively.

Furthermore, to estimate the effect size of the proposed CSHDA technique and assess its practical significance the Cohen’s d is computed as:

$$d = \frac{\bar{X}_{\text{CSHDA}} - \bar{X}_{\text{baseline}}}{s_d} \quad (24)$$

where $\bar{X}_{\text{CSHDA}}$ and $\bar{X}_{\text{baseline}}$ denote the mean performance scores of the augmented and baseline models, respectively, and s_d represents the standard deviation of the paired differences. A 95% confidence interval (CI) was computed for the mean paired difference to quantify the reliability of the observed performance improvements. The statistical results are summarized in Table 15.

Table 15: Statistical significance analysis comparing baseline models with the proposed Class-Aware Semantic Hybrid Data Augmentation (CSHDA) framework using Macro F1-scores (CUFC dataset).

Model	Baseline	Proposed (CSHDA)	p-value	Cohen’s d	95% CI
BERT	0.608 ± 0.066	0.811 ± 0.041	0.0031**	2.858	(0.115, 0.291)
ALBERT	0.592 ± 0.058	0.766 ± 0.008	0.0034**	2.776	(0.096, 0.252)
DistilBERT	0.554 ± 0.095	0.800 ± 0.025	0.0084**	2.167	(0.105, 0.387)
RoBERTa	0.708 ± 0.063	0.812 ± 0.016	0.0169*	1.763	(0.031, 0.178)

Notes: Results are reported as Mean $\pm$ Standard Deviation of the Macro F1-score over five independent runs. Statistical significance was evaluated using a paired Student’s t -test. * $p < 0.05$ and ** $p < 0.01$. Cohen’s d denotes the standardized effect size, and 95% CI represents the confidence interval of the paired mean difference.

The paired t -test was performed using the Macro F1-scores obtained across five independent runs using $\alpha = 0.05$. Additionally, the Wilcoxon signed-rank test was also evaluated as a non-parametric robustness check on the five paired observations. However, Wilcoxon test produced an exact two-sided p -value of 0.0625 for all evaluated models. Although this value does not satisfy the conventional significance threshold ($p < 0.05$), it is expected because the exact Wilcoxon test has limited statistical resolution when only five paired observations are available. In contrast, the paired t -test yielded statistically significant improvements for all transformer models. Together with the consistently large Cohen’s d values and 95% confidence intervals that exclude zero, these results provide strong evidence that the proposed CSHDA framework yields statistically reliable and practically meaningful performance improvements.

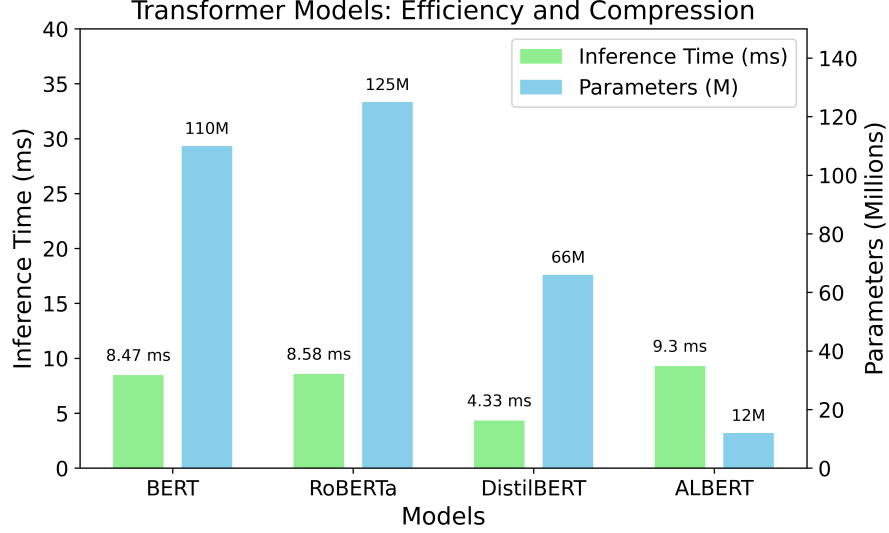


Fig. 7: Inference Efficiency and Parameter Complexity Comparison of Transformer Models on the CUFC dataset.

In addition to classification performance, the computational efficiency of each transformer architecture was evaluated on the CUFC dataset using inference time (milliseconds per sample) and the number of trainable parameters (millions). These measures provide complementary insights into the trade-off between predictive performance and deployment efficiency, as illustrated in Fig. 7. Among the evaluated models, DistilBERT achieves the lowest inference time (4.33 ms) while requiring only 66M parameters, making it well suited for latency-sensitive applications. BERT and RoBERTa require 8.47 ms and 8.58 ms per sample, respectively, reflecting the additional computational cost associated with their larger architectures. Although ALBERT contains substantially fewer parameters because of parameter sharing, it exhibits the highest inference time (9.30 ms), indicating that parameter count alone does not necessarily determine inference latency. Overall, the results demonstrate that the proposed CSHDA framework consistently improves classification performance while preserving the computational characteristics of the underlying transformer architectures, allowing practitioners to select an appropriate model according to the desired balance between predictive accuracy and computational efficiency.

6 Discussion

The experimental results confirm that the proposed Class-Aware Semantic Hybrid Data Augmentation (CSHDA) framework effectively addresses the challenges associated with class-imbalanced sentiment classification. As reported in result section, the proposed framework consistently improves macro-level performance for both datasets

i.e., the CUFC and Twitter Airline datasets, by generating semantically consistent and diverse training samples for minority sentiment classes.

The advantage of proposed framework is that it apply augmentation selectively on minority classes using class-aware probability distributions. This results in reducing the learning bias toward majority classes while preserving the original sentiment polarity. Whereas in conventional augmentation strategies apply identical transformations across all classes. These improvements also suggest that augmentation quality is more important than augmentation quantity i.e., increasing the number of minority samples may introduce semantic drift or redundant instances which may provide limited learning benefit. In contrast, the proposed framework combines rule-based perturbation techniques with transformer-based paraphrasing, SBERT-based semantic similarity validation, and duplicate-aware filtering to ensure that only high-quality augmented samples are retained. Consequently, the augmented dataset contains diverse yet semantically faithful training instances that improve the generalization capability of transformer-based sentiment classifiers.

Across both datasets, the proposed CSHDA framework consistently improves Macro F1-score and minority-class recall. The largest gains are observed for the neutral and negative sentiment classes, which contain considerably fewer training samples than the majority positive class. This behaviour is also reflected in the confusion matrices, where fewer minority-class instances are incorrectly assigned to the positive class after augmentation. The stronger diagonal entries indicate that the augmented training data enables the transformer models to distinguish sentiment classes more effectively instead of merely increasing overall accuracy.

The proposed CSHDA framework remains effective even when applied to datasets with substantially different characteristics. The CUFC corpus consists of relatively structured food reviews, whereas the Twitter Airline dataset contains short and noisy social media posts with informal language and abbreviations. Despite these differences, the framework improves classification performance on both datasets. This behaviour indicates that integrating class-aware augmentation with semantic similarity filtering and duplicate-aware validation is not restricted to a particular writing style or domain, making the framework suitable for cross-domain sentiment classification.

Another factor contributing to the effectiveness of the proposed framework is the use of class-aware augmentation probabilities. Instead of assigning identical probabilities to all sentiment classes, the probability distributions were determined empirically according to the characteristics of each dataset and its minority classes. For the CUFC corpus, where sentiment is expressed using relatively structured language, higher probabilities were assigned to semantically preserving transformations such as BERT-based paraphrasing. In contrast, the Twitter dataset contains informal expressions, abbreviations, and lexical variations, requiring a more diverse augmentation strategy. Accordingly, the probability distributions were adjusted to introduce greater lexical diversity while preserving the underlying sentiment. This adaptive strategy enables the augmentation process to better reflect dataset-specific characteristics than a fixed probability distribution.

The ablation study provides additional insight into the contribution of individual augmentation techniques. No single augmentation method consistently achieves the

highest performance across all transformer architectures. For example, Random Deletion performs competitively for certain models but does not maintain the same level of effectiveness across different architectures. By combining multiple complementary augmentation techniques within a class-aware framework, CSHDA produces more stable improvements across the evaluated transformer models. These findings suggest that augmentation quality depends not only on the choice of transformation but also on the interaction between different augmentation strategies while preserving semantic consistency.

From a practical perspective, the proposed CSHDA framework can be readily integrated into existing transformer-based sentiment analysis pipelines without modifying the underlying classifier architecture. Since augmentation is performed exclusively during the data preparation stage, the framework remains model-agnostic and can be readily integrated with different transformer architectures as well as other low-resource sentiment classification tasks.

Overall, the experimental findings indicate that combining class-aware augmentation, semantic similarity validation, and duplicate-aware filtering produces higher-quality minority-class training samples than conventional augmentation techniques. The consistent improvements observed across multiple transformer architectures and two datasets suggest that the proposed CSHDA framework provides an effective and robust solution for imbalanced sentiment classification.

7 Conclusion

In this paper, we proposed a Class-Aware Semantic Hybrid Data Augmentation (CSHDA) framework to address the class imbalance problem in transformer-based sentiment classification. The proposed framework combines class-aware augmentation probabilities with semantic consistency filtering and duplicate-aware validation to generate diverse yet semantically faithful minority-class samples. The effectiveness of the proposed framework was evaluated using four transformer-based models on two sentiment datasets, namely the CUFC corpus and the Twitter Airline dataset. Experimental results demonstrate consistent improvements in Macro F1-score and minority-class performance across most evaluated models. Furthermore, statistical significance analysis confirms that the observed improvements are statistically reliable, while cross-dataset evaluation demonstrates that the proposed framework generalizes well across both structured review data and noisy social media text.

Although the proposed CSHDA framework demonstrates consistent improvements in imbalanced sentiment classification, several limitations remain. First, the semantic similarity thresholds and class-aware probability configurations were empirically determined through pilot experiments and may not generalize optimally across all datasets and domains. Future work will investigate adaptive thresholding strategies that automatically adjust similarity bounds according to dataset characteristics or model feedback. Second, the augmentation pipeline incorporates multiple constraints, including sentence-length control, semantic similarity filtering, and duplicate-aware validation. Although these steps improve the quality of the generated samples, they

introduce additional computational overhead due to SBERT-based similarity computation. Future work may investigate more efficient semantic filtering techniques or approximate similarity measures to improve scalability. Third, although the proposed framework effectively addresses class imbalance by augmenting only minority classes, future work will explore instance-level augmentation strategies and active learning.

Finally, the proposed CSHDA framework has been evaluated only on English-language sentiment datasets and transformer-based sentiment classification models. Future work will investigate its applicability to multilingual corpora, other natural language processing tasks, and emerging large language models to further assess its generalization capability.

Author contributions statement

The following individuals made contributions to the formulation of conceptualization and methodology that was necessary in carrying out the augmentation pipeline: P.U. The following made contributions to the analysis and visualization of the data: G.L.S. Contributions to the reviewing and editing of the manuscript were made by A.M.

Declarations

Funding:

The authors did not receive any funding from outside sources to conduct this study. The Article Processing Charges (APC) were waived off as per the institutional policy of Manipal University Jaipur.

Conflicts of Interest:

The authors declare that they have no conflict of interest.

Data availability

The datasets used and/or analyzed during the current study are available from the following source: CUFC DataSet <https://www.kaggle.com/datasets/jhansiprs/cufc-chandigarh-university-food-court-review> Twitter DataSet <https://www.kaggle.com/datasets/crowdfower/twitter-airline-sentiment/data>

Code Availability:

The implementation of the proposed Class-Aware Semantic Hybrid Data Augmentation (CSHDA) framework, together with the scripts used for data augmentation, model training, and experimental evaluation, is publicly available on GitHub: https://github.com/jhansiprs/Smart_Augmentation_Twitter_Dataset.

Consent to Participate:

Not applicable.

Consent to Publish:

Not applicable.

Ethics Approval:

Not applicable.

Clinical Trial Number:

Not applicable.

References

- [1] Sun, T., Jing, L., Wei, Y., Song, X., Cheng, Z., Nie, L.: Dual consistency-enhanced semi-supervised sentiment analysis towards covid-19 tweets. *IEEE Transactions on Knowledge and Data Engineering* **35**(12), 12605–12617 (2023) <https://doi.org/10.1109/TKDE.2023.3270940>
- [2] Gupta, A., Prabhu, A., Pallavi, N., Phalapradha, M.J., Yashaswini, P.: Sentiment analysis of product reviews: Deep learning approach. In: 2025 12th International Conference on Computing for Sustainable Global Development (INDIACom), Delhi, India, pp. 01–06 (2025). <https://doi.org/10.23919/INDIACom66777.2025.11115522>
- [3] Manning, C.D., Raghavan, P., Schütze, H.: *Introduction to Information Retrieval*. Cambridge University Press, ??? (2008)
- [4] Pang, B., Lee, L.: Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval* **2**(1-2), 1–135 (2008)
- [5] Nair, A.J., G., V., Vinayak, A.: Comparative study of twitter sentiment on covid - 19 tweets. In: 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), pp. 1773–1778 (2021). <https://doi.org/10.1109/ICCMC51019.2021.9418320>
- [6] Moghadasi, M.N., Safari, Z., Zhuang, Y.: A sentimental and semantical analysis on facebook comments to detect latent patterns. In: 2020 IEEE International Conference on Big Data (Big Data), pp. 4665–4671 (2020). <https://doi.org/10.1109/BigData50022.2020.9378425>
- [7] Yang, L., Li, Y., Wang, J., Sherratt, R.S.: Sentiment analysis for e-commerce product reviews in chinese based on sentiment lexicon and deep learning. *IEEE Access* **8**, 23522–23530 (2020) <https://doi.org/10.1109/ACCESS.2020.2969854>
- [8] Gu, T., Xu, G., Luo, J.: Sentiment analysis via deep multichannel neural networks with variational information bottleneck. *IEEE Access* **8**, 121014–121021 (2020) <https://doi.org/10.1109/ACCESS.2020.3006569>

- [9] Vanaja, S., Belwal, M.: Aspect-level sentiment analysis on e-commerce data. In: 2018 International Conference on Inventive Research in Computing Applications (ICIRCA), pp. 1275–1279 (2018). <https://doi.org/10.1109/ICIRCA.2018.8597286>
- [10] Jagdale, R.S., Deshmukh, S.S.: Sentiment classification on twitter and zomato dataset using supervised learning algorithms. In: 2020 International Conference on Smart Innovations in Design, Environment, Management, Planning and Computing (ICSIDEMPC), pp. 330–334 (2020). <https://doi.org/10.1109/ICSIDEMPC49020.2020.9299582>
- [11] Upadhyay, A., Rai, S., Shukla, S.: Sentiment analysis of zomato and swiggy food delivery management system. In: Luhach, A.K., Poonia, R.C., Gao, X.-Z., Singh Jat, D. (eds.) Second International Conference on Sustainable Technologies for Computational Intelligence, pp. 39–46. Springer, Singapore (2022)
- [12] Abdellaoui, B., Remaida, A., Sabri, Z., Abdellaoui, M., El Hafidy, A., El Bouzekri El Idrissi, Y., Moumen, A.: Analyzing emotions in online classes: Unveiling insights through topic modeling, statistical analysis, and random walk techniques. *International Journal of Cognitive Computing in Engineering* **5**, 221–236 (2024) <https://doi.org/10.1016/j.ijcce.2024.05.003>
- [13] Kastrati, Z., Imran, A.S., Kurti, A.: Weakly supervised framework for aspect-based sentiment analysis on students’ reviews of moocs. *IEEE Access* **8**, 106799–106810 (2020) <https://doi.org/10.1109/ACCESS.2020.3000739>
- [14] Park, K., Park, S., Joung, J.: Contextual meaning-based approach to fine-grained online product review analysis for product design. *IEEE Access* **12**, 4225–4238 (2024) <https://doi.org/10.1109/ACCESS.2023.3343501>
- [15] Nkomo, L.M., Ndukwe, I.G., Daniel, B.K.: Social network and sentiment analysis: Investigation of students’ perspectives on lecture recording. *IEEE Access* **8**, 228693–228701 (2020) <https://doi.org/10.1109/ACCESS.2020.3044064>
- [16] Mao, J., Qian, Z., Lucas, T.: Sentiment analysis of animated online education texts using long short-term memory networks in the context of the internet of things. *IEEE Access* **11**, 109121–109130 (2023) <https://doi.org/10.1109/ACCESS.2023.3321303>
- [17] Naveed, M.S., Sajid, M., Sanjrani, A.A., Malik, S.S.: Sentiment analysis of diabetes patients’ experiences using machine learning techniques. *Journal of Computing & Biomedical Informatics* **6**(2), 172–181 (2024)
- [18] Alattar, F., Shaalan, K.: Using artificial intelligence to understand what causes sentiment changes on social media. *IEEE Access* **9**, 61756–61767 (2021) <https://doi.org/10.1109/ACCESS.2021.3073657>

- [19] Mao, Y., Liu, Q., Zhang, Y.: Sentiment analysis methods, applications, and challenges: A systematic literature review. *Journal of King Saud University - Computer and Information Sciences* **36**(4), 102048 (2024) <https://doi.org/10.1016/j.jksuci.2024.102048>
- [20] Abdelgwad, M.M., Soliman, T.H.A., Taloba, A.I., Farghaly, M.F.: Arabic aspect based sentiment analysis using bidirectional gru based models. *arXiv preprint arXiv:2101.10539* (2021)
- [21] Kang, H., Yoo, S.J., Han, D.: Senti-lexicon and improved naïve bayes algorithms for sentiment analysis of restaurant reviews. *Expert Systems with Applications* **39**, 6000–6010 (2012)
- [22] Tripathy, A., Agrawal, A., Rath, S.K.: Classification of sentimental reviews using machine learning techniques. *Procedia Computer Science* **57**, 821–829 (2015) <https://doi.org/10.1016/j.procs.2015.07.523> . 3rd International Conference on Recent Trends in Computing 2015 (ICRTC-2015)
- [23] Upadhyaya, P., Bhalla, S., Padhi, A., Rathi, K., Andey, A., Hooda, R.: Food items prediction using sentimental analysis. In: *Proceedings of the Third International Conference on Electronics and Sustainable Communication Systems (ICESC 2022)*, Chennai, India, pp. 751–756 (2022). <https://doi.org/10.1109/ICESC54411.2022.9885508> . IEEE
- [24] Singh, S., Krishnan, D., Sehgal, P., Sharma, H., Surani, T., Singh, J.: Gradient boosting approach for sentiment analysis for job recommendation and candidate profiling. In: *2022 IEEE Bombay Section Signature Conference (IBSSC)*, pp. 1–6 (2022). <https://doi.org/10.1109/IBSSC56953.2022.10037443>
- [25] Cheng, C., Tsai, S.-L.: Deep learning for automated sentiment analysis of social media. In: *2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pp. 1001–1004 (2019). <https://doi.org/10.1145/3341161.3344821>
- [26] Bommasani, R., Liang, P., *et al.*: Holistic evaluation of language models. *Transactions on Machine Learning Research (TMLR)* (2023) <https://doi.org/10.48550/arXiv.2211.09110> [arXiv:2211.09110](https://arxiv.org/abs/2211.09110)
- [27] Naveed, S., Husnain, M., Alsubaie, N.: Hybriddldr: A hybrid deep learning-based drug resistance prediction system of glioblastoma (gbm) using molecular descriptors and gene expression data. *Computer Methods and Programs in Biomedicine* **270**, 108913 (2025) <https://doi.org/10.1016/j.cmpb.2025.108913>
- [28] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: *Advances in Neural Information Processing Systems*, vol. 30, pp. 5998–6008 (2017). <https://arxiv.org/abs/1706.03762>

- [29] Vavekanand, R., Das, B., Kumar, T.: Dausindhi: a data augmentation approach for enhancing sindhi language text classification. *Discover Data* **3**(1), 22 (2025) <https://doi.org/10.1007/s44248-025-00040-8>
- [30] Melamud, O., Levy, O., Dagan, I.: A simple word embedding model for lexical substitution. In: *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1–7. Association for Computational Linguistics, Denver, Colorado (2015). <https://aclanthology.org/W15-1501>
- [31] Roller, S., Erk, K.: Pic a different word: A simple model for lexical substitution in context. In: *Proceedings of the NAACL-HLT 2016*, pp. 1121–1126. Association for Computational Linguistics, San Diego, California (2016)
- [32] Crowdfunder: Twitter US Airline Sentiment Dataset. Kaggle (2015)
- [33] Go, A., Bhayani, R., Huang, L.: Twitter Sentiment Classification using Distant Supervision. <http://twittersentiment.appspot.com/>. Stanford University project. (2009)
- [34] Tang, D., Qin, B., Liu, T.: Document modeling with gated recurrent neural network for sentiment classification. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Lisbon, Portugal, pp. 1422–1432 (2015). <https://doi.org/10.3115/v1/D15-1167>
- [35] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: *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)*, Minneapolis, Minnesota, pp. 4171–4186 (2019)
- [36] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019)
- [37] Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., Le, Q.V.: Xlnet: Generalized autoregressive pretraining for language understanding. In: *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, pp. 5754–5764 (2019)
- [38] Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R.: Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942* (2020)
- [39] Wei, J., Zou, K.: Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196* (2019)

- [40] Saeedi, D., Saeedi, S., Panahi, A., Fong, A.C.M.: Cs/nlp at semeval-2022 task 4: Effective data augmentation methods for patronizing language detection and multi-label classification with roberta and gpt3. In: Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pp. 503–508 (2022). Association for Computational Linguistics
- [41] Chen, J., Ding, Y., Han, X., Yin, W., Wang, W.Y., Qiu, X.: An empirical survey of data augmentation for limited data learning in nlp. Transactions of the Association for Computational Linguistics **11**, 191–211 (2023) <https://doi.org/10.1162/tacl.a.00542>
- [42] Sennrich, R., Haddow, B., Birch, A.: Improving neural machine translation models with monolingual data. In: Erk, K., Smith, N.A. (eds.) Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 86–96. Association for Computational Linguistics, Berlin, Germany (2016). <https://doi.org/10.18653/v1/P16-1009> . <https://aclanthology.org/P16-1009/>
- [43] Hugging Face: Pretrained models — Transformers 2.4.0 Documentation. https://huggingface.co/transformers/v2.4.0/pretrained_models.html. Accessed: 2025-12-19 (2019)